Towards Learning Structured Latent Visual Representations

by

Yibing Wei

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Computer Science)

at the

UNIVERSITY OF WISCONSIN–MADISON

2026

Date of final oral examination: 05/01/2026

The dissertation is approved by the following members of the Final Oral Committee:
Pedro Morgado, Member of Technical Staff, Skild AI (Chair)
Yong Jae Lee, Professor, Computer Science
Frederic Sala, Assistant Professor, Computer Science
Ramya Korlakai Vinayak, Assistant Professor, Electrical & Computer Engineering

To my family.

ACKNOWLEDGMENTS

I would first like to express my deepest gratitude to my advisor, Dr. Pedro Morgado. I am especially grateful that he welcomed me into his lab in 2022 and gave me the opportunity to become his first Ph.D. student. Through our many discussions, he taught me how to formulate research questions, design experiments, think critically about results, and navigate the uncertainties and challenges of research. His mentorship helped me grow into a more independent researcher, from early collaborative work to leading projects. Beyond research, Pedro has been a source of encouragement, kindness, and support. His scientific rigor, intellectual curiosity, optimism, and warmth have had a lasting influence on me, both professionally and personally.

I would also like to thank my committee members, Professor Yong Jae Lee, Professor Fred Sala, and Professor Ramya Korlakai Vinayak, for their time, feedback, and support throughout my Ph.D. journey. I am grateful for their guidance during my qualification exam, preliminary exam, and dissertation defense, as well as for the thoughtful questions and suggestions that helped improve my work. I also thank Professor Abhinav Gupta from Carnegie Mellon University for his support on my first project, LatentMIM. I am deeply grateful to my collaborators from PAII, including Zhicheng Yang, Hang Zhou, Mei Han, Jui-Hsin Lai, Andy Wong, and Chen Du, and to my collaborators from Adobe, including Sudeep Katakol, Manuel Brack, Yu-Teng Li, Richard Zhang, Eli Shechtman, Hareesh Ravi, and Ajinkya Kale, for their insights, discussions, and support. These collaborations broadened my perspective and greatly enriched my Ph.D. research experience.

I am also grateful to my colleagues at the University of Wisconsin-Madison, including Cheng-En Wu, Lin Zhang, Jinhong Lin, and Haoyue Bai, for their companionship, support, and many inspiring conversa-

tions along the way. Being part of this community, where we could exchange ideas, share challenges, and grow together, has been one of the most meaningful parts of my Ph.D. experience. I would also like to thank Samuel Church and Victor Suciu, whom I had the opportunity to mentor through our work on TrackVerse. Working with them was a deeply rewarding experience and helped me better appreciate the value of collaboration, guidance, and shared growth.

Finally, I would like to thank my family and my partner for their unconditional love, trust, and encouragement. Their belief in me has been one of my greatest sources of strength throughout this journey. I could not have reached this point without their support.

CONTENTS

Contents iv

List of Tables vi

List of Figures viii

Abstract x

1 Introduction 1

2 Towards Latent Masked Image Modeling for Self-Supervised Visual Representation Learning 6
- *2.1 Introduction* 7
- *2.2 Latent Masked Image Modeling* 11
- *2.3 Challenges of Latent MIM* 14
- *2.4 Scaling to ImageNet-1k* 24
- *2.5 Conclusion* 28

3 TrackVerse: A Large-Scale Dataset of Object Tracks for Visual Representation Learning 29
- *3.1 Introduction* 29
- *3.2 Related Work* 32
- *3.3* TrackVerse *Dataset* 33
- *3.4 Pretraining Protocol* 39
- *3.5 Downstream Evaluations* 42
- *3.6 Conclusion* 49

4 Queryable Attribute Representation Extraction from Frozen Vision-Language Models 50
- *4.1 Introduction* 50

4.2 Related Work 53
4.3 The QARE *Benchmark* 54
4.4 Method 61
4.5 Experiments 64
4.6 Discussion 67
4.7 Summary 68

5 Conclusion 69

References 73

LIST OF TABLES

2.1 Target encoder optimization. In addition to downstream performance, we report the average pairwise cosine similarity between (mean pooled) latents Z. 17
2.2 Loss Impact. 19
2.3 Improvement from reducing semantic correlation mitigation with strategies of high mask ratio, non-contiguous stochastic maskin and patch-wise similarity constraint. 21
2.4 Comparison between different decoder designs. Other configurations follow the optimal settings found in previous sections. 24
2.5 Top-1 NN and LP classification accuracy on ImageNet. 25
2.6 **Video object segmentation** on DAVIS-2017. $\mathcal{J}$ and $\mathcal{F}$ quantify region similarity and boundary alignment, respectively. 27
2.7 **Few-shot learning on various datasets.** We report the mean and standard deviation of the top-1 accuracy across 3 different runs. Each run is trained on a set of 16 training images per class. 27

3.1 `TrackVerse` **vs. Existing Video Datasets.** Full Dyn.: full dynamics that supports random frame sampling at any time point. The R2V2 dataset only contains 4 fixed frames per video. Obj Centric: object-centric dataset. Class Bal: Class Balance where the high-frequency classes were capped. Dur: total duration of the dataset. 31

3.2 **Evaluation of representations across diverse downstream tasks**: LVIS-IN and TrackVerse object classification, SSv2 action recognition, MIT-States object/attribute classification, and DAVIS-2017 video object segmentation. Features were pre-trained using MoCo on ImageNet and variance-aware MoCo on videos, both with equivalent training budget of 220M forward-backward passes. 43
3.3 **Scalability to Dataset Size.** Increasing the dataset size of TrackVerse significantly improves downstream performance. 46
3.4 **Comparison of static frames (MoCo) vs. tracks (Variance-Aware MoCo) pretraining on TrackVerse** 48

4.1 **Comparison of different methods on the QARE benchmark.** We evaluate three distinct families of methods: (1) VLM2Vec variants that fine-tune VLMs to produce queryable embeddings; (2) Standard visual encoders that output a single, entangled global embedding; and (3) Our proposed training-free approach TF-QARE directly extracts disentangled attribute features from frozen VLMs and consistently achieves substantial gains, demonstrating the effectiveness of prompt-guided, attribute-aware embedding extraction. Higher mAP (↑) and lower AIS (↓) indicate better performance, and the gray row highlights our default model. 64
4.2 Ablations on layer selection. 67

LIST OF FIGURES

2.1 **Challenges of Latent MIM.** The representations learned by MIM approaches fail to capture high-level semantics, as shown by the poor performance in nearest neighbor and linear probe evaluation. 8
2.2 **Latent Masked Image Modeling Overview**. Models are trained to reconstruct the latent representations generated by a target encoder at withheld locations. Four major challenges for effectively deploying Latent MIM are identified in this chapter, as well as potential solutions. These challenges relate to joint encoder optimization, direct reconstruction loss, the semantic correlation between visible and target patches, and the decoder design. 12
2.3 Training Collapse of the Naive Latent MIM. This solution achieves a zero reconstruction loss but fails to capture any meaningful information about the input images. As a result, the nearest neighbor (NN) evaluation yields random performance. Top: NN accuracy; Bottom: training loss. 16
2.4 Patch Generation Strategies. Left: Masking contiguous grids. Right: Non-contiguous stochastic masking. 21
2.5 **Comparison of Three Different Decoder Designs.** Self-attention decoder is commonly used for low-level MIM models. Cross-attention decoder provides direct conditioning at each layer on the visible latents. 22
2.6 **Unsupervised segmentation maps** by hierarchical clustering of patch representations within each image. Better viewed digitally with zoom. 26
2.7 **Visualization**. Col 1: ground truth; Cols 2–4: predictions at 25%, 50%, and 100% timesteps. 27

3.1 **Examples from TrackVerse** TRACKVERSE comprises object-centric video tracks with a diverse range of objects. 34
3.2 **Overview of TrackVerse Collection Pipeline.** 35
3.3 **Characteristics of the Full TrackVerse.** (a) The number of videos per class, illustrates the long-tail distribution of the dataset; (b) Distribution of object track durations; (c) Distribution of track size where size is the short side of the object's bounding box; (d) Distribution of aspect ratios of all object tracks. 38
3.4 **TrackVerse's Scalability to Training Budget.** Compared to SSL pretraining on static images, TRACKVERSE dynamic tracks prevent overfitting and enhance representation learning when using longer training schedules. 44

4.1 Overview of QARE: from entangled global embeddings to queryable attribute-specific representations. 51
4.2 QARE-BENCH synthetic set. 56
4.3 Examples from the QARE-BENCH Real Set. 57
4.4 QARE-BENCH Real Set image source. 59
4.5 Overview of TF-QARE with attribute-focused prompting. . . . 62

ABSTRACT

Visual representation learning has advanced rapidly through self-supervised learning, contrastive methods, and vision-language pre-training. Yet most existing approaches produce global, unstructured embeddings that compress visual content into a holistic summary, conflating distinct visual factors and discarding the rich internal organization of images and videos. This dissertation argues that progress requires moving beyond global semantic summaries toward structured latent representations that explicitly capture this internal structure, and addresses three complementary dimensions in which current methods fall short.

The first dimension is spatial structure. This thesis investigates Latent Masked Image Modeling, which predicts representations of masked patches in latent space rather than raw pixels. A systematic analysis identifies representational collapse, inter-patch redundancy, and poor decoder design as core failure modes, and proposes principled remedies that yield representations that are simultaneously spatially grounded and semantically rich, with substantial gains in retrieval, linear probing, and unsupervised segmentation.

The second dimension is state structure. Standard contrastive objectives collapse semantically distinct object states under invariance, and existing video datasets are not object-centric. This thesis introduces TrackVerse, a large-scale object-centric video dataset capturing the natural evolution of everyday objects across poses, interactions, and state changes, together with a relaxed contrastive objective that preserves sensitivity to semantic-changing transformations while enforcing invariance to semantic-preserving ones.

The third dimension is compositional attribute structure. This thesis formalizes Queryable Attribute Representation Extraction, where an

embedding must be sensitive to a specified attribute — such as foreground object, background, or style — while remaining invariant to others. A new benchmark enables rigorous attribute-level evaluation, and a training-free method that pools vision-language model hidden states from attribute-conditioned reply tokens substantially outperforms fine-tuned and global embedding baselines.

Together, these contributions advance a unified vision: structured latent representations, spanning spatial, state-aware, and compositional dimensions, are essential for the next generation of visual understanding systems.

1 INTRODUCTION

Visual perception lies at the heart of artificial intelligence. The ability to understand and interpret visual data underpins a wide range of real-world applications, from autonomous systems and medical imaging to content retrieval and human-computer interaction. Central to this capability is the quality of visual representations — the learned encodings of images and videos that downstream models rely upon to perform recognition, reasoning, and generation tasks.

The past decade has witnessed remarkable progress in visual representation learning. Early approaches relied on large-scale supervised training on datasets such as ImageNet [40], using human-annotated labels to shape the learned feature space. The advent of self-supervised learning removed this dependency on manual annotation, enabling models to learn powerful representations directly from unlabeled visual data. Masked Image Modeling (MIM) approaches learn by reconstructing masked regions of an image, encouraging spatially aware local representations [6, 8]. Contrastive learning methods learn by pulling together different augmented views of the same image, encouraging semantic invariance [2]. More recently, vision-language models have extended representation learning to the multimodal setting, aligning visual features with natural language through large-scale image-text pretraining [47, 48, 101, 98]. Together, these advances have produced general-purpose visual representations that transfer effectively across a broad range of downstream tasks.

Despite this progress, a fundamental limitation persists across existing approaches: the representations they learn are predominantly **global and unstructured**. Whether through reconstruction objectives, contrastive objectives, or vision-language alignment, these methods optimize for a holistic summary of visual content — compressing an

image into a single vector or a set of loosely organized patch features that conflate distinct visual factors. This design choice, while effective for coarse-grained recognition tasks, leaves a significant gap: it fails to capture the rich **internal structure** inherent in visual data. Images and videos are not monolithic signals; they are composed of spatially organized parts, objects that evolve and interact over time, and distinct semantic attributes — such as object identity, background context, and visual style — that co-exist within a single scene. When these structural dimensions are entangled or ignored in the learned representation, the resulting features are ill-suited for tasks that require fine-grained understanding, attribute-level control, or structured reasoning.

This dissertation addresses this gap. We argue that advancing visual representation learning requires moving beyond global semantic summaries toward **structured latent representations** — representations that explicitly capture the internal organization of visual data across multiple dimensions of structure. Specifically, we identify three fundamental and complementary dimensions that current methods inadequately model: the **spatial structure** of local image regions, the **state structure** of objects as they evolve across time and interaction, and the **compositional attribute structure** of images as a mixture of distinct semantic factors. Each of these dimensions reflects a different facet of how visual data is internally organized, and each demands a distinct perspective on how representations should be learned.

This dissertation investigates the problem of learning structured latent visual representations — representations that go beyond holistic semantic summaries to capture the internal organization of visual data. We approach this problem along three complementary structural dimensions. The first concerns **spatial structure**: how can self-supervised learning methods be designed to yield patch-level representations that are both spatially aware and semantically rich, rather than sacrificing

one for the other? The second concerns **state structure**: how can video data — which naturally records how objects evolve, interact, and change state over time — be leveraged to learn object representations that are sensitive to semantically meaningful state variation, rather than collapsing it away? The third concerns **compositional attribute structure**: given that a visual scene is composed of distinct factors such as object identity, background, and style, how can we learn representations that are queryable with respect to a specific attribute, rather than entangling all factors into a single undifferentiated embedding? Together, these three lines of inquiry constitute a systematic exploration of structured visual representation learning, each advancing the goal of learning latent spaces that are not merely semantically informative, but structurally interpretable and compositionally controllable.

Chapter 2 addresses the spatial structure of visual representations through a systematic investigation of Latent Masked Image Modeling (Latent MIM) for self-supervised visual representation learning. Standard MIM methods learn spatially grounded patch-level representations by reconstructing masked image regions from visible context [6, 8]. However, because these methods reconstruct raw pixels, the learned representations are constrained to retain low-level appearance details, limiting their ability to encode high-level semantics — a limitation evidenced by their poor performance under linear probing and nearest-neighbor evaluation. A natural remedy is to perform masked image modeling in latent space, predicting the representations of masked patches rather than their pixel values. Yet naive adaptations of this idea are unstable: they are prone to representational collapse, suffer from high inter-patch redundancy that allows trivial prediction by copying nearby visible features, and are sensitive to decoder design. This chapter conducts an in-depth analysis of these challenges, identifying four core failure modes of Latent MIM and proposing principled solutions to each. The resulting framework

substantially improves over pixel-level MIM baselines [6] and prior Latent MIM methods [14], achieving significant gains in nearest-neighbor classification, linear probing, unsupervised scene segmentation, and video object segmentation — demonstrating that Latent MIM, when carefully designed, can learn representations that are simultaneously spatially structured and semantically rich.

Chapter 3 addresses the state structure of object representations through the introduction of TrackVerse, a large-scale object-centric video dataset designed for visual representation learning. Video data offers a rich source of structural information that static image datasets cannot provide: it naturally records how objects change pose, undergo state transitions, and interact with their environment over time. However, existing video datasets are poorly suited for leveraging this information [55, 56, 58, 45]. Natural video is not object-centric, making it difficult to isolate consistent object signals; and existing contrastive learning methods, designed around invariance objectives [2], tend to collapse the representations of different object states rather than preserve them. TrackVerse addresses the data challenge by providing a large-scale, balanced, object-centric collection of object tracks, constructed via an automated pipeline using open-vocabulary detection and tracking. Beyond the dataset, this chapter demonstrates that effectively exploiting object tracks for representation learning requires a relaxed contrastive objective — one that enforces invariance to semantic-preserving transformations while remaining sensitive to semantic-changing ones, such as the distinct states an object passes through over time. Representations learned under this framework outperform those trained on non-object-centric video datasets and equivalent static image datasets [40], demonstrating the value of state-aware structure for object representation learning.

Chapter 4 addresses the compositional attribute structure of visual representations in the multimodal setting. A visual scene is not a mono-

lithic signal but a composition of distinct semantic factors — object identity, background context, and visual style — that are entangled in the representations produced by current vision-language models [47, 101, 98]. Existing multimodal embeddings, including those derived from powerful VLMs such as VLM2Vec [96], produce global features that conflate these factors, making it impossible to isolate a specific attribute conditioned on a user query. This chapter formalizes this limitation as the problem of Queryable Attribute Representation Extraction (QARE): given an image and an attribute query, the goal is to produce an embedding that is sensitive to the specified attribute and invariant to all others. To support rigorous evaluation of this capability, we introduce a new benchmark covering three orthogonal attributes — object, background, and style — with both synthetic and real-image evaluation sets, addressing gaps in existing evaluation protocols [96, 105, 107, 108, 109]. We further propose a surprisingly effective training-free solution: extracting attribute-specific embeddings by pooling the hidden states of VLM reply tokens generated in response to a structured attribute query. This approach exploits the implicit attribute structure already encoded in modern VLMs without any fine-tuning, and substantially outperforms both post-trained and global embedding baselines across all attributes and VLM backbones — establishing a strong foundation for compositionally controllable multimodal representation learning.

Chapter 5 concludes the dissertation, summarizing the contributions across these three structural dimensions and outlining open questions and future directions for structured visual representation learning.

2 TOWARDS LATENT MASKED IMAGE MODELING FOR SELF-SUPERVISED VISUAL REPRESENTATION LEARNING

Masked Image Modeling (MIM) has emerged as a promising method for deriving visual representations from unlabeled image data by predicting missing pixels from masked portions of images. It excels in region-aware learning and provides strong initializations for various tasks, but struggles to capture high-level semantics without further supervised fine-tuning, likely due to the low-level nature of its pixel reconstruction objective. A promising yet unrealized framework is learning representations through masked reconstruction in latent space, combining the locality of MIM with the high-level targets. However, this approach poses significant training challenges as the reconstruction targets are learned in conjunction with the model, potentially leading to trivial or suboptimal solutions. Our study is among the first to thoroughly analyze and address the challenges of such a framework, which we refer to as Latent MIM. Through a series of carefully designed experiments and extensive analysis, we identify the source of these challenges, including representation collapse from joint online/target optimization, learning objectives, high inter-region correlation in latent space, and decoder conditioning. By sequentially addressing these issues, we demonstrate that Latent MIM can indeed learn high-level representations while retaining the benefits of MIM models. Code is available at `https://github.com/yibingwei-1/LatentMIM`.

2.1 Introduction

Masked Image Modeling (MIM), a learning framework that derives visual representations from unlabeled image data, has recently gained prominence. This technique masks a substantial part of an image and trains a model to predict the missing pixels using the surrounding context. Despite the simple learning objective, MIM has been shown to learn powerful representations, which, when fine-tuned, can achieve state-of-the-art performance on a variety of downstream tasks, including object classification, detection, and segmentation [6, 8]. MIM approaches offer several key advantages over other self-supervised visual representation learning methods. By requiring the model to accurately reconstruct all masked patches, MIM incentivizes the model to maintain distinct local representations of each image region and forces the model to reason over the spatial layout of object subparts. These benefits make MIM a popular approach for learning visual representations in a self-supervised manner.

However, the representations learned through this framework fail to capture high-level semantics without further supervised fine-tuning, as shown by their lower performance in linear probing and nearest-neighbor classification. We hypothesize that MIM does not directly learn high-level semantics due to the use of low-level learning targets like raw pixels. To effectively reconstruct the high-frequency details, the learned representations must retain a low-level description of the image, which limits their ability to encode higher-level semantics.

A natural solution to this challenge is to perform *masked image modeling in latent space*, a general framework that we refer to as Latent MIM. This means that instead of reconstructing raw pixels, Latent MIM methods should learn representations by defining local (patch-wise) latent representations for each image and reconstructing the representations of masked regions from visible ones. By bypassing low-level pixel targets and focusing on distinguishing the latent representations of

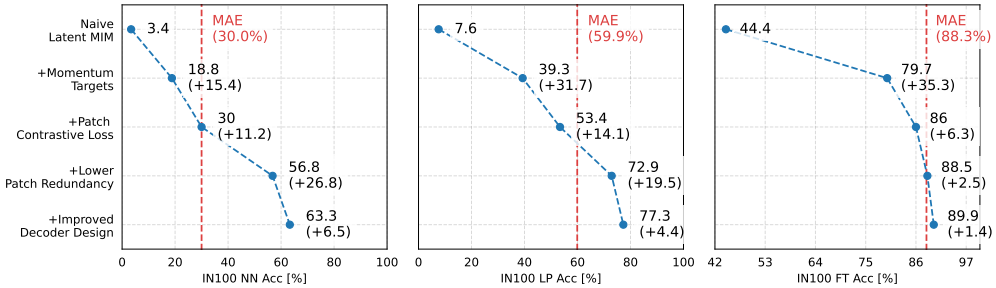


Figure 2.1: **Challenges of Latent MIM.** The representations learned by MIM approaches fail to capture high-level semantics, as shown by the poor performance in nearest neighbor and linear probe evaluation.

masked regions, Latent MIM methods hold the promise of substantially enhancing the semantics obtained through pixel-level MIM approaches and potentially their downstream performance.

However, intuitive adaptations of MIM to latent space reconstruction are unstable due to (1) the *existence of trivial solutions* and (2) the *high correlation between the semantics of nearby patches*. First, trivial solutions occur when the learned latent devolves into easy-to-predict but uninformative representations of the input images; for example, using a mean squared error (MSE) loss, the decoder might generate constant latent representations across all regions. Second, even if latent representations are indicative of the contents of each region, the semantics of nearby patches are more correlated with each other than the pixels themselves. This can also lead to poor representations, as the model can learn to predict masked regions by simply copying from nearby visible regions. As a result, despite its great potential, these challenges have significantly impeded the advancement of Latent MIM.

In this chapter, we conduct an in-depth investigation into the challenges of pure latent MIM. Our objective is to provide a thorough analysis, identifying, characterizing, and exploring potential solutions to the core challenges. Fig. 2.1 provides an overview of our analysis, showing the

downstream classification performance as we progress through four challenges, with significant improvements in nearest neighbor, linear probe, and fine-tuning accuracy.

(1) *Joint optimization of visible and masked regions representations leads to representation collapse.* Instead, as popularized in contrastive learning methods, creating asymmetries between the two representations, while avoiding the target encoder to contribute to the gradient computation, is crucial for learning meaningful representations.

(2) *Direct reconstruction results in poor representations, regardless of the specific loss used*. Beyond direct reconstruction, we explore contrastive predictive coding within image patches[1], which not only encourages the model to predict representations that are similar to those of target patches (thus achieving the reconstruction objective) but also encourages richer and spatially diverse representations across the image.

(3) *Controlling the redundancy between visible and target patches is critical.* Specifically, the optimal masking strategies for MIM in pixel space (*i.e.*, 75% masking of regular grids) can result in a set of target patches whose representations are too similar to (and thus easily predicted from) visible patches. To reduce the redundancy between patches, we investigate different procedures to generate the visible and target token sequences, including the use of non-contiguous grids, where nearby patches are separated by a random number of pixels, higher mask ratios and explicit patch similarity constraints.

(4) *The decoder design must be carefully crafted since Latent MIM decoders predict high-level representations of mask regions from those of visible regions.*

[1]We explicitly avoid batch-wise contrastive objectives to focus our study on pure masked reconstruction.

The decoder needs to have enough capacity to fill in the missing information but is simultaneously shallow enough to avoid taking on the role of the encoder (*i.e.*, computing the image semantics). To this end, we investigate a variety of decoder architectures. We observe that cross-attention decoders with modified mask tokens that directly encode the representation of nearby visible regions were particularly effective.

By explicitly investigating the inherent challenges of Latent MIM, we demonstrate that the Latent MIM framework can indeed be used to learn richer semantics and diverse patch-wise representations compared to existing MIM approaches, without the need for supervised fine-tuning. We further validate our findings by scaling up an instantiation of the Latent MIM informed by our findings to ImageNet-1K. The learned representations achieved 50.1% nearest neighbor (a 37.9% gain over MAE) and 72.0% linear probing (+4.2% gain over MAE) top-1 classification accuracy. The semantics and localizability of the learned representations are further demonstrated in three distinct tasks that require robust local representations: unsupervised scene segmentation, video object segmentation, and few-shot transfer learning across a variety of tasks. Remarkably, representations learned through Latent MIM approaches are capable of generating accurate segmentation of visual scenes, even in the absence of supervision. In video object segmentation and few-shot transfer learning, Latent MIM surpasses both the pixel-level MIM techniques, such as MAE, and previous Latent MIM methodologies like data2vec.

2.2 Latent Masked Image Modeling

Framework Overview

Latent MIM is a self-supervised learning framework that aims to learn visual representations through masked image modeling in latent space. As illustrated in Fig. 2.2, Latent MIM models comprise three components: an *online encoder* $f(\cdot)$, a *target encoder* $f_T(\cdot)$, and a *decoder* $g(\cdot)$. Similar to pixel-based MIM approaches [6, 8], each image is first divided into a set of L small patches x_i along with their location within the image p_i. This set $X = \{(x_i, p_i)\}_{i=1}^{L}$ is randomly split into two disjoint sets: the *visible* $X_\mathcal{V} = \{(x_i, p_i)\}_{i\in\mathcal{V}}$ and *target* $X_\mathcal{T} = \{(x_i, p_i)\}_{i\in\mathcal{T}}$ patches, where $\mathcal{V}$ and $\mathcal{T}$ are non-overlapping index sets of grid locations. As the name suggests, the visible patches are used to generate the latent representation of an image, while the target patches are used as the reconstruction targets. To accomplish this, the online encoder extracts latent representations of the visible patches

$$\{z_i\}_{i\in\mathcal{V}} = Z_\mathcal{V} = f(X_\mathcal{V}), \tag{2.1}$$

which are then used to inform the decoder for predicting patches at the target locations $P_\mathcal{T} = \{p_i\}_{i\in\mathcal{T}}$

$$\hat{Z}_\mathcal{T} = g(Z_\mathcal{V}, P_\mathcal{T}). \tag{2.2}$$

However, instead of predicting pixel values, Latent MIM imposes the latent representations obtained from the *target encoder* as the reconstruction targets,

$$\{z_i\}_{i\in\mathcal{T}} = Z_\mathcal{T} = f_T(X_\mathcal{T}). \tag{2.3}$$

Latent MIM models can then be trained to minimize the discrepancy Δ between the predicted target representations $\hat{Z}_\mathcal{T}$ and those obtained from

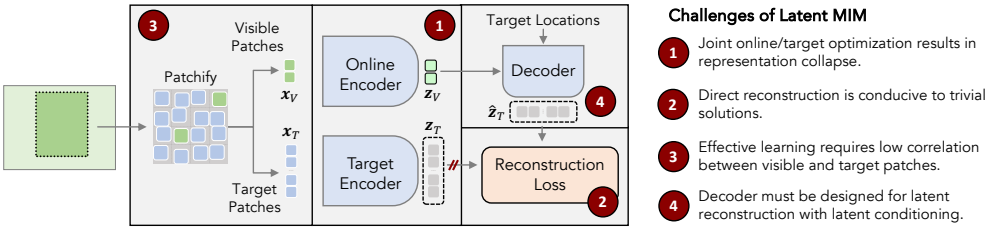


Figure 2.2: **Latent Masked Image Modeling Overview**. Models are trained to reconstruct the latent representations generated by a target encoder at withheld locations. Four major challenges for effectively deploying Latent MIM are identified in this chapter, as well as potential solutions. These challenges relate to joint encoder optimization, direct reconstruction loss, the semantic correlation between visible and target patches, and the decoder design.

the target encoder $Z_{\mathcal{T}}$. This is achieved by a reconstruction loss

$$\mathcal{L}_{\mathrm{rec}} = \mathbb{E}\left[\Delta\left(Z_{\mathcal{T}}, \hat{Z}_{\mathcal{T}}\right)\right], \tag{2.4}$$

where the expectation is taken over the training dataset. In pixel-based MIM, the mean squared error (MSE) is a popular choice for Δ, $\Delta = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}} \|\hat{z}_i - z_i\|^2$. However, as shown in Section 2.3, the MSE loss is not effective for Latent MIM.

Distinctions and similarities from related frameworks

Masked Image Modeling (MIM) methods learn by predicting masked parts of an image using targets derived from simple transformations of the original image. Targets can be raw pixels [6, 25], hand-crafted features such as HOG [7] or pre-trained features [8]. Thus, low-level MIM is required to maintain localized representations of an image. However, the model capacity is partially consumed by low-level details, limiting its capacity for high-level semantics.

Contrastive Learning (CL) In contrast, CL learns representations of each image in the context of the dataset in which they occur [2, 1, 3]. While CL can achieve outstanding semantics without finetuning, learning in context can be undesirable, for example, when the data distribution does not provide appropriate negative samples that highlight meaningful semantic distinctions between images. Batch dependencies also make contrastive objectives less flexible and reproducible, as they depend on the available compute resources. Instead, pure Latent MIM objectives learn representations from the image itself, and independently from other images.

Previous explorations for Latent MIM The potential of Latent MIM is tied to its ability to learn from high-level and region-aware targets, which are continuously improved throughout training. However, prior works attempting to deploy Latent MIM [27, 15, 28, 16, 14, 29] do not directly tackle the core optimization challenges inherent in Latent MIM discussed in Section 2.3. These challenges have been mitigated by treating latent MIM as a supplementary objective to other techniques, such as global contrastive learning [15, 28], low-level reconstruction [16], or alternatively, by using fixed pre-trained features as the latent targets [8, 27], instead of jointly learning the target and online representations.

The most closely related prior works focused on pure latent MIM, include data2vec [14], ConMIM [29], and I-JEPA [32]. Despite their contribution, data2vec and ConMIM use mask tokens instead of patch removal to hide image regions, causing a mismatch between pre-training and real-world deployment with unmasked images, which leads to poor performance in downstream tasks without finetuning. I-JEPA learns by predicting contiguous regions in latent space via an MSE loss, potentially leading to low-resolution ("blurred") or less localizable semantics (one of the major benefits of pixel-based MIM models).

While these works have advanced beyond pixel-level MIM, they have not thoroughly analyzed or addressed the inherent challenges specific to Latent MIM. Therefore, we argue that the full potential of this framework remains untapped, positioning Latent MIM as a significant, high-reward research area. Latent MIM has the potential to generate rich and localizable high-level semantic representations while maintaining the diverse patch-wise representations characteristic of MIM approaches.

2.3 Challenges of Latent MIM

To better focus on the various difficulties of Latent MIM optimization, we introduce and analyze each challenge separately. We begin by describing a naive implementation that fails to learn meaningful representations. We then introduce each of the four challenges (Section 2.3–2.3) and conduct a thorough analysis of potential strategies to address them. Each section builds on the findings of the preceding one, yielding increasingly effective instantiations of the Latent MIM framework. Fig. 2.1 shows the overview of performance through this progression.

Experimental Design

Pre-training For a comprehensive experimental analysis within an acceptable compute budget, we conduct experiments using the standard ViT-B transformer backbone for both the online and target encoders and trained the model on the ImageNet-100 (IN100) dataset, a subset of ImageNet-1k, containing 100 classes selected at random. The class partition used follows that of [5, 12]. This dataset contains approximately 125,000 images, sufficient for executing experiments with statistical significance. Section 2.4 further discusses the scaling properties of the proposed framework in the context of larger datasets, like ImageNet-1k.

All models were pre-trained for 300 epochs using the AdamW [33] optimizer with a batch size of 1024, a base learning rate of 1.5×10^{-4}, following 30 warm-up epochs and a cosine decay schedule. Data augmentations were applied to the input images, including random horizontal flipping and random image cropping with a minimum crop area of 0.2.

Downstream tasks We evaluated all models using nearest-neighbor, linear probing, and fine-tuning on the same dataset. For *nearest-neighbor* evaluation, we extract the representations for all images in the test set and report the fraction of samples whose nearest neighbor shares the same class. For *linear probing*, we train for 20 epochs using the LARS optimizer [18] with a base learning rate of 0.5, 2 warm-up epochs, and a batch size of 1024. For *fine-tuning*, we train for 50 epochs using an AdamW [33] optimizer with a base learning rate of 0.001, a weight decay of 0.05, 5 warm-up epochs, and a batch size of 1024.

Naive Latent MIM

To fully appreciate the complexities involved in learning meaningful representations through Latent MIM, we initiate our discussion with a simple and intuitive implementation, by closely following those of pixel-based MIM models [6]. Input images of resolution 224×224 are divided into a regular 14×14 grid of patches and partitioned into a 25/75% split to create the sets of visible/target patches, respectively. We then use a standard ViT-B transformer [9] to encode both the visible and target patches (f and f_T) and a transformer decoder g to reconstruct the target patches from the visible latent representations. We use the same self-attention decoder as in MAE [6], but reduce its depth to only 3 layers since, in Latent MIM, the decoder task (predicting latents from latents) is less complex than in pixel-based MIM (predicting pixels from latents). Both online and target encoders and the decoder are jointly trained using

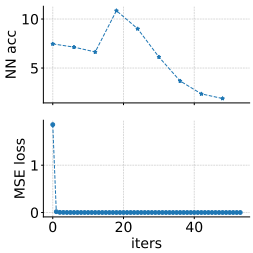


Figure 2.3: Training Collapse of the Naive Latent MIM. This solution achieves a zero reconstruction loss but fails to capture any meaningful information about the input images. As a result, the nearest neighbor (NN) evaluation yields random performance. Top: NN accuracy; Bottom: training loss.

the mean squared error (MSE) loss between the predicted and target latents.

As shown in Fig. 2.3, the Naive Latent MIM model is highly unstable. We observe that within a short number of iterations, the model collapses into a degenerate solution. The subsequent sections seek to understand the underlying causes of this finding and propose strategies to mitigate it.

Challenge 1: Latent Target Optimization

Representation collapse in latent MIM is consistent with the findings of negative-free contrastive learning methods like BYOL [3]. Unsurprisingly, because both the online and target encoders are trained to minimize the discrepancy between their representations, they can easily lead to a degenerate solution where all images are mapped to the same latent. Common strategies to prevent collapse involve introducing asymmetries between encoders and detaching the target encoder from gradient computation [3]. We explore three strategies:

Stand-alone target encoder. The first strategy is to treat the target encoder independently of the online encoder, without weight sharing.

Weight-sharing with stop gradient. Alternatively, we can use a siamese architecture with shared weights, but avoid using the target's gradients to update the online encoder.

Table 2.1: Target encoder optimization. In addition to downstream performance, we report the average pairwise cosine similarity between (mean pooled) latents Z.

Method	Sim	NN	LP	Ft
Naive Latent-MIM	1.00	0.9	7.9	—
No weight sharing	0.96	3.4	7.6	44.4
Weight sharing & stop-grad	0.99	11.3	26.6	56.9
Momentum targets	**0.50**	18.8	39.3	79.7
MAE [6]	0.67	**30.0**	**70.4**	**88.3**

Momentum targets. Using Momentum encoder [13] is also a common strategy to create asymmetries, and enhance the target encoder simultaneously. As $f(\cdot)$ evolves, the exponential moving average of its weights is tracked and used as the weights of the momentum encoder $\bar{\theta} \leftarrow m\theta + (1-m)\bar{\theta}$.

Study Results We report the performance of the various models in Table 2.1. To assess representation collapse, we computed the pairwise similarity between samples in the latent space. As expected, the naive latent-MIM model collapses, with all samples mapped to the same latent (similarity of 1.0). As indicated by the lower cosine similarities, all three strategies help prevent full representation collapse, with momentum targets consistently outperforming the other strategies on downstream classification tasks. However, none of the strategies are able to match the performance of MAE, suggesting that latent target optimization is not the only factor hampering the success of latent MIM.

Challenge 2: Reconstruction Objective for Latent MIM

In pixel-based MIM, patch reconstruction is enforced by minimizing the mean squared error (MSE) between the decoder's output and the pixel intensities at the target locations. In order to accurately reconstruct the target pixels from a limited visible context, the model is encouraged to

learn representations that capture both global semantics as well as patch-specific information. However, since in Latent MIM, the targets are the learned latent representations, we hypothesize that direct reconstruction objectives can also contribute to the optimization challenges, as there are no negative samples to stabilize the learning process. Thus, we investigate the impact of different reconstruction objectives on the effectiveness of Latent MIM.

Direct reconstruction We study three loss functions that directly minimize the discrepancy between predicted and target representations, namely MSE, L1, and Huber losses. While the MSE is widely used due to its simplicity and effectiveness, the L1 loss is robust to outliers. The Huber loss combines the best properties of the MSE and L1 losses by being quadratic for small errors and linear for large ones, thus providing a balance between robustness and efficiency. Mathematically, the reconstruction losses for the k-th target patch are

$$\Delta_{L2}^{k} = \|\hat{\boldsymbol{z}}_k - \boldsymbol{z}_k\|_2^2, \quad \Delta_{L1}^{k} = \|\hat{\boldsymbol{z}}_k - \boldsymbol{z}_k\|_1, \tag{2.5}$$

$$\Delta_{Huber}^{k} = \begin{cases} \frac{1}{2}\Delta_{L2}^{k} & \text{if } \Delta_{L2}^{k} < \delta^2 \\ \delta \cdot \left(\Delta_{L1}^{k} - \delta/2\right) & \text{otherwise.} \end{cases} \tag{2.6}$$

Patch discrimination The main drawback of direct reconstruction for Latent MIM is its inability to explicitly incentivize the model to learn diverse representations across the image. This is unlike in pixel-based MIM where the pixel intensities are guaranteed to vary across the image. To circumvent this limitation, we propose a *patch discrimination* objective, where the model is trained to distinguish between target patches using an InfoNCE loss [4]. Specifically, for each target patch k, the predicted latent

Table 2.2: Loss Impact.

Loss	NN	LP	Ft
MSE	18.8	39.3	79.7
L1	15.2	36.7	81.7
Huber	24.3	46.4	82.1
PatchDisc	**30.0**	53.4	86.0
MAE [6]	**30.0**	**59.9**	**88.3**

$\hat{z}_k$ is contrasted with the latents of all target patches z_l

$$\Delta^k_{PatchDisc} = -\log \frac{\exp(\texttt{sim}(\hat{z}_k, z_k)/\tau)}{\sum_{l \in \mathcal{T}} \exp(\texttt{sim}(\hat{z}_k, z_l)/\tau)}, \quad \texttt{sim}(\hat{z}, z) = \frac{\hat{z}^T z}{\|\hat{z}\|\|z\|} \tag{2.7}$$

where τ is a temperature hyper-parameter.[2] To minimize this loss, the model must not only align the predicted and target latents accurately but also ensure sufficient diversity among the latents within the image.

Study Results We compare each of the aforementioned loss functions for latent reconstruction. Building on the findings of Section 2.3, targets are computed from a momentum encoder. The results shown in Table 2.2 indicate that, although still insufficient for effective representation learning through latent MIM, the patch discrimination loss (PatchDisc) can learn better representations than with direct reconstruction losses. In particular, we highlight the significant improvements in the nearest neighbor and linear probing accuracy, which are more sensitive to the quality of the learned representations.

[2]The patch discrimination loss has no batch dependencies, as the negative samples are derived from the image itself.

Challenge 3: Semantic Correlation between Nearby Patches

Image content displays high correlation within proximate regions. This can render mask reconstruction a trivial task, as the model can interpolate missing information from nearby visible patches. To counter this, pixel-based MIM masks a substantial portion of the image (up to 75%) [6]. Latent representations, which are expected to encapsulate high-level semantics, exhibit even higher correlations across patches compared to their corresponding pixels. This correlation potentially undermines the effectiveness of the task for representation learning.

High mask ratio. Latent MIM also benefits from *high mask ratios*. Beyond reducing patch correlation and enhancing representation learning, this strategy also enables faster training and a lower GPU memory footprint. However, it has to keep enough visible patches to capture critical features within the image.

Non-contiguous grids. To further minimize the correlation between visible and target patches without reducing the amount of visible cues, we experiment with stochastic non-contiguous grids (Fig. 2.4). This strategy increases the distance between patches by separating each patch from its neighbors by a random number of unused pixels. Specifically, let P represent the patch size and G the average gap between consecutive patches. Stochastic non-contiguous grids can be conveniently generated by initially splitting the image into a regular grid of patches, each of size $(P + G) \times (P + G)$, and then extracting a $P \times P$ patch at random from each grid location.

Patch Similarity Constraints While the previous strategies reduce correlation by refining the patch selection process, we can also impose ex-

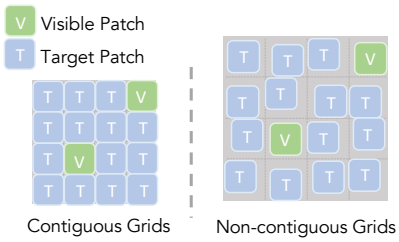


Figure 2.4: Patch Generation Strategies. Left: Masking contiguous grids. Right: Non-contiguous stochastic masking.

Mask	Gap	Sim.	NN	LP	Ft
0.75	0	✗	30.0	53.4	86.0
0.9	0	✗	53.3	70.4	87.6
0.9	4	✗	54.6	71.2	88.0
0.9	4	✓	**56.8**	**72.9**	**88.5**
MAE [6]			30.0	59.9	88.3

Table 2.3: Improvement from reducing semantic correlation mitigation with strategies of high mask ratio, non-contiguous stochastic maskin and patch-wise similarity constraint.

plicit constraints to avoid correlation in the latent space. This is especially important since, as highlighted in the previous challenge (Table 2.2), patch representations tend to cluster together when trained with no constraints. To counteract this, we impose a constraint on both the visible and predicted latents, $Z_{\mathcal{V}}$ and $\hat{Z}_{\mathcal{T}}$,

$$\mathcal{R} = (\gamma - \mathbb{E}_{i,j\in\mathcal{T}}\left[\texttt{sim}(\hat{z}_i, \hat{z}_j)\right])^2 + (\gamma - \mathbb{E}_{i,j\in\mathcal{V}}\left[\texttt{sim}(z_i, z_j)\right])^2, \tag{2.8}$$

where γ is a predefined desired inter-patch similarity.

Study Results Once again, we build on the findings of the previous challenge and use patch discrimination for latent reconstruction. Table 2.3 shows the incremental improvement in representation quality from each different strategy for mitigating semantic correlation. Latent MIM shows substantial improvements from even higher mask ratios than pixel-based MIM, with the optimal ratio being 90%. Exceeding 90% masking is counterproductive, as not enough visible patches would be available to capture critical features within the image. Using non-contiguous stochastic masking with gap=4 can further improve the representation quality by lowering the spatial redundancy while keeping enough visible information.

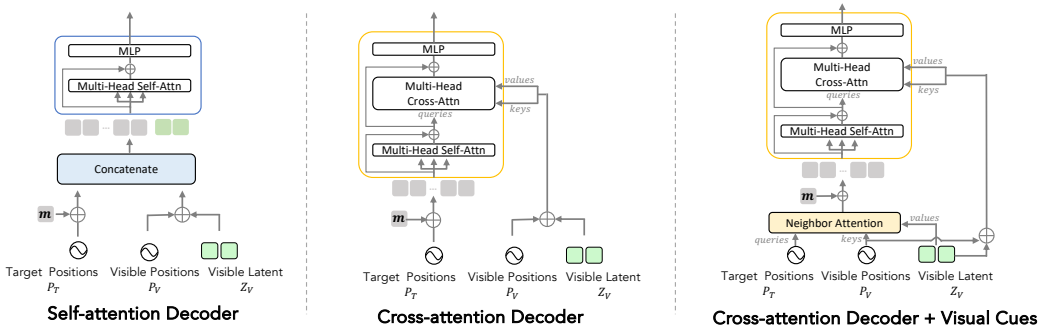


Figure 2.5: **Comparison of Three Different Decoder Designs.** Self-attention decoder is commonly used for low-level MIM models. Cross-attention decoder provides direct conditioning at each layer on the visible latents.

Finally, adding a regularizer that explicitly constrains the mean similarity among patches also enhances performance.

Challenge 4: Decoder Design for Latent Reconstruction

The final important component of Latent MIM is the decoder, responsible for predicting target representations from visible patches. The decoder design is crucial for the model's ability to effectively utilize the high-level semantics extracted from the encoder. We explore the impact of different decoder architectures.

Self-attention decoder Pixel-based MIM models employ a self-attention transformer to predict the target pixels $X_{\mathcal{T}}$. To accomplish this, the decoder receives two sets of inputs: the latents of visible patches $Z_{\mathcal{V}}$ and a sequence of learnable mask tokens $\mathbf{m}$, marked by the fixed SinCos positional embeddings $\mathbf{p}$ of their corresponding locations (*i.e.*, $\mathbf{m} + \mathbf{p}_t \; \forall t \in \mathcal{T}$ and $\mathbf{z}_v + \mathbf{p}_v \; \forall v \in \mathcal{V}$). After processing this sequence through a series of self-attention blocks, the decoder outputs the target representations using a linear head.

Cross-attention decoder Unlike pixel-based MIM, in Latent MIM the targets $Z_{\mathcal{T}}$ are representations with a similar level of abstraction than the features obtained from the encoder $Z_{\mathcal{V}}$. Thus, the decoder should be able to condition on these visible representations $Z_{\mathcal{V}}$ more directly. While self-attention only conditions once $Z_{\mathcal{T}}$ through the input sequence, cross-attention allows the decoder to condition on $Z_{\mathcal{V}}$ at every layer. A standard cross-attention architecture [10] with alternating self-attention, cross-attention, and feed-forward MLP blocks is used to update the prediction tokens $\mathbf{m} + \mathbf{p}_t \; \forall t \in \mathcal{T}$.

Visual cues from neighboring visible patches As discussed in Section 2.3, neighboring patch latents can be highly correlated. To prevent the decoder from focusing excessively on interpolating between these patches, we embed visual cues directly into its input sequence. This allows the decoder to better focus on more fine-grained spatial reasoning. Specifically, with $P_{\mathcal{V}}$ and $P_{\mathcal{T}}$ as the positional embeddings for the visible and target patches, respectively, we initialize the prediction tokens as $\mathbf{m}_i = \mathbf{m} + \mathbf{p}_t + \mathit{Softmax}_t\left(P_{\mathcal{T}} P_{\mathcal{V}}^{T}\right) Z_{\mathcal{V}}$ at each target location $i \in \mathcal{T}$. This setup equips the mask tokens $M_{\mathcal{T}} = \{\mathbf{m}_i\}_{i \in \mathcal{T}}$ with a weighted blend of latents from the nearest visible patches, providing precise location and visual cues right from the start.

Latent Projector Following contrastive learning methods, we experiment with non-linear projection heads $h(\cdot)$ to prepare the latents from visible patches $Z_{\mathcal{V}}$ for the decoder. Through a series of experiments with various non-linear projectors h, we found that a simple multi-layer perceptron (MLP) with three layers, GELU activations, and layer normalization yielded the most effective representations.

Study Results Table 2.4 assesses the impact of different decoder designs on the quality of the learned representations. Pixel-based MIM models,

Table 2.4: Comparison between different decoder designs. Other configurations follow the optimal settings found in previous sections.

Decoder	Proj.	Depth	NN	LP	Ft
Self-attn	none	3	56.8	72.9	88.5
Self-attn	mlp	3	59.7	76.8	89.4
Cross-attn	mlp	3	61.1	77.0	89.4
Cross-attn w/ Vis Cues	mlp	3	**63.3**	**77.3**	**89.9**
Cross-attn w/ Vis Cues	mlp	8	35.9	65.8	86.2
MAE [6]		8	30.0	59.9	88.3

such as MAE, employ a lightweight self-attention decoder with a depth of 8 layers. Our findings underscore the necessity for a distinct decoder design for Latent MIM, given the different nature of the decoder's task. Specifically, we found that cross-attention provides a better mechanism for conditioning the reconstruction on the visible patches, leading to superior performance in all downstream tasks, especially when initializing the prediction tokens with visual cues from neighboring visible patches and when combined with a non-linear projection head to process the input visible patches. Furthermore, since both the inputs and targets are high-level, the Latent MIM decoder can be remarkably lightweight. Optimal performance is achieved using a shallow (3-layer) transformer, considerably smaller than the 8-layer MAE decoder. When combined, these strategies allow the decoder to better utilize the high-level semantic features extracted from the encoder, surpassing the performance of pixel-based MIM models in all downstream tasks.

2.4 Scaling to ImageNet-1k

The previous section provided a detailed analysis of the optimization challenges and design decisions for Latent MIM on a medium-sized dataset. In this section, we show the scalability of Latent MIM to larger datasets, specifically ImageNet-1k, and compare it to prior work. We also highlight the strongly localized semantics learned by Latent MIM

Table 2.5: Top-1 NN and LP classification accuracy on ImageNet.

Method	Epochs	NN	LP
Low-level			
MAE [6]	1600	12.2	67.8
SimMIM [31]	800	-	56.7
Latent			
ConMIM [29]	800	-	39.3
data2vec [14]	800	25.7	60.3
Latent MIM	800	**50.1**	**72.0**

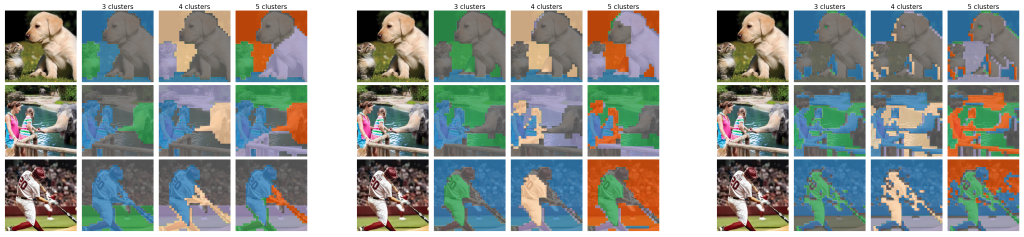
by evaluating the trained model on unsupervised scene segmentation, video object segmentation, and few-shot transfer learning.

Implementation We scaled up pretraining to ImageNet-1k using the optimal Latent MIM configuration from Section 2.3. The model is pretrained for 800 epochs using the Adam optimizer.

ImageNet-1k Classification

Following [6], we evaluate the learned representations on ImageNet-1K using Nearest Neighbor (NN) and linear probing (LP) protocols. Table 2.5 compares our Latent MIM with low-level MIM methods and prior latent MIM methods with ViT-B/16. For MAE and data2vec, we use the features (either cls-token or average-pooling) that yield the best performance.

Latent MIM surpasses low-level MIM methods by large margins on NN classification (+37.2% on MAE), and LP (+3.3% on MAE and +14.4% on SimMIM), demonstrating the effectiveness of Latent MIM for learning improved semantics from latent masked reconstruction. Our method also outperforms related prior work, data2vec and ConMIM, by large margins.



(a) Latent MIM (b) data2vec (c) MAE

Figure 2.6: **Unsupervised segmentation maps** by hierarchical clustering of patch representations within each image. Better viewed digitally with zoom.

Properties beyond Classification

The Latent MIM framework is designed to learn localizable and semantically rich representations. We showcase these properties on unsupervised segmentation, semi-supervised video object segmentation, and transfer learning.

Unsupervised segmentation An emerging property of Latent MIM is its capacity for semantic clustering of local representations, which enables impressive segmentation and scene parsing outcomes without the need for supervised fine-tuning. Fig. 2.6 illustrates the unsupervised segmentation maps generated by hierarchical clustering of patch-level representations. Compared to both lower-level MIM approaches, such as MAE, and earlier latent MIM methods like data2vec, our Latent MIM model learns better semantic and localizable representations.

Video Object Segmentation Our Latent MIM model can also maintain both semantic integrity and localization accuracy, even in complex, dynamic video sequences. DAVIS-2017 semi-supervised video object segmentation benchmark evaluates the ability to generate precise object segmentation masks in videos, starting from ground truth masks of the ini-

Method	epochs	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
MAE	1600	57.5	54.8	60.2
data2vec	800	28.5	27.8	29.2
Latent MIM	800	**65.5**	**63.1**	**68.0**

Table 2.6: **Video object segmentation** on DAVIS-2017. $\mathcal{J}$ and $\mathcal{F}$ quantify region similarity and boundary alignment, respectively.

Figure 2.7: **Visualization**. Col 1: ground truth; Cols 2–4: predictions at 25%, 50%, and 100% timesteps.

Table 2.7: **Few-shot learning on various datasets.** We report the mean and standard deviation of the top-1 accuracy across 3 different runs. Each run is trained on a set of 16 training images per class.

Method	Caltech101 [26]	DTD [20]	Oxford Flowers [21]	Oxford Pets [22]	Stanford Cars [19]	SUN397 [23]	UCF101 [24]
MAE	80.5± 0.4	51.8± 2.2	62.7± 1.8	60.2± 3.6	7.8± 1.0	21.9± 0.4	53.2± 0.4
data2vec	76.6± 2.1	52.0± 2.1	74.1± 2.2	58.5± 2.0	8.8± 0.4	29.0± 0.5	52.6± 1.2
Latent MIM	**89.2**± 1.0	**55.9**± 2.0	**84.0**± 1.0	**79.8**± 2.6	**16.3**± 2.0	**48.4**± 0.6	**75.8**± 2.8

tial frame. We follow the experimental protocol in [30], which segments scenes through a nearest-neighbor strategy applied between consecutive frames. Crucially, this benchmark tests the robustness and adaptability of the pre-trained features without the need for additional training. Table 2.6 shows that Latent MIM surpassed both pixel-level and previous Latent MIM models in performance. Fig. 2.7 shows sample qualitative results.

Few-shot transfer learning To assess generalization beyond ImageNet-1k, we transfer the learned representations to a variety of datasets. Following the setting of [17], we perform 16-shot transfer learning experiments by training a linear classifier on top of frozen representations. All models are trained using SGD with a momentum of 0.9, a learning rate of 0.1,

a batch size of 128, and updated for a total of 2500 iterations. Table 2.7 demonstrates the superior performance of Latent MIM over MAE and data2vec across all datasets, highlighting its robustness, versatility, and potential for various recognition tasks.

2.5 Conclusion

We identified and addressed the key training challenges in Latent MIM, demonstrating its capacity to generate spatially diverse, high-level semantic representations. This is evidenced by significant improvements in nearest neighbor and linear probe evaluation on ImageNet, fewshot transfer learning, as well as in segmentation tasks requiring minimal or no supervision. We hope this chapter will inspire further exploration into Latent MIM for learning fine-grained semantics without human supervision.

3 TRACKVERSE: A LARGE-SCALE DATASET OF OBJECT TRACKS FOR VISUAL REPRESENTATION LEARNING

Video data inherently captures rich, dynamic contexts that reveal objects in varying poses, interactions, and state transitions, offering rich potential for unsupervised visual representation learning. However, existing natural video datasets are not well-suited for effective object representation learning due to their lack of object-centricity and class diversity. To address these challenges, we introduce TrackVerse, a novel large-scale video dataset for learning object representations. TrackVerse features diverse, common objects tracked over time, capturing their evolving states. To leverage temporal dynamics in TrackVerse, we extend contrastive learning with a variance-aware predictor that conditions on data augmentations, enabling models to learn state-aware representations. Extensive experiments demonstrate that representations learned from TrackVerse with variance-aware contrastive learning significantly outperform those from non-object-centric natural video and static image datasets across multiple downstream tasks including object/attribute recognition, action recognition and video instance segmentation, highlighting the rich semantic and state content in TrackVerse features.

3.1 Introduction

Video data, with its dynamic and rich contextual cues, offers a unique perspective into the world of everyday objects, capturing their ever-changing poses, interactions, and changes of state. This source of

information holds immense potential for enhancing unsupervised visual representation learning. However, a key question arises: do our datasets showcase these rich contexts and can our methods effectively leverage them to better understand objects in the real world? This work delves into this question, proposing a new video dataset and exploring the potential of video data for unsupervised visual representation learning of everyday objects.

Despite its potential, natural video datasets [55, 56, 58, 45] are not well-suited for effective object representation learning. Why? First, natural video is not object-centric, and thus crops of a video can depict unrelated visual signals. Successful representation learning methods of everyday objects have been developed primarily leveraging static, object-centric datasets like ImageNet [40]. These methods [2] fall short when applied to natural video data, failing to accurately represent individual objects. While unsupervised methods designed for simultaneous localization and representation could potentially be developed, they [58, 56] are often tailored to specific datasets and do not generalize to different video sources. Second, natural videos exhibit a long-tail distribution of object categories, resulting in limited diversity within finite datasets. This imbalance negatively impacts representation learning, as models are less exposed to a broad and balanced set of objects, ultimately limiting their generalization ability.

In this chapter, we introduce a new video dataset, TrackVerse, specifically designed for representation learning of objects in their natural environments, featuring a *diverse set of common objects tracked over time*. TrackVerse mitigates the challenges of natural video by providing a balanced, object-centric dataset that captures the evolution of object states as they interact with the environment. Since manually localizing and tracking all objects in a video would be prohibitively expensive, we introduce an automated data collection pipeline that leverages an

Table 3.1: **`TrackVerse` vs. Existing Video Datasets.** Full Dyn.: full dynamics that supports random frame sampling at any time point. The R2V2 dataset only contains 4 fixed frames per video. Obj Centric: object-centric dataset. Class Bal: Class Balance where the high-frequency classes were capped. Dur: total duration of the dataset.

Dataset	Domain	Full Dyn.	Obj. Centric	Class Bal.	Open Source	# Clips	Dur (HR)
MOT20 [63]	Pedestrian	✓	✓	✗	✓	2,332	14.8
VOTS2023 [79]	Free-form	✓	✓	✗	✓	341	6.5
TAO [70]	LVIS(488), free-form(345)	✓	✓	✗	✓	1,787	46.6
R2V2 [55]	ImageNet	✗	✗	✓	✓	696K	—
Ego4D [57]	Human activities	✓	✗	✗	✓	923	3,670
Walking Tours [58]	Urban	✓	✗	✗	✓	10	23
VideoNet [56]	ImageNet	✓	✗	✓	✗	1.2M	3,055
Full TrackVerse (Ours)	LVIS Objects	✓	✓	✗	✓	31.9M	45,582
Curated TrackVerse (Ours)	LVIS Objects	✓	✓	✓	✓	3.8M	5,328

open-vocabulary object detector and tracker to create the dataset.

Through a series of experiments, we demonstrate the value of TrackVerse for unsupervised visual representation learning, and in particular contrastive learning. We show that the learned representations outperform those learned from non-object-centric video datasets, and those from equivalent static image datasets across a range of downstream tasks. We also demonstrate that to exploit the full potential of object tracks for representation learning, contrastive methods should relax the invariance objective to enable models to be sensitive to semantically distinct views of an object, such as the different states an object might go through over time. By learning representations that are invariant to semantic-preserving transformations (*e.g.*, color augmentations) while being predictable to semantic-changing transformations (*e.g.*, spatio-temporal augmentations), the proposed extension enhances downstream task performance, outperforming similar contrastive methods that rely on invariance objectives alone.

3.2 Related Work

Self-supervised object representation learning Self-supervised learning (SSL) aims to learn general object representations that can be effectively transferred to various downstream tasks without requiring human annotations. Existing SSL methods for visual representation learning can be broadly categorized into two frameworks: (1) joint embedding learning, which encodes different views of a sample into aligned representations [34, 35, 36, 37, 38, 3], and (2) masked image modeling (MIM), which learns by reconstructing masked regions of the input data [6, 15, 28, 14, 39]. While these approaches demonstrate remarkable success on ImageNet-1K, their application to natural videos presents significant challenges. One key challenge is that natural videos exhibit dynamic scenes with multiple objects and complex backgrounds. Therefore, to better study and evaluate SSL methods for learning object representations from videos, a large-scale, object-centric video dataset with diverse objects and rich motion patterns is needed.

Video datasets Numerous video datasets have been developed over the years, driven by progress in video tasks such as action recognition [59, 60, 24, 61, 62], object tracking [63, 79, 70], video object classification [45] and segmentation [64]. However, these datasets are often confined to specific domains such as human actions [59, 60, 24, 61], egocentric views [62, 65] or urban scenes [58], lacking the broad diversity needed for effective self-supervised learning (SSL).

A few attempts have been made to construct large-scale video datasets specifically for SSL (R2V2 [55], VideoNet [56], Walking Tours [58]). However, these datasets have their own limitations as well. Although they display a wide range of scenes and objects, the distribution remains long-tailed, predominantly featuring human-centric activities [57] or urban environments [58]. Furthermore, although class balanced, R2V2 [55] and

VideoNet [56] use YouTube Search and classification-based filtering to construct the dataset. As a result, they fail to guarantee that each frame displays the intended object without major distracting elements. This lack of object-centricity adds a significant burden to the SSL process, as the model must learn to focus on the relevant object amidst the clutter.

Learning object representations from videos Video representation learning has evolved significantly over the years, with early approaches focusing on fundamental visual cues like temporal coherence [71, 72] or object tracking [73, 42, 41]. Recent advances have predominantly leveraged modern self-supervised techniques, including contrastive learning and masked autoencoding, to learn robust spatio-temporal features [43, 75, 49, 50, 74, 81, 82, 83, 84, 85, 86, 87]. While these methods have shown promising results, they typically rely on specialized video architectures and primarily transfer to video-specific tasks such as action recognition [83, 86, 87], without explicitly considering object-centric representations. These methods struggle to learn object-centric representations due to the nature of natural video data, typically depicting complex scenes with a long-tail distribution of object categories. Some methods have explored cycle-consistency [30, 76, 42] and optical flow [77, 78] to establish correspondences between local (potentially object related) regions across time. However, they still struggle with the trade-off between temporal understanding and spatial scene comprehension since simultaneously localizing, tracking and learning representations of objects as they move and change their state and appearance is challenging.

3.3 TrackVerse Dataset

We constructed a video dataset, denoted TrackVerse, where a diverse set of objects are tracked over time, capturing the evolution of their state as

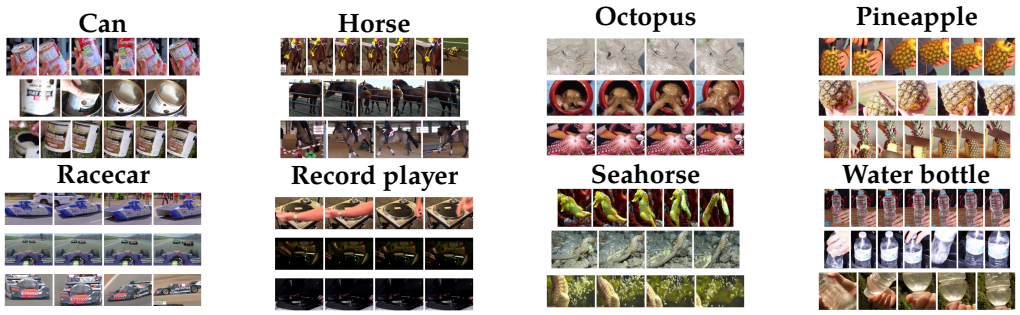


Figure 3.1: **Examples from `TrackVerse`** TRACKVERSE comprises object-centric video tracks with a diverse range of objects.

they interact with the environment. Since manually localizing and tracking all objects in a video is too costly, we leverage an open-vocabulary object detector and tracker to automate the process. In addition to the large-scale dataset of object tracks generated by our automated collection pipeline, we manually verified the tracking accuracy for a subset of the data and manually annotate the categories of the depicted objects for downstream evaluation. The proposed dataset is designed to provide richer sources of context, beyond static object appearance, where the principles of unsupervised representation learning from object dynamics can be investigated. Examples from TRACKVERSE can be seen in Fig. 3.1. The dataset will be publicly released.

Dataset acquisition and tracking pipeline

To build the TRACKVERSE dataset, we developed a data collection pipeline illustrated in Fig. 3.2.

Step 1. Data source: We downloaded 600,000 YouTube videos randomly sampled from the HD-VILA-100M [44] dataset, each acquired at 720p resolution and original frame rate.

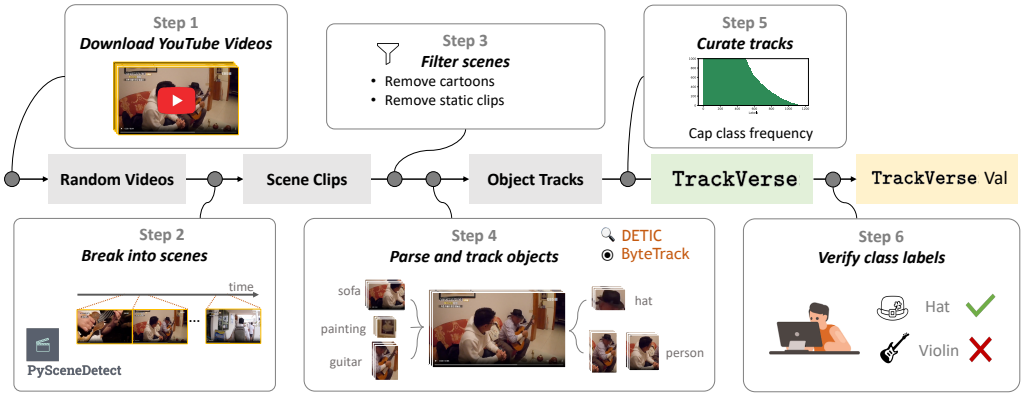


Figure 3.2: **Overview of `TrackVerse` Collection Pipeline.**

Step 2. Scene segmentation: Each video was segmented into scenes using PySceneDetect [69] and split into video segments with consistent visual content.

Step 3. Filtering: We filtered out segments with predominantly static visual content as well as segments identified as cartoons. Cartoon detection was performed using an ImageNet-pretrained ResNet18 model, fine-tuned for cartoon recognition. For static content, we calculate the average pixel change rate between consecutive frames to identify minimal motion.

Step 4. Object detection and tracking: We detected and tracked objects in the remaining segments using the DETIC object detector [66] and the ByteTrack tracker [67]. DETIC performs open-vocabulary object detection using a SWIN transformer aligned with CLIP representations. To ensure diversity of object categories, we deployed DETIC using the LVIS ontology [68] comprised of 1203 classes. We then merge DETIC's detections into object tracks. ByteTrack is a tracking-by-detection algorithm robust to occlusions or failures of the underlying detector by combining both high and low-confidence detections, together with a Kalman filter

of object motion and appearance (as represented by DETIC's ROI pooling features).

Step 5. Data curation: To increase sample diversity within a finite dataset, we curated the dataset to balance the distribution of object categories, by subsampling high-frequency classes. We capped the number of tracks for high-frequency classes at 1,000, 2,500, 4,000 and 8,000 tracks per class, selecting tracks with the highest object detection scores. This yielded curated subsets of 0.4M, 1.1M, 2.4M, and 3.8M tracks, respectively. These class-balanced subsets of varying scales enabled us to investigate the impact of dataset size on the learned representations. For large-scale experiments, we use the 3.8M-CB8000 subset, the largest class-balanced subset, which was shown to be the most effective.

Step 6. Test set: Finally, we build a test set of TrackVerse with human-verified labels as a few-shot video classification benchmark. The verification process was conducted through manual visual inspection by the authors. For each class, one annotator selected the tracks from a pool of candidates randomly sampled from DETIC predictions, ensuring accurate class representation while excluding near-duplicates and visually similar tracks. To ensure the quality, we also run a multi-annotator validation on a smaller subset of tracks, achieving 100% agreement among all annotators in their independent evaluations. The resulting test set contains 4,188 object tracks spanning 698 distinct categories with 6 tracks per category. To prevent data leakage, videos containing these tracks were rigorously excluded from the TrackVerse dataset.

Dataset Characteristics

Table 3.1 shows an overview of the collected TrackVerse dataset in comparison to related databases. Our dataset distinguishes itself by being

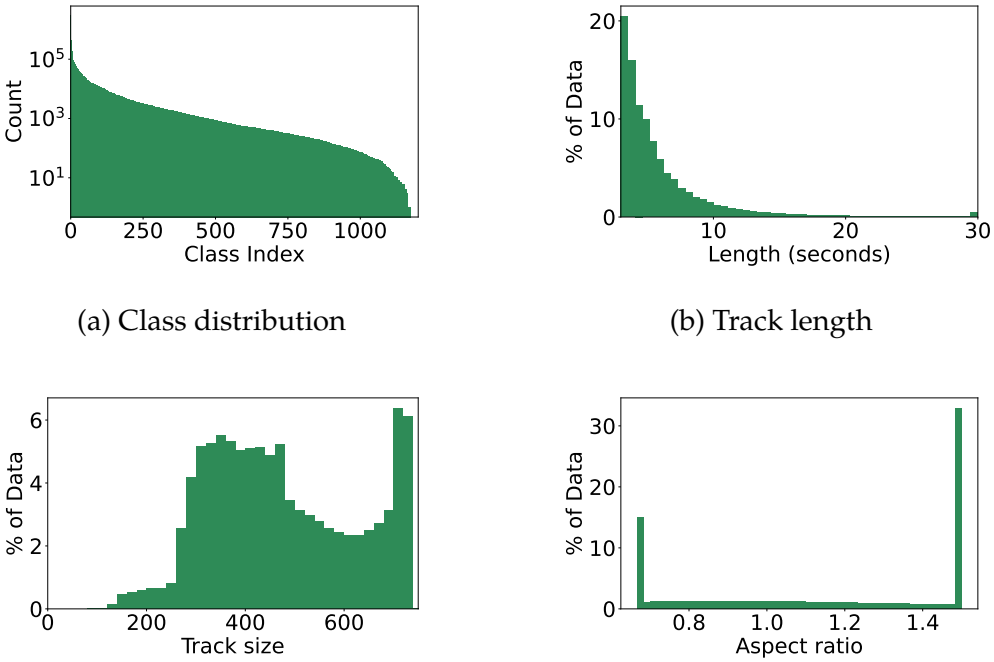
the only *large-scale object-centric video dataset*. The most related dataset is VideoNet, a dataset compiled by searching YouTube for videos containing objects from the ImageNet dataset. However, VideoNet does not locate and track objects over time, each clip can depict cluttered scenes (*i.e.*, not fully object centered). Perhaps more importantly, VideoNet has not been open-sourced and thus is not available for research. Other datasets, such as MOT20, VOTS2023 and TAO, do provide object track annotations, but are too small for representation learning purposes. In contrast, TrackVerse provides over 2000× more samples than the largest of the tracking datasets (TAO), and over 100× more content (in hours).

Since the data collection pipeline is fully automated, TrackVerse can be easily scaled. In the current work, we applied the tracking pipeline to 600,000 YouTube videos, yielding an unbalanced set of 31.9 million object tracks (45,582 hours) distributed across the 1203 LVIS classes. As can be seen in Fig. 3.3a, the distribution of object categories is long-tailed, with the most common category, "person", accounting for 31.9% of all tracks and being 261 times more prevalent than the 100th most common category, "computer keyboard". The dataset H-index is 578 (*i.e.*, there are 578 classes with at least 578 samples). Fig. 3.3b–3.3d further show the distribution of track lengths, track sizes (*i.e.*, the shorter side length of the object's bounding box), and aspect ratios among TrackVerse samples.

To avoid short tracks which may not provide enough variation in object appearance, we discarded tracks with less than 3 seconds. We also forced the aspect ratio of each track bounding box to be between 2:3 and 3:2 to avoid overly elongated tracks.

Pipeline optimization and validation

To ensure accurate tracking, we optimized the DETIC+ByteTrack pipeline on the DAVIS [64] dataset for tracking performance and on ImageNet [40] for classification performance. After optimization, our tracker obtains



(a) Class distribution

(b) Track length

(c) Track size

(d) Aspect ratio

Figure 3.3: **Characteristics of the Full `TrackVerse`.** (a) The number of videos per class, illustrates the long-tail distribution of the dataset; (b) Distribution of object track durations; (c) Distribution of track size where size is the short side of the object's bounding box; (d) Distribution of aspect ratios of all object tracks.

an IDF1 score of 74.7 and 62.6% MOTA score on the DAVIS dataset and achieved a classification accuracy of 62.7% on the ImageNet-1k dataset.

We also manually assessed the quality of DETIC's assigned labels on TrackVerse clips by inspecting a random sample of 1000 tracks from each subset. DETIC's precision on the full TrackVerse and the largest class-balanced subset (TrackVerse-3.8M-CB8000) were 78.7% and 52.3%, respectively. The higher precision in the full TrackVerse is due to its long-tailed distribution and DETIC's improved reliability on common objects. Since the category itself is not used in self-supervised learning, misclas-

sifications do not introduce faulty supervision into training. On the contrary, the use of an automated pipeline (even though not perfect) allows scaling up the data available for self-supervised training both in terms of number of samples and diversity which are critical for effective representation learning.

3.4 Pretraining Protocol

The TrackVerse dataset was built for self-supervised learning of object representations, so as to learn representations that are descriptive of not only their category but also of their state and dynamics. To demonstrate its value, we assess the representations learned from TrackVerse using contrastive learning for pretraining. Contrastive learning has been shown effective for a variety of data modalities from static images of objects [34, 35, 36, 37, 1, 38], to videos of human actions [83, 86, 87], or multimodal data including vision-language data [47, 48] and audio-visual data [46, 88, 89, 74].

However, despite its demonstrated success, contrastive learning has limitations when applied to object-centric video. The main advantages of learning from this type of data are two-fold: (1) the natural data augmentations caused by camera and object motion, which enables learning invariances to pose, view-point and lighting conditions, and (2) the natural interactions between objects and the consequent changes in object states, which enables learning representations that further capture object states and their dynamics. Contrastive learning is designed to learn invariances, and thus is well-suited to capture the former. However, enforcing view invariance discourages the model from learning the semantic differences across views, such as the changing states of objects over time. To effectively utilize TrackVerse, we extend contrastive learning so as to capture both invariance to appearance-preserving augmentations as well as the

variations from the natural dynamics of object states across time. We refer to this as *variance-aware contrastive learning*.

Background: Contrastive Learning

Contrastive learning [34, 35, 36, 37] learns to encode different views of a sample into aligned representations. These methods generally adhere to the following procedure. First, two (base and target) views $x_b = \mathcal{T}(x; v_b)$ and $x_t = \mathcal{T}(x; v_t)$ of a sample x are computed from $\mathcal{T}(\cdot; v)$ — an augmentation procedure parameterized by a random variable v. We refer to x_b, x_t as augmentations of sample x, and refer to v_b, v_t as augmentation parameters. The two views are then encoded using two (base and target) neural networks, f_b and f_t, into vectorized representations $z_b = f_b(x_b)$ and $z_t = f_t(x_t)$. Finally, the base representation z_b is fed to a predictor $h_{b\to t}$ whose output $\hat{z}_t = h_{b\to t}(z_b)$ is required to be aligned with the target representation z_t. Contrastive methods often accomplish this by minimizing the InfoNCE loss [4]

$$\mathcal{L} = -\log \frac{e^{\mathtt{sim}(\hat{z}_t, z_t)/\tau}}{e^{\mathtt{sim}(\hat{z}_t, z_t)/\tau} + \sum_n e^{\mathtt{sim}(\hat{z}_t, z_n)/\tau}} \tag{3.1}$$

where $\mathtt{sim}(\cdot, \cdot)$ is the cosine similarity and τ a temperature coefficient. Following [37], the base and target encoders, f_b and f_t, have the same architecture composed of a backbone model and a non-linear projection. Predictor $h_{b\to t}$ is a multi-layer perceptron (MLP). While f_b and $h_{b\to t}$ are updated by gradient descent, f_t is updated by an exponential moving average of f_b. After pretraining, only the base encoder f_b is used to extract features for downstream tasks.

Variance-Aware Contrastive Learning

Since view-invariance requires $h(z_b) = z_t$ for all possible views, the representations z_b and z_t are encouraged to be identical (up to an involutory transformation h). Since invariance can be too restrictive for object-centric video data, we relax this requirement by conditioning the predictor $h_{b\to t}(\cdot)$ on the augmentation parameters ν_b and ν_t. This enables the model to leverage the contextual relationship between the two views for the prediction of $\hat{z}_t$. To this end, we encode the augmentation parameters ν_b and ν_t using an MLP f_ν, *i.e.*, $\boldsymbol{\nu}_b = f_\nu(\nu_b)$ and $\boldsymbol{\nu}_t = f_\nu(\nu_t)$, and use the embeddings to condition the predictor $\hat{z}_t = h_{b\to t}(z_b, \boldsymbol{\nu}_b, \boldsymbol{\nu}_t)$. Conditioning is done by simple concatenation of all predictor inputs.

Data augmentation is deployed to generate diverse views of an object from each track. To leverage natural object motion, we perform *temporal* jittering, where the two views x_b and x_t are sampled at random timesteps t_b and t_t, separated by a time gap $\delta = t_b - t_t$. We also apply spatial and color augmentations following prior work [35, 36, 3], including random resized cropping, horizontal flipping, random Gaussian blur, color jittering, grayscale conversion, and solarization. One characteristic of TrackVerse is that the typical size/resolution of a track differs significantly across object categories, and thus their natural resolution can provide a cue for the model to "cheat". To mitigate this issue, we use adaptive Gaussian Blur, where the blurriness level is controlled by the crop resolution. Low-resolution crops are blurred less than high-resolution crops.

Augmentation parameters Semantics-altering transformations include spatial cropping, which can yield views of different object subparts, and temporal jittering, which may modify the object's state. Conversely, color augmentations do not modify the semantics of the view, so the represen-

tations should ideally exhibit invariance to them. Therefore, we define the augmentation parameter v to describe only the spatio-temporal location, *i.e.*, $v = (t, \frac{x_c}{W}, \frac{y_c}{H}, \frac{w}{W}, \frac{h}{H})$, where t denotes the timestep, (x_c, y_c) the center coordinates of the crop, $w \times h$ its width and height, and $W \times H$ the original frame size.

Implementation Details

Modeling Following MoCo-v3 [2], we adopt ViT-Base/16 as the base encoder, followed by a three-layer MLP projector with 256 output dimension. The target encoder has the same architecture but its weights are updated by the exponential moving average of the base encoder. To better leverage video data as described in Section 3.4, we employ adaptive Gaussian blur (with maximum blurriness of 1) for spatial augmentations and apply temporal jittering (with a time gap of $\delta = 2$). The augmentation parameters $\boldsymbol{v}$ are defined as the addition of spatial and temporal embeddings of each view. They are encoded into a 256-dim embedding through a 2-layer MLP projector with 4096 hidden units. The spatio-temporal augmentation embeddings are then used to condition the contrastive learning process. It is worth noting that the variance-aware adaptations account for only 0.9% additional parameters compared to standard view-invariance MoCo, ensuring that performance gains are attributable primarily to the dataset's characteristics rather than model capacity.

Optimization We pre-train the models using the AdamW [33] optimizer with a batch size of 1024, a base learning rate of 1.5×10^{-4}.

3.5 Downstream Evaluations

To assess the quality and generalizability of the representational features extracted from our dataset, we conduct comprehensive transfer learning

Table 3.2: **Evaluation of representations across diverse downstream tasks**: LVIS-IN and TrackVerse object classification, SSv2 action recognition, MIT-States object/attribute classification, and DAVIS-2017 video object segmentation. Features were pre-trained using MoCo on ImageNet and variance-aware MoCo on videos, both with equivalent training budget of 220M forward-backward passes.

Pretrain Dataset	LVIS-IN	TrackVerse		SSv2	MIT-States				DAVIS-2017		
	NN	NN	FSL	Ft	Obj NN	Attr NN	Pair NN	Pair Ft.	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Image dataset for MoCo pretraining:											
ImageNet-1K [40]	48.8	32.0	47.7	56.8	55.8	38.9	11.1	46.8	64.8	62.1	67.4
Video datasets for augmentation-aware MoCo pretraining:											
TAO [70]	14.8	13.3	19.8	46.7	45.7	38.6	5.8	25.9	60.6	57.9	63.3
Walking Tours [58]	17.0	12.5	16.3	47.0	38.4	37.6	6.3	26.1	62.7	60.2	65.2
Scene Clips	35.3	29.3	41.5	57.0	52.4	39.1	10.6	40.2	66.3	63.4	69.2
TrackVerse	**50.1**	**56.2**	**71.1**	**61.8**	**61.0**	**45.3**	**13.7**	**50.0**	**67.1**	**64.1**	**70.1**

experiments across diverse evaluation tasks. These experiments aim to validate whether features learned from TrackVerse capture more comprehensive object representations compared to existing datasets, particularly in their ability to encapsulate both high-level semantic information and dynamic state changes.

In this section, we first describe the evaluation tasks and their training protocols. We then show that our dataset provides unique value for object representation pretraining by comparing features extracted from TrackVerse with those extracted from existing datasets. Finally, we investigate the scaling properties of TrackVerse, examining how performance evolves as the dataset size and training budget increase.

Downstream Tasks and Evaluation Protocols

We evaluate the learned representation across various tasks, categorized into two groups based on their requirements: tasks focused on object-level semantics and those requiring state-aware semantics. Evaluations such as nearest neighbor (NN) and few-shot learning (FSL) on LVIS-IN,

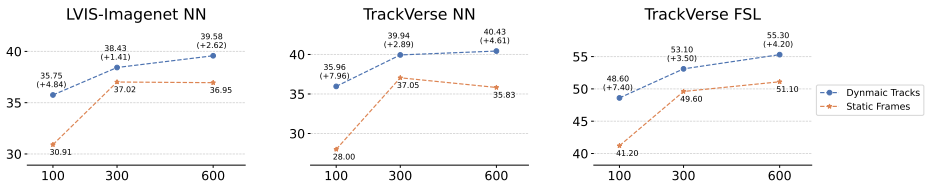


Figure 3.4: **TrackVerse's Scalability to Training Budget.** Compared to SSL pretraining on static images, TrackVerse dynamic tracks prevent overfitting and enhance representation learning when using longer training schedules.

TrackVerse, and MIT-States [53] primarily assess the understanding of object semantics. In contrast, action recognition on Something-Something v.2 (SSv2) [51], and attribute classification on MIT-States, demand that features encapsulate both object semantics and state-aware representation.

Nearest neighbor recognition We evaluate image recognition on the TrackVerse evaluation set and LVIS-IN test set using a nearest-neighbor (NN) protocol. The LVIS-IN test set is a subset of the ImageNet-1k evaluation subset, including 266 classes that overlap with the LVIS vocabulary. NN classification relies solely on the learned representations, without additional tuning, making it a strong indicator of the semantic quality captured during pre-training.

Few-shot learning We perform few-shot learning (FSL) experiments on TrackVerse evaluation set by training a linear classifier atop the frozen representations, following the setting of [17]. We repeat this evaluation on 6 different sets of 5 training images per class and average the final performance.

SSv2 The Something-Something dataset [51] is an egocentric action recognition dataset with 174 classes describing actions independent of

objects (e.g., 'throwing something,' 'putting something into something'). This task emphasizes temporal cues and state transitions. All experiments used the TimeSformer architecture [52], a video Transformer with self-attention across patches over time. Spatial parameters were initialized from the pre-trained model, while temporal parameters were trained from scratch. We report top-1 accuracy of an ensemble over 15 inputs generated for 5 different frame selections and 3 distinct cropping regions.

MIT-States The MIT-States dataset [53] contains 63,440 images labeled with objects (e.g., 'apple,' 'tomato') and attributes (e.g., 'ripe,' 'cooked'), enabling assessment of object-level semantics and fine-grained attributes. We evaluate our pre-trained model on three tasks using nearest neighbor: (1) object recognition, (2) attribute recognition with a known object, and (3) object-attribute pair recognition, using balanced subsets with 50 images per object, 5 per attribute, and 5 per object-attribute pair. We also perform fine-tuning with the CLIP-Adapter framework [54], initializing the image encoder with our pre-trained model. Fine-tuning results are reported as Harmonic Mean (HM) ($HM = 2 * \frac{S*U}{S+U}$, where S and U are the accuracies on seen and unseen examples, respectively).

DAVIS-2017 The DAVIS-2017 [80] semi-supervised video object segmentation benchmark evaluates the ability to generate precise object segmentation masks in videos, given the ground truth mask of the initial frame. We follow the experimental protocol in [30], which segments scenes using a nearest-neighbor strategy applied between consecutive frames. This benchmark assesses both the semantic integrity of the pre-trained features and their sensitivity to object state changes.

Table 3.3: **Scalability to Dataset Size.** Increasing the dataset size of TrackVerse significantly improves downstream performance.

Subset	# Tracks	Max Tracks per Class	LVIS-IN	TrackVerse		SSv2	MIT-States				DAVIS-2017		
			NN	NN	FSL	Ft	Obj NN	Attr NN	Pair NN	Pair Ft.	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
0.4M-CB1000	0.4M	1000	39.6	40.3	55.3	59.3	55.1	44.2	11.8	40.8	65.4	62.5	68.3
1.1M-CB2500	1.1M	2500	46.8	51.8	66.0	59.6	58.5	44.7	12.7	47.1	65.8	62.7	68.8
2.4M-CB4000	2.4M	4000	48.6	54.9	69.1	60.5	59.9	45.1	13.1	**50.0**	66.8	63.8	70.0
3.8M-CB8000	3.8M	8000	**50.1**	**56.2**	**71.1**	**61.8**	**61.0**	**45.3**	**13.7**	**50.0**	**67.1**	**64.1**	**70.1**

Advantages of TrackVerse-Derived Features

In this section, we compare features learned from TrackVerse against those from other datasets on the downstream tasks outlined above. We trained view-invariance MoCo on ImageNet-1k [40] and variance-aware MoCo on video datasets (TrackVerse, Walking Tours [58], TAO [70], and Scene Clips), with an equivalent training budget of 220M forward-backward passes. The experimental results show that TrackVerse enables learning more generalizable, state-aware representations, thanks to its rich object motions, diverse states, object-centricity, and large scale.

Object tracks yield more state-aware features. TrackVerse video dataset captures rich dynamics and state changes of various objects. To demonstrate the critical role of motion dynamics in learning object representations, we compare features learned from ImageNet-1K using MoCo with those from TrackVerse using variance-aware MoCo which are pre-trained under the equivalent training budget. As shown in Table 3.2, features learned from TrackVerse not only capture better semantics, as evidenced by superior performance on object classification tasks including LVIS-IN, TrackVerse, and MIT-States, but also exhibit significantly better performance across all tasks requiring dynamic and fine-grained attribute understanding, including SSv2 action recognition, MIT-States attribute recognition, and DAVIS-2017 video object segmentation. This performance gap highlights TrackVerse's effectiveness in

learning features with both semantics and spatiotemporal evolution of object states.

Object centricity facilitates more effective representation learning. The TrackVerse is object-centric, which eliminates contextual distractions including irrelevant background or non-target object information. This object-centric focus enables models to more effectively capture object-specific features, including semantic attributes and state changes. To demonstrate the benefits of object centricity, we compared features learned from two natural video datasets: Walking Tours [58] and Scene Clips. Walking Tours is a video dataset of urban scenes with a long-tail distribution of object categories (with humans being the most prevalent). Scene Clips consists of the original videos in TrackVerse without object tracking, i.e., without implementing Step 4 shown in Fig. 3.2, thus retaining the complete spatial context. As shown in Table 3.2, features learned from TrackVerse consistently outperform those from both Walking Tours and Scene Clips across all evaluation tasks. Notably, Walking Tours, the least object-centric dataset, yielded the lowest performance across all tasks. This result confirms that object-centric datasets mitigate the challenges of scene complexity and spurious correlations, thereby enhancing feature learning efficiency.

Dataset scale and object diversity enable more generalizable features. Our automatic collection pipeline facilitates the acquisition of data at an unprecedented scale. While existing tracking video datasets also offer object-centric dynamics, and even provide more accurate human-annotated category labels, these costly annotation processes inherently limit their scale, hampering their effectiveness for object representation learning. In self-supervised learning, where labels are not used during pretraining, dataset scale and object variety become crucial factors for learning general object representations. To compare, we run the pretrain-

ing on TAO [70], currently the largest tracking dataset available, which only contains 17,287 tracks across 833 classes (see details in Table 3.1). This small scale results in limited within-class variety, insufficient coverage of objects' possible state changes, and class imbalance. By contrast, TrackVerse includes 31.9M object tracks spanning 1203 classes, with the largest class-balanced subset containing 3.8M object tracks. As shown in Table 3.2, features extracted from TrackVerse notably outperform those learned from TAO across all downstream tasks.

Table 3.4: **Comparison of static frames (MoCo) vs. tracks (Variance-Aware MoCo) pretraining on `TrackVerse`**

Pretraining Data (Method)	LVIS-IN NN	TrackVerse NN	TrackVerse FSL	SSv2 Ft	MIT-States Pair Ft.
Frames (MoCo)	34.6	33.7	46.8	55.1	40.3
Tracks (VA-MoCo)	**38.4**	**39.9**	**53.1**	**57.6**	**41.1**

Variance-aware contrastive learning Finally, we compared augmentation invariance MoCo (MoCo) with variance-aware MoCo (VA-MoCo) on TrackVerse (0.4M-CB1000 subset). As shown in Table 3.4, the VA-MoCo framework consistently enhanced learned features across all downstream tasks.

Scalability

An ideal SSL dataset should improve representation quality with more data and longer training. In this section, we examine the scaling properties of TrackVerse, focusing on tasks of NN and FSL.

Training budget We evaluated the scalability of TrackVerse to the training budget by training on the 0.4M-CB1000 subset for 100, 300, and 600 epochs (*i.e.*, 38K, 114K and 228K iterations, respectively). We also

compared this with static frame training to assess the advantages of dynamic content in extended schedules. As shown in Fig. 3.4, representations from dynamic tracks improve consistently with increased training, while static frame training plateaus at 300 epochs.

Dataset size To assess scalability to dataset size, we run variance-aware MoCo pretraining on five TrackVerse subsets of different sizes, while controlling for the training budget (228K training iterations). The results in Table 3.3 show that TrackVerse is capable of effectively leveraging larger datasets to enhance model performance. Given that the TrackVerse data acquisition pipeline is fully automated, scaling up the dataset size is a straightforward process, making it a promising resource for future research in SSL.

3.6 Conclusion

We build TrackVerse, a large-scale dataset of object tracks using an automated collection pipeline. TrackVerse is the largest video dataset ensuring object-centricity, class diversity and rich object motions and states, offering a unique playground for learning general object representations from object dynamics beyond static appearance. In this chapter, we extend contrastive learning with a variance-aware predictor conditioning on data augmentations, maximizing the learning potential from videos. Empirical results validate the effectiveness of TrackVerse for scalable self-supervised learning, outperforming representations learned from static images or natural videos.

4 QUERYABLE ATTRIBUTE REPRESENTATION EXTRACTION FROM FROZEN VISION-LANGUAGE MODELS

While powerful, existing multimodal embeddings are predominantly global, entangling distinct visual factors such as object, style, and background into a single holistic representation. This entanglement fundamentally limits attribute-level control for downstream tasks like fine-grained retrieval or controllable editing. Even embeddings distilled from powerful vision-language models (VLMs), such as VLM2Vec, still struggle to isolate specific attributes on demand. To address this, we introduce Queryable Attribute Representation Extraction (QARE), a new task focused on generating embeddings that are sensitive only to a queried attribute. To enable rigorous evaluation, we present QARE-BENCH, the first benchmark designed for QARE, featuring both synthetic compositions and challenging real-world data. We further propose TF-QARE, a simple yet remarkably effective training-free method that extracts attribute-specific features from frozen VLMs by pooling the hidden states of reply tokens generated in response to a structured prompt. Strikingly, our experiments show that this zero-shot approach is not merely competitive; it substantially outperforms fine-tuned methods like VLM2Vec across a range of VLM backbones on our benchmark.

4.1 Introduction

Multimodal embeddings, which capture semantic correspondences between vision and language, support many core applications such as cross-modal retrieval [47, 48, 98, 96], controllable content creation [116, 117, 118],

Figure 4.1: **Overview of QARE: From Entangled Global Embeddings to Queryable Attribute-Specific Representations.** The left side illustrates the limitation of VLM2Vec [96]: it fails to follow attribute prompts, producing nearly identical retrieval results for an image paired with different attribute queries. The retrieved neighbors show that its embeddings primarily capture overall appearance rather than the requested attribute, preventing true attribute-specific representation. Conversely, our method produces separate attribute-specific representations that support precise query-conditioned retrieval. Prompt texts are simplified here for clarity.

and robotic perception [119]. With the rise of powerful vision–language models (VLMs), recent work further derives general-purpose multimodal embeddings directly from frozen VLMs [96, 97].

However, many tasks require *attribute-level control*: retrieving images that match a specific style but not content, or separately reasoning about object appearance versus background. Existing embeddings are predominantly *global*. They entangle object, background, and style cues in a single holistic representation, making it difficult to isolate the visual factor specified by a user query. Even VLM-based embedding extractors—most notably VLM2Vec [96], which fine-tunes VLMs to learn general-purpose embeddings—still output a global feature that entangles multiple visual factors, lacking the ability to isolate attribute-specific representations conditioned on demand (see Fig. 4.1).

To address this limitation, we introduce the problem of Queryable Attribute Representation Extraction (QARE): given an image I and an attribute a, the goal is to produce an embedding $E(I, a)$ that (i) is sensitive to the specified attribute (and not just the image as a whole), and (ii) is invariant to all other components of the image that are unrelated to the attribute. Such queryable representations would enable more controllable retrieval, editing, and reasoning over individual factors of variation.

Although many datasets evaluate general multimodal alignment or

text-conditioned editing [96, 105, 106, 107, 108, 109], they do not measure whether a method can (i) disentangle intrinsic visual factors in the feature space, or (ii) extract attribute-specific embeddings conditioned on a query. Existing evaluations therefore cannot reveal whether a representation is truly queryable or simply globally entangled.

To fill this gap, we introduce QARE-BENCH, the first benchmark designed explicitly for QARE. It consists of: (i) a synthetic set spanning three orthogonal attributes—object, style, background—constructed images, and (ii) a real-image set containing 6,184 object crops and 2,758 background crops, grouped into 325 and 243 query groups, respectively, with challenging positives and hard negatives. Our evaluation protocol measures both attribute-conditioned retrieval (mAP) and query specificity via intra-image dissimilarity, providing a rigorous testbed for assessing attribute-level embeddings.

We further introduce a surprisingly effective *training-free* method for QARE: extracting attribute-specific embeddings by pooling VLM hidden states from only the reply tokens generated in response to a structured attribute query. This exploits the implicit attribute structure already present in modern VLMs, requiring no fine-tuning or auxiliary training. Our method, TF-QARE, delivers significantly stronger attribute-level retrieval and query sensitivity than both post-trained and global embedding baselines, across all tested VLM backbones on QARE-BENCH.

Our main contributions are summarized as follows:

- **New Task: Queryable Attribute Representation Extraction (QARE).** We identify a core limitation of current multimodal embeddings and formulate QARE to explicitly evaluate *query sensitivity* and *attribute invariance*.
- **A Benchmark for Attribute-Level Evaluation.** We introduce QARE-BENCH, the first dataset designed for QARE, covering object,

background, and style with both synthetic compositions and challenging real-image groups.

- **A Simple and Training-Free Method.** We propose a zero-shot approach that isolates attribute-specific embeddings from frozen VLMs using reply-conditioned token features, requiring no fine-tuning.
- **Strong and Consistent Empirical Gains.** Our method substantially outperforms post-trained and global embedding baselines across all attributes and VLM backbones.

4.2 Related Work

Multimodal embeddings. Early multimodal embedding models, such as VisualBERT [100], learned joint vision–language spaces via cross-modal attention. Dual-encoder methods like CLIP, ALIGN, LiT, and SigLIP [47, 48, 98, 99] scaled this paradigm using large image–text corpora, achieving strong global retrieval features. Models such as BLIP-2 [101] further explored modular alignment between frozen vision encoders and language models. Recent work shifts from training from scratch to extracting embeddings from pretrained VLMs. VLM2Vec [96] fine-tunes VLMs with contrastive learning to obtain competitive global multimodal embeddings. However, existing approaches largely produce entangled features that cannot be conditioned on specific semantic attributes, limiting applications such as attribute-specific editing or retrieval. Beyond general multimodal methods, prior work in fine-grained fashion analysis has explored attribute-specific embeddings [113, 114, 115], but these methods are restricted to the fashion domain and rely on task-specific training. In contrast, our approach

provides training-free, general-purpose attribute querying across diverse objects, backgrounds, and visual styles.

Vision-Language Models (VLMs). Modern VLMs, such as Qwen-VL [102], InternVL [103], and Gemma 3 [104], achieve remarkable multimodal reasoning via instruction tuning on massive corpora. However, these models are optimized for autoregressive text generation rather than producing structured, disentangled semantic embeddings. Consequently, extracting explicit feature representations from them is underexplored. Our `TF-QARE` exploits the richness of VLM representations, reformulating their hidden activations into queryable, attribute-specific embeddings.

Benchmarks for feature disentanglement. Current evaluation protocols do not adequately assess attribute-level isolation. General benchmarks like MMEB [96] focus on global tasks such as classification or holistic matching. Meanwhile, Composed Image Retrieval (CIR) benchmarks (*e.g.*, Fashion-IQ [106], CIRR [108], GeneCIS [109]) measure a model's ability to *modify* a reference image based on text, rather than to *query* and isolate its intrinsic attributes. There is no standard for evaluating how well a model separates visual factors like color, texture, or object identity within a single image's embedding. We address this with `QARE-Bench`, a dedicated benchmark designed to rigorously evaluate attribute-level grounding and disentanglement directly in the feature space, independent of image modification or text generation metrics.

4.3 The `QARE` Benchmark

Here, we define the Queryable Attribute Representation Extraction (`QARE`) task and introduce the `QARE-Bench` benchmark. In addition, we establish

the evaluation protocol based on multi-target retrieval metrics, providing a standardized framework for future research.

Problem Formulation

We introduce Queryable Attribute Representation Extraction (QARE), the task of producing a multimodal embedding of isolated visual attributes. From a given image and attribute, the produced embedding should be disentangled from all other image components.

Formally, let $I \in \mathcal{I}$ be an image and let $a \in \mathcal{A}$ denote an attribute (*e.g.*, $\mathcal{A} = \{\text{object}, \text{style}, \text{background}\}$). The objective of QARE is to build an encoder function $E : \mathcal{I} \times \mathcal{A} \to \mathbb{R}^d$ that maps an image-attribute pair to a d-dimensional embedding vector $\mathbf{v}_a = E(I, a)$ that represents only the specified attribute. An ideal QARE encoder E should thus satisfy two critical properties:

1. Sensitivity. The encoder must be sensitive to the specified attribute. For a given image I, querying different attributes must yield distinct representations. For instance, the embedding for an image's salient object, $\mathbf{v}_{\text{object}} = E(I, \text{object})$, should significantly differ from the embedding of the background, $\mathbf{v}_{\text{background}} = E(I, \text{background})$.

2. Invariance. The representation of a specific attribute should be invariant to all other changes in the image. For example, consider two images, I_1 and I_2, that depict the *same object* but with different styles and backgrounds. A successful QARE encoder should produce highly similar object embeddings:

$$E(I_1, \text{object}) \approx E(I_2, \text{object}). \tag{4.1}$$

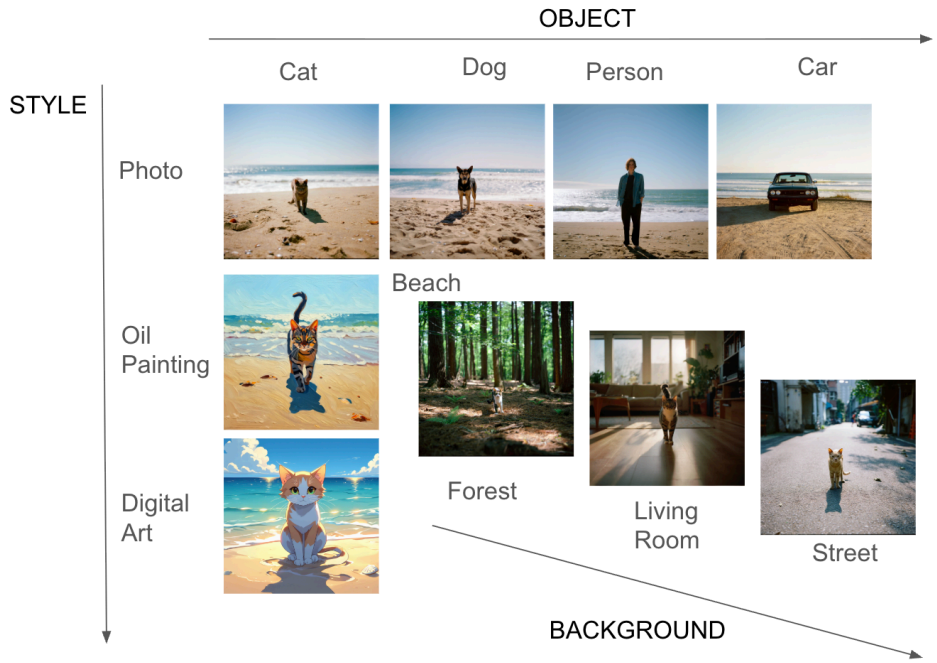
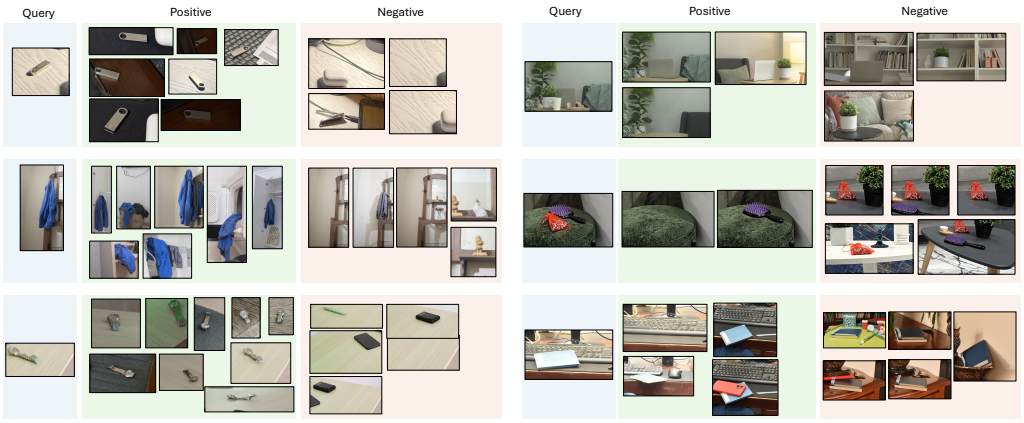


Figure 4.2: **QARE-Bench Synthetic Set.** The synthetic set is created by combinatorially composing instances from three orthogonal attribute categories: **Object** (4 types), **Style** (3 types), and **Background** (4 types). The figure illustrates the attribute axes and shows representative examples.

Conversely, if two images I_1 and I_3 contain different objects, their object embeddings should be dissimilar, *i.e.*, $E(I_1, \text{object}) \not\approx E(I_3, \text{object})$, even if they share the same style or background.

QARE-Bench

The QARE-BENCH benchmark is specifically designed to instantiate and measure these properties, providing a concrete framework for evaluating progress in QARE. Following established frameworks for embedding models, we adopt a retrieval-based evaluation. Each test instance for an attribute a consists of a query image I exhibiting a, with a corresponding set of positives that share a, and a set of hard negatives that do not.



(a) 'Object' examples. (b) 'Background' examples.

Figure 4.3: **Examples from the QARE-Bench Real Set.** Each row illustrates a query group. (a) For object queries, positives contain the identical object instance in varied contexts, while negatives feature different objects, testing for fine-grained identification. (b) The same principle applies to backgrounds, where positives show the target scene across diverse compositions, forcing models to learn an invariant representation.

QARE-BENCH consists of two complementary sets: a synthetic set with a perfectly factorial design for controlled analysis, and a real-world set with challenging, authentic visual scenarios.

Synthetic Set. The synthetic set is intended as a precise *diagnostic tool* designed to probe the core capabilities of models for QARE in a controlled environment. This set is a carefully constructed collection of generated images and exhibits all permutations of **objects**, **styles**, and **backgrounds** illustrated in Fig. 4.2. This $4 \times 3 \times 4$ design results in 48 unique images, where each represents a distinct combination of one object, one style, and one background. Each image yields three test instances (one per attribute), resulting in $48 \times 3 = 144$ total test instances, each having a rich positive and negative set associated with it. For instance, for a query image of a 'cat' with 'object' attribute as the condition, positives

are all (I, object) tuples where I depicts a cat. Conversely, negatives are all images that don't contain cats, including hard negative images with cats when conditioned on other attributes, *i.e.*, style or background. This factorial design eliminates confounding correlations in real data, yielding truly independent attributes and an unambiguous testbed for attribute disentanglement.

Real Set. The real set is built from original, high-resolution photos and extends the evaluation to complex real-world scenarios. This set focuses on object and background attributes. As visualized in Fig. 4.4, we collect images that capture each object in various compositions with other distractor objects in multiple, distinct real-world scenes. From these images, we create a test set where each instance consists of one query image, 2–30 positive images, and 2–71 hard negative images.

Object. To ensure **query** images are unambiguous, we select crops containing a single, clear primary object. Using an open-vocabulary object detector (Detic [110]), we identify all distractor objects and extract the largest crop that isolates the target object while excluding distractors. **Positives** are crops that capture the *identical* target object instance, photographed across diverse backgrounds, contexts, poses, and lighting conditions. Our two types of **negatives** are (i) crops from a photograph taken with a fixed camera position after removing the query object, and (ii) crops of other "distractor" objects that co-occurred in the same original scene.

Background. We choose **queries** as crops from an image, which may contain one or more foreground objects. **Positives** are crops from the *exact same scene* under two conditions: (1) different foreground objects but an identical viewpoint and lighting, and (2) varied shooting angles and lighting. **Negative** crops feature different backgrounds containing one or more foreground objects in the query crop. We locate images of different

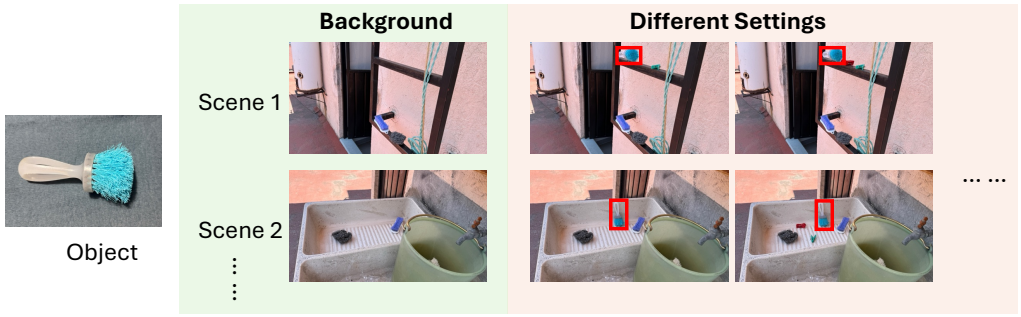


Figure 4.4: **`QARE-Bench` Real Set Image Source.** Each target object instance was photographed across multiple distinct scenes. Within each scene, the object was arranged in various compositions, often alongside other distractor objects. Crucially, we also captured corresponding images of the scene with the target object removed, enabling the creation of verifiably accurate positive and negative examples.

scenes containing the same object and extract crops that maintain similar relative object positions.

In total, this curation yields 325 unique objects and 243 background query groups, with a total of 6,184 and 2,758 crops, respectively. This set spans diverse scenes and resolutions, with challenging positives and hard negatives by design. Fig. 4.3 provides visual examples of the query groups. For each query, the positives represent the ground-truth match for the target attribute, while the hard negatives are specifically selected to create challenging scenarios for disentanglement.

Evaluation Protocol

To comprehensively assess the capabilities of `QARE` models, we design a two-part evaluation protocol that directly measures the sensitivity and invariance properties defined in Section 4.3.

1. Attribute-Conditioned Retrieval. This protocol directly tests whether an embedding for a specific attribute (*e.g.*, an object) remains

constant when other attributes (*e.g.*, style, background) vary.

Given a query image I_q and attribute query a_q, a model computes the attribute-specific embedding $E(I_q, a_q)$. All images I_r in a retrieval set $\mathcal{R}$ are then ranked based on the cosine similarity $\cos\big(E(I_q, a_q), E(I_r, a_q)\big)$.

For quantifying performance, we use **Mean Average Precision (mAP)** based on positives and negatives defined in each set. Let R_q be the set of all relevant items for a query q, with size $t_q = |R_q|$, and let $rel_q(k) \in \{0, 1\}$ be an indicator function that is 1 if the item at rank k is relevant. The Average Precision (AP) for a single query is defined as:

$$AP(q) = \frac{1}{t_q} \sum_{k=1}^{|\mathcal{R}|} P_q(k) \cdot rel_q(k), \tag{4.2}$$

where $P_q(k)$ is the precision at rank k. The final mAP score is the mean of AP scores over all queries.

The mAP metric provides a single, robust score that accounts for both precision and the rank of retrieved items. Crucially, in our benchmark, the number of positive samples t_q can vary significantly from one query to another. Metrics like Recall@K, when averaged globally, can be skewed by this variance. In contrast, mAP is inherently normalized by the number of ground-truth positives for each query via the AP calculation.

2. Intra-Image Similarity. When evaluating specificity, we assess if embeddings for different attributes derived from the *same* image are distinct. For example, similar embeddings for "object" and "background" queries indicate insufficient disentanglement.

For each image I in the test set, we extract the embeddings for its set of core attributes $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ (in our case, object, style, background). We then compute the average pairwise cosine similarity among these em-

beddings:

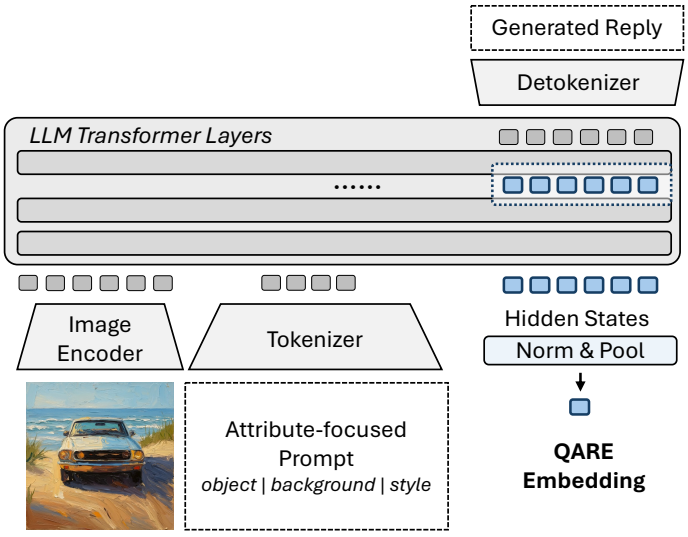
$$S(I) = \frac{1}{\binom{|\mathcal{A}|}{2}} \sum_{1 \leqslant i < j \leqslant |\mathcal{A}|} \text{sim}\big(E(I, a_i), E(I, a_j)\big). \tag{4.3}$$

Our reported metric is the **Average Intra-Image Similarity (AIS)**, which is $S(I)$ averaged over all images in the test set. A low score indicates strong query specificity and effective disentanglement, as the attribute representations lie in different directions in the embedding space. Conversely, a high score suggests that attribute information is entangled, and the model produces a generic, query-agnostic image representation.

4.4 Method

We introduce `TF-QARE` (Training-Free `QARE`), a simple yet effective approach that leverages a frozen pre-trained VLM as a zero-shot encoder to generate prompt-guided, attribute-disentangled embeddings. An overview is shown in Fig. 4.5.

VLM as a zero-shot `QARE` encoder. We argue that general-purpose VLM training on tasks like detailed captioning and visual question answering produces well-grounded latent representations. Our method, `TF-QARE`, is based on the premise that key visual attributes are implicitly disentangled in existing VLM activations. Consequently, we need a method to identify and isolate relevant features. Given an image I and a structured text prompt q (detailed below) targeting an attribute (*e.g.*, object, background, or style), the VLM generates a textual reply $\mathcal{R}$. We then extract the attribute embedding $\mathbf{z}_q$ from the VLM's hidden states. The final embedding $\mathbf{z}_q$ is produced by applying *average pooling* over the *normalized* hidden state vectors corresponding only to the generated reply tokens $\mathcal{R}$. This ensures that the resulting representation is conditioned on both the



(a) TF-QARE overview

Example Replies

Object:
Main object is a car, a classic white coupe with round headlights and a slightly weathered body.

Style:
Visual style is impressionistic, characterized by loose, textured brushstrokes and vibrant, blended colors that capture the essence of the scene rather than its precise details.

Background:
Background is a beach, with a sandy shore, gentle waves, and green grasses in the foreground under a clear blue sky.

(b) Attribute-specific replies (object, style, background) for the image in (a).

Figure 4.5: **Overview of TF-QARE with attribute-focused prompting.** (a) We treat a frozen VLM as a zero-shot QARE encoder: given an image and an attribute-focused prompt (object/background/style), the VLM generates a reply and we pool the normalized hidden states of the reply tokens to obtain an attribute embedding. (b) Example attribute-specific replies from Qwen2-VL-7B for the image in (a), illustrating how different prompts isolate object, style, and background information.

image and the specific attribute query, effectively isolating the target attribute's features.

Attribute-focused prompting. We design structured prompts that compel the model to focus exclusively on a single attribute of the given image while suppressing irrelevant details. Each prompt combines three components: a *direct command* that specifies the target attribute, a *strict output format* that enforces predictable structure, and a set of *negative constraints* that exclude other attributes. As an example, the prompt designed for the *object* attribute is shown below.

Prompt for Object Attribute Extraction

Describe ONLY the main object in the image using a two-part structured format.
FORMAT MUST MATCH EXACTLY:
Main object is [main summary], [detailed description].

Rules:
- [main summary]: 1–3 words describing the object's category, color, and general appearance.
- [detailed description]: 15–30 words expanding on shape, material, surface, parts, or posture.
- Focus on ONE main foreground object only; ignore background, scene, or style.
- Write up to TWO sentences; no lists, no line breaks, no quotes.

Prompts for the *style* and *background* attributes follow the same structure. This combination of positive and negative constraints forces the VLM to generate text that is highly specific to the queried attribute, resulting in distinct, disentangled descriptions for the same input image when different attribute prompts are used.

Method	Backbone	QARE-Bench Synthetic					QARE-Bench Real			
		mAP (↑)				AIS (↓)	mAP (↑)			AIS (↓)
		obj	sty	bg	all		obj	bg	all	
(1) Post-Trained, Queryable										
VLM2VecV1 [96]	Qwen2-VL-7B	8.9	29.6	11.6	16.7	0.97	35.8	36.6	36.2	0.96
VLM2VecV2 [97]	Qwen2-VL-2B	7.9	27.0	11.2	15.4	0.82	46.2	44.8	45.5	0.81
(2) Zero-Shot, Non-Queryable										
Vision Encoder	CLIP	9.4	13.1	8.8	4.5	1.0	32.2	23.2	27.7	1.0
	SigLIP	10.0	11.0	10.1	4.4	1.0	33.4	24.2	28.8	1.0
	DINOv2	13.5	6.8	10.0	4.2	1.0	31.9	23.5	27.7	1.0
	DINOv3	12.1	7.1	11.2	4.1	1.0	30.8	22.3	26.6	1.0
(3) Zero-Shot, Queryable (Ours)										
TF-QARE	Qwen2-VL-2B	8.7	20.5	37.1	22.1	0.63	49.4	43.1	46.2	0.69
	Qwen2-VL-7B	69.7	73.9	91.7	78.4	0.68	66.8	61.9	64.3	0.59
	Qwen2.5-VL-3B	38.7	45.6	91.5	58.6	0.78	62.7	58.6	60.7	0.72
	Qwen2.5-VL-7B	83.9	56.9	90.1	77.0	0.73	65.5	63.7	64.6	0.70
	Qwen2.5-VL-32B	79.0	55.2	91.7	75.3	0.81	63.8	62.0	62.9	0.73
	InternVL3-1B	47.8	23.5	65.6	45.6	0.74	59.7	59.2	59.4	0.80
	InternVL3-2B	46.9	58.0	90.2	65.0	0.75	57.6	55.0	56.3	0.75
	InternVL3-8B	78.0	56.8	91.7	75.5	**0.55**	64.2	61.9	63.1	**0.55**
	InternVL3-14B	**85.8**	55.4	91.7	77.6	0.78	**67.1**	**64.1**	**65.6**	0.78
	Gemma3-4B	55.6	70.4	83.9	70.0	0.88	56.2	58.9	57.6	0.87
	Gemma3-12B	82.9	**75.4**	91.7	**83.3**	0.88	63.0	62.6	62.8	0.88

Table 4.1: **Comparison of different methods on the QARE benchmark.** We evaluate three distinct families of methods: (1) VLM2Vec variants that fine-tune VLMs to produce queryable embeddings; (2) Standard visual encoders that output a single, entangled global embedding; and (3) Our proposed training-free approach TF-QARE directly extracts disentangled attribute features from frozen VLMs and consistently achieves substantial gains, demonstrating the effectiveness of prompt-guided, attribute-aware embedding extraction. Higher mAP (↑) and lower AIS (↓) indicate better performance, and the gray row highlights our default model.

4.5 Experiments

In this section, we comprehensively evaluate our TF-QARE framework on the QARE-Bench. We report comparisons across VLM architectures and scales, analyze performance against post-trained queryable models and zero-shot global visual encoders, and provide ablations on the selection

of optimal VLM layers.

Comparison Results

Across VLMs and scales. We first evaluated our TF-QARE approach across diverse VLM backbones and scales (see Panel (3) of Table 4.1). We observe a general trend where performance improves with model scale. For instance, InternVL3's mAP on the synthetic set increases from 45.6 (1B) to 77.6 (14B). This scaling suggests that larger VLMs possess richer latent representations. Notably, this scaling is not strictly monotonic; the Qwen2.5-VL-32B model underperforms its 7B counterpart (75.3 vs. 77.0 mAP). We attribute this to a potential "alignment tax," where the instruction-tuning of ultra-large models prioritizes complex reasoning and dialogue over the strict adherence to the rigid, descriptive format required by our prompts. This suggests mid-scale models can offer a better trade-off for this specific task. Despite this nuance, the overarching results strongly validate TF-QARE as a powerful strategy for unlocking features from large VLMs.

Compared with post-trained methods. Panel (1) of Table 4.1 compares our TF-QARE with post-trained, queryable embeddings from VLM2VecV1 [96] and VLM2VecV2 [97]. VLM2Vec fine-tunes a frozen VLM on a large collection of multimodal tasks (classification, retrieval, VQA, *etc.*) to produce a single generic embedding per input, and V2 further extends this to more modalities such as video and long documents. On QARE, however, both V1 and V2 perform poorly: they achieve low attribute mAP and high AIS, indicating that object, background, and style remain strongly entangled. This is unsurprising, since the fine-tuning objective optimizes a global task-level representation without explicit pressure to preserve separate attribute factors. Though trained on more data and modalities, V2 still fails to improve QARE scores. This

indicates that simply scaling generic fine-tuning does not guarantee attribute-disentangled embeddings. Directly querying frozen VLMs in a zero-shot manner, as in QARE, can more effectively expose and exploit the latent attribute structure already encoded in these models (see Section 4.6 for further discussion).

Compared with global visual encoders. Panel (2) of Table 4.1 compares QARE with zero-shot global visual encoders such as CLIP [47], SigLIP [99], DINOv2 [111], and DINOv3 [112]. Although these models are trained for image–text alignment, they provide only a single global embedding and lack any instruction-tuned, queryable mechanism. As a result, they cannot isolate object-, background-, or style-specific information, and all extracted features remain fully entangled. This is reflected in both their low attribute mAP and AIS scores fixed at 1.0, indicating that changing the query does not alter the retrieved ranking at all. In contrast, QARE leverages prompt-guided extraction from frozen VLMs to produce genuinely attribute-specific embeddings, leading to significantly stronger performance.

Ablation Study

To understand where attribute information is best encoded within a large VLM, we analyze how QARE performs when extracting representations from different decoder layers using backbone Qwen2-VL-7B. We evaluate layers spanning early, middle, and high depths of the decoder (see Table 4.2). High decoder layers generally yield stronger attribute separation than middle or early layers, and the penultimate decoder layer yields the best overall performance across object, style, and background attributes. We therefore adopt this layer as our default configuration in all main experiments.

Table 4.2: Ablations on layer selection. Backbone: Qwen2-VL-7B.

		Syn. mAP			
	layer	obj.	sty.	bg.	all
high	28 (-1)	62.3	71.1	91.7	75.0
	27 (-2)	**69.7**	**73.9**	**91.7**	**78.4**
	26 (-3)	69.5	72.8	91.7	78.0
middle	21 (-8)	44.9	55.5	88.5	63.0
	13 (-16)	50.7	53.3	83.9	62.6
early	9 (-20)	53.5	46.7	81.5	60.5
	5 (-24)	54.9	43.0	82.4	60.1

4.6 Discussion

A key result from our study is the question of dedicated fine-tuning for disentangled VLM representations compared to zero-shot representation extraction. Our experiments highlight the challenges of training-based approaches like VLM2Vec [96, 97], which consistently underperform our zero-shot approach (see Table 4.1).

However, we argue that these results do not necessarily uncover any fundamental issues with task-specific VLM tuning. Instead, they paint a more nuanced picture of the efficacy of general-purpose training and the availability of dedicated data. For one, VLMs are trained on large amounts of general-purpose data and tasks [102, 103, 104]. Consequently, the models' internal representations are likely to reasonably encode and disentangle various concepts. Our experimental results demonstrate that training-free extraction methods lead to strong performance. Further, such approaches directly benefit from continuous improvements to available models with no additional overhead. Conversely, the success of fine-tuning approaches largely depends on the quality and quantity of the available data. Curating a sufficiently large, dedicated dataset with disentangled image attributes is challenging.

Limitations. `QARE` focuses on three primary visual attributes—object,

background, and style—which cover many common use cases but do not span the full attribute space (*e.g.*, geometry, material, or lighting). Although extracting additional attribute-specific embeddings is straightforward within our framework—requiring only appropriate modifications to the prompt—rigorous evaluation of such attributes would require expanding the benchmark with corresponding curated data. We leave these dataset extensions and broader attribute coverage to future work.

4.7 Summary

In this chapter, we addressed a critical limitation of existing multimodal embeddings: their entangled, global nature, which hinders fine-grained, attribute-level control. To tackle this, we formalized the problem of Queryable Attribute Representation Extraction (QARE) and introduced QARE-Bench, the first benchmark designed to rigorously evaluate attribute isolation and query sensitivity. To solve the task, we proposed TF-QARE, a simple, effective, and training-free method that repurposes frozen vision-language models to extract attribute-specific features via prompt-guided generation.

Across diverse backbones and attribute types, TF-QARE consistently and substantially outperforms post-trained models such as VLM2Vec, despite requiring no fine-tuning or additional supervision. These results indicate that modern VLMs already contain rich attribute-relevant signals, and that structured prompting provides an effective mechanism for isolating them without additional training.

We believe QARE opens new directions for building attribute-controllable multimodal systems, with potential impact on fine-grained retrieval, content creation, and robot perception. We hope our findings encourage further exploration of prompt-guided representation extraction as a lightweight yet powerful alternative to large-scale fine-tuning.

5 CONCLUSION

This dissertation studies how to learn more structured, general, and queryable visual representations from images and videos. While modern vision representation models have achieved strong performance across many downstream tasks, they still face important limitations in capturing fine-grained spatial structure, object-centric temporal dynamics, and attribute-level semantic variation. To address these challenges, this dissertation explores three complementary directions:

In Chapter 2, this dissertation investigated masked image modeling from the perspective of latent-space representation learning. Instead of reconstructing raw pixels, which often contain substantial low-level redundancy, the proposed approach encourages the model to predict meaningful latent representations. This design reduces the emphasis on superficial pixel-level details and helps the model focus on more semantically useful visual structure. Through this study, Chapter 2 shows that the choice of reconstruction target is crucial for masked image modeling, and that latent-space objectives can provide a more effective learning signal for visual representation learning.

In Chapter 3, this dissertation moved from static images to object-centric videos. The central motivation was that many important visual concepts are not fully observable from individual images alone, but emerge through temporal continuity, motion, and state changes. To study this problem, Chapter 3 introduced an object-centric video dataset and used it to improve image-level representation learning with temporally grounded visual information. The results suggest that object-centric videos provide a valuable source of supervision for learning more dynamic and state-aware visual features, complementing image-only pretraining.

In Chapter 4, this dissertation focused on queryable and attribute-

wise visual representation learning. While general-purpose visual embeddings often encode rich information, different visual attributes are usually entangled within a single feature space. Chapter 4 studied how to extract and evaluate attribute-specific representations, with the goal of making visual embeddings more interpretable, controllable, and useful for fine-grained retrieval or reasoning. The findings show that attribute-wise representation learning is a promising direction, but also reveal that current training-based methods do not fully disentangle different visual factors, leaving substantial room for improvement.

The three lines of research presented in this dissertation point toward several promising directions for future work. A central challenge left open is how to scale structured visual representation learning in a principled and general way. The three structural dimensions explored here — spatial, state-aware, and compositional — were each addressed through bespoke frameworks, but a more ambitious goal is to integrate them into a unified model capable of simultaneously capturing multiple facets of visual structure. Recent work on joint-embedding predictive architectures such as I-JEPA [32] and V-JEPA [91] demonstrates that prediction in abstract latent space, rather than pixel space, is a promising direction for learning structured and scalable representations; our findings in Chapter 2 are consistent with and complementary to this line of work.

However, a key challenge that distinguishes visual representation learning from language pretraining remains unresolved: unlike language, which possesses a natural discrete token vocabulary that enables near-exhaustive coverage at scale, visual data is continuous and exhibits high pixel-level redundancy. As a result, visual structure cannot emerge organically from data scale alone — it must be explicitly defined or discovered [92]. Recent empirical work has begun to examine the conditions under which data scaling leads to compositional generalization in vision models [93], finding that scale alone is insufficient and that represen-

tational structure must be deliberately encouraged. This motivates a deeper investigation into what constitutes a *scale-friendly* definition of visual structure: one that is expressive enough to capture meaningful variation across diverse visual domains, yet stable and general enough to benefit consistently from increasing data and model capacity. How the spatial, state-aware, and compositional dimensions of structure explored in this dissertation can be organically unified into such a scalable framework remains an important open question.

The second direction concerns attribute disentanglement in the multimodal setting. Chapter 4 demonstrated that a training-free approach to attribute-specific representation extraction yields surprisingly strong results, exploiting the implicit attribute structure already encoded in frozen VLMs. However, the experiments also revealed that training-free methods do not fully disentangle visual attributes — residual entanglement between object, background, and style persists across all tested backbones. This motivates developing efficient fine-tuning strategies that can further sharpen attribute boundaries in VLM representations without sacrificing the broad multimodal knowledge acquired during pretraining. Recent work on parameter-efficient fine-tuning of VLMs, including lightweight adapter-based methods [94] and vision-specific low-rank adaptation strategies, suggests that targeted fine-tuning of a small subset of model parameters can meaningfully reshape internal representations while preserving general-purpose capabilities [95]. A promising direction is to design such objectives to explicitly optimize for attribute invariance and query sensitivity — for instance, using contrastive supervision derived from our QARE benchmark — while regularizing against catastrophic forgetting of pretrained visual-semantic alignment. We anticipate that the QARE benchmark introduced in Chapter 4 will serve as a natural foundation for evaluating such methods, and more broadly, for advancing controllable, attribute-disentangled

multimodal representation learning.

REFERENCES

[1] Caron, Mathilde, Touvron, Hugo, Misra, Ishan, Jégou, Hervé, Mairal, Julien, Bojanowski, Piotr, Joulin, Armand. Emerging Properties in Self-Supervised Vision Transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[2] Xinlei Chen*, Saining Xie*, Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[3] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, Michal Valko. Bootstrap Your Own Latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Oord, Aaron van den, Li, Yazhe, Vinyals, Oriol. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[5] Xiao, Tete, Wang, Xiaolong, Efros, Alexei A, Darrell, Trevor. What Should Not Be Contrastive in Contrastive Learning. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[6] He, Kaiming, Chen, Xinlei, Xie, Saining, Li, Yanghao, Dollár, Piotr, Girshick, Ross. Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition (CVPR), pp. 16000–16009, 2022.

[7] Wei, Chen, Fan, Haoqi, Xie, Saining, Wu, Chao-Yuan, Yuille, Alan, Feichtenhofer, Christoph. Masked feature prediction for self-supervised visual pre-training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14668–14678*, 2022.

[8] Bao, Hangbo, Dong, Li, Piao, Songhao, Wei, Furu. BEiT: BERT Pre-Training of Image Transformers. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[9] Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, Jakob Uszkoreit, Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[10] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, Polosukhin, Illia. Attention is all you need. *Advances in neural information processing systems, vol. 30*, 2017.

[11] Bao, Hangbo, Dong, Li, Piao, Songhao, Wei, Furu. BEIT: BERT pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[12] Tian, Yonglong, Krishnan, Dilip, Isola, Phillip. Contrastive multiview coding. *Proceedings of the European Conference on Computer Vision (ECCV), pp. 776–794*, 2020.

[13] Tarvainen, Antti, Valpola, Harri. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems (NeurIPS), vol. 30*, 2017.

[14] Baevski, Alexei, Hsu, Wei-Ning, Xu, Qiantong, Babu, Arun, Gu, Jiatao, Auli, Michael. Data2vec: A general framework for self-supervised learning in speech, vision and language. *International Conference on Machine Learning, pp. 1298–1312*, 2022.

[15] Zhou, Jinghao, Wei, Chen, Wang, Huiyu, Shen, Wei, Xie, Cihang, Yuille, Alan, Kong, Tao. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

[16] Chen, Xiaokang, Ding, Mingyu, Wang, Xiaodi, Xin, Ying, Mo, Shentong, Wang, Yunhao, Han, Shumin, Luo, Ping, Zeng, Gang, Wang, Jingdong. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision, pp. 1–16*, 2023.

[17] Zhou, Kaiyang, Yang, Jingkang, Loy, Chen Change, Liu, Ziwei. Conditional prompt learning for vision-language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16816–16825*, 2022.

[18] You, Yang, Gitman, Igor, Ginsburg, Boris. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[19] Krause, Jonathan, Stark, Michael, Deng, Jia, Fei-Fei, Li. 3d object

representations for fine-grained categorization. *Proceedings of the IEEE international conference on computer vision workshops, pp. 554–561*, 2013.

[20] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, A. Vedaldi. Describing Textures in the Wild. *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[21] Nilsback, Maria-Elena, Zisserman, Andrew. Automated flower classification over a large number of classes. *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp. 722–729*, 2008.

[22] Parkhi, Omkar M, Vedaldi, Andrea, Zisserman, Andrew, Jawahar, CV. Cats and dogs. *2012 IEEE conference on computer vision and pattern recognition, pp. 3498–3505*, 2012.

[23] Xiao, Jianxiong, Ehinger, Krista A, Hays, James, Torralba, Antonio, Oliva, Aude. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision, vol. 119, pp. 3–22*, 2016.

[24] Soomro, Khurram, Zamir, Amir Roshan, Shah, Mubarak. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[25] Zhang, Renrui, Guo, Ziyu, Gao, Peng, Fang, Rongyao, Zhao, Bin, Wang, Dong, Qiao, Yu, Li, Hongsheng. Point-M2AE: multi-scale masked autoencoders for hierarchical point cloud pre-training. *arXiv preprint arXiv:2205.14401*, 2022.

[26] Fei-Fei, Li, Fergus, Rob, Perona, Pietro. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *2004 conference on computer vision and pattern recognition workshop, pp. 178–178*, 2004.

[27] Dong, Xiaoyi, Bao, Jianmin, Zhang, Ting, Chen, Dongdong, Zhang, Weiming, Yuan, Lu, Chen, Dong, Wen, Fang, Yu, Nenghai, Guo, Baining. Peco: Perceptual codebook for bert pre-training of vision transformers. *Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 552–560*, 2023.

[28] Tao, Chenxin, Zhu, Xizhou, Su, Weijie, Huang, Gao, Li, Bin, Zhou, Jie, Qiao, Yu, Wang, Xiaogang, Dai, Jifeng. Siamese image modeling for self-supervised vision representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2132–2141*, 2023.

[29] Yi, Kun, Ge, Yixiao, Li, Xiaotong, Yang, Shusheng, Li, Dian, Wu, Jianping, Shan, Ying, Qie, Xiaohu. Masked image modeling with denoising contrast. *arXiv preprint arXiv:2205.09616*, 2022.

[30] Jabri, Allan, Owens, Andrew, Efros, Alexei. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems, vol. 33, pp. 19545–19560*, 2020.

[31] Xie, Zhenda, Zhang, Zheng, Cao, Yue, Lin, Yutong, Bao, Jianmin, Yao, Zhuliang, Dai, Qi, Hu, Han. Simmim: A simple framework for masked image modeling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9653–9663*, 2022.

[32] Assran, Mahmoud, Duval, Quentin, Misra, Ishan, Bojanowski, Piotr, Vincent, Pascal, Rabbat, Michael, LeCun, Yann, Ballas, Nicolas. Self-supervised learning from images with a joint-embedding predictive architecture. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[33] Loshchilov, Ilya, Hutter, Frank. Decoupled Weight Decay Regularization. *International Conference on Learning Representations*, 2018.

[34] Dosovitskiy, Alexey, Fischer, Philipp, Springenberg, Jost Tobias, Riedmiller, Martin, Brox, Thomas. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI), vol. 38, pp. 1734–1747*, 2015.

[35] Wu, Zhirong, Xiong, Yuanjun, Yu, Stella X, Lin, Dahua. Unsupervised feature learning via non-parametric instance discrimination. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3733–3742*, 2018.

[36] Chen, Ting, Kornblith, Simon, Norouzi, Mohammad, Hinton, Geoffrey. A simple framework for contrastive learning of visual representations. *Proceedings of the International Conference on Machine Learning (ICML), pp. 1597–1607*, 2020.

[37] He, Kaiming, Fan, Haoqi, Wu, Yuxin, Xie, Saining, Girshick, Ross. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9729–9738*, 2020.

[38] Caron, Mathilde, Misra, Ishan, Mairal, Julien, Goyal, Priya, Bojanowski, Piotr, Joulin, Armand. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[39] Wei, Yibing, Gupta, Abhinav, Morgado, Pedro. Towards latent masked image modeling for self-supervised visual representation learning. *European Conference on Computer Vision, pp. 1–17*, 2024.

[40] Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255*, 2009.

[41] Pathak, Deepak, Girshick, Ross, Dollár, Piotr, Darrell, Trevor, Hariharan, Bharath. Learning features by watching objects move. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2701–2710*, 2017.

[42] Wang, Xiaolong, Gupta, Abhinav. Unsupervised learning of visual representations using videos. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2794–2802*, 2015.

[43] Qian, Rui, Meng, Tianjian, Gong, Boqing, Yang, Ming-Hsuan, Wang, Huisheng, Belongie, Serge, Cui, Yin. Spatiotemporal contrastive video representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6964–6974*, 2021.

[44] Xue, Hongwei, Hang, Tiankai, Zeng, Yanhong, Sun, Yuchong, Liu, Bei, Yang, Huan, Fu, Jianlong, Guo, Baining. Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5036–5045*, 2022.

[45] Abu-El-Haija, Sami, Kothari, Nisarg, Lee, Joonseok, Natsev, Paul, Toderici, George, Varadarajan, Balakrishnan, Vijayanarasimhan, Sudheendra. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[46] Arandjelovic, Relja, Zisserman, Andrew. Look, listen and learn. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 609–617*, 2017.

[47] Radford, Alec, Kim, Jong Wook, Hallacy, Chris, Ramesh, Aditya, Goh, Gabriel, Agarwal, Sandhini, Sastry, Girish, Askell, Amanda, Mishkin, Pamela, Clark, Jack, Gretchen Krueger, Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[48] Jia, Chao, Yang, Yinfei, Xia, Ye, Chen, Yi-Ting, Parekh, Zarana, Pham, Hieu, Le, Quoc, Sung, Yun-Hsuan, Li, Zhen, Duerig, Tom. Scaling up visual and vision-language representation learning with noisy text supervision. *International conference on machine learning, pp. 4904–4916*, 2021.

[49] Feichtenhofer, Christoph, Fan, Haoqi, Li, Yanghao, He, Kaiming. Masked Autoencoders As Spatiotemporal Learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,

2022.

[50] Morgado, Pedro, Vasconcelos, Nuno, Misra, Ishan. Audio-visual instance discrimination with cross-modal agreement. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12475–12486*, 2021.

[51] Goyal, Raghav, Ebrahimi Kahou, Samira, Michalski, Vincent, Materzynska, Joanna, Westphal, Susanne, Kim, Heuna, Haenel, Valentin, Fruend, Ingo, Yianilos, Peter, Mueller-Freitag, Moritz, others. The" something something" video database for learning and evaluating visual common sense. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5842–5850*, 2017.

[52] Gedas Bertasius, Heng Wang, Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding?. *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[53] Phillip Isola, Joseph J. Lim, Edward H. Adelson. Discovering States and Transformations in Image Collections. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[54] Zheng, Zhaoheng, Zhu, Haidong, Nevatia, Ram. CAILA: Concept-Aware Intra-Layer Adapters for Compositional Zero-Shot Learning. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1721-1731*, 2024.

[55] Gordon, Daniel, Ehsani, Kiana, Fox, Dieter, Farhadi, Ali. Watching the World Go By: Representation Learning from Unlabeled Videos. 2020.

[56] Parthasarathy, Nikhil, Eslami, SM, Carreira, João, Hénaff, Olivier J. Self-supervised video pretraining yields human-aligned visual representations. *arXiv preprint arXiv:2210.06433*, 2022.

[57] Grauman, Kristen, Westbury, Andrew, Byrne, Eugene, Chavis, Zachary, Furnari, Antonino, Girdhar, Rohit, Hamburger, Jackson, Jiang, Hao, Liu, Miao, Liu, Xingyu, others. Ego4d: Around the world in 3,000 hours of egocentric video. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18995–19012*, 2022.

[58] Venkataramanan, Shashanka, Rizve, Mamshad Nayeem, Carreira, João, Asano, Yuki M, Avrithis, Yannis. Is ImageNet worth 1 video? Learning strong image encoders from 1 long unlabelled video. *arXiv preprint arXiv:2310.08584*, 2023.

[59] Karpathy, Andrej, Toderici, George, Shetty, Sanketh, Leung, Thomas, Sukthankar, Rahul, Fei-Fei, Li. Large-scale video classification with convolutional neural networks. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1725–1732*, 2014.

[60] Kuehne, Hildegard, Jhuang, Hueihan, Garrote, Estíbaliz, Poggio, Tomaso, Serre, Thomas. HMDB: a large video database for human motion recognition. *2011 International conference on computer vision, pp. 2556–2563*, 2011.

[61] Kay, Will, Carreira, Joao, Simonyan, Karen, Zhang, Brian, Hillier, Chloe, Vijayanarasimhan, Sudheendra, Viola, Fabio, Green, Tim, Back, Trevor, Natsev, Paul, others. The kinetics human action video dataset.

arXiv preprint arXiv:1705.06950, 2017.

[62] Damen, Dima, Doughty, Hazel, Farinella, Giovanni Maria, Furnari, Antonino, Kazakos, Evangelos, Ma, Jian, Moltisanti, Davide, Munro, Jonathan, Perrett, Toby, Price, Will, others. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision, pp. 1–23*, 2022.

[63] Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. *arXiv:1504.01942 [cs]*, 2015.

[64] Perazzi, Federico, Pont-Tuset, Jordi, McWilliams, Brian, Van Gool, Luc, Gross, Markus, Sorkine-Hornung, Alexander. A benchmark dataset and evaluation methodology for video object segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 724–732*, 2016.

[65] Zhou, Luowei, Xu, Chenliang, Corso, Jason J. Towards Automatic Learning of Procedures from Web Instructional Videos. *arXiv preprint arXiv:1703.09788*, 2017.

[66] Zhou, Xingyi, Girdhar, Rohit, Joulin, Armand, Krähenbühl, Philipp, Misra, Ishan. Detecting twenty-thousand classes using image-level supervision. *European Conference on Computer Vision, pp. 350–368*, 2022.

[67] Zhang, Yifu, Sun, Peize, Jiang, Yi, Yu, Dongdong, Weng, Fucheng, Yuan, Zehuan, Luo, Ping, Liu, Wenyu, Wang, Xinggang. Bytetrack: Multi-object tracking by associating every detection box. *European confer-*

ence on computer vision, pp. 1–21, 2022.

[68] Gupta, Agrim, Dollar, Piotr, Girshick, Ross. LVIS: A dataset for large vocabulary instance segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5356–5364*, 2019.

[69] Brandon Castellano. PySceneDetect. *GitHub repository*,

[70] Dave, Achal, Khurana, Tarasha, Tokmakov, Pavel, Schmid, Cordelia, Ramanan, Deva. Tao: A large-scale benchmark for tracking any object. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[71] Wiskott, Laurenz, Sejnowski, Terrence J. Slow feature analysis: Unsupervised learning of invariances. *Neural computation, vol. 14, pp. 715–770*, 2002.

[72] Hurri, Jarmo, Hyvärinen, Aapo. Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation, vol. 15, pp. 663–691*, 2003.

[73] Agrawal, Pulkit, Carreira, Joao, Malik, Jitendra. Learning to see by moving. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

[74] Recasens, Adria, Luc, Pauline, Alayrac, Jean-Baptiste, Wang, Luyu, Strub, Florian, Tallec, Corentin, Malinowski, Mateusz, Pătrăucean, Viorica, Altché, Florent, Valko, Michal, others. Broaden your views for self-supervised video learning. *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), 2021.

[75] Feichtenhofer, Christoph, Fan, Haoqi, Xiong, Bo, Girshick, Ross, He, Kaiming. A large-scale study on unsupervised spatiotemporal representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[76] Bian, Zhangxing, Jabri, Allan, Efros, Alexei A, Owens, Andrew. Learning pixel trajectories with multiscale contrastive random walks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[77] Sharma, Yash, Zhu, Yi, Russell, Chris, Brox, Thomas. Pixel-level correspondence for self-supervised learning from video. *arXiv preprint arXiv:2207.03866*, 2022.

[78] Xiong, Yuwen, Ren, Mengye, Zeng, Wenyuan, Urtasun, Raquel. Self-supervised representation learning from flow equivariance. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[79] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, Luka Čehovin. A Novel Performance Evaluation Methodology for Single-Target Trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, pp. 2137-2155*, 2016.

[80] Pont-Tuset, Jordi, others. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[81] Tschannen, Michael, Djolonga, Josip, Ritter, Marvin, Mahendran, Aravindh, Houlsby, Neil, Gelly, Sylvain, Lucic, Mario. Self-supervised learning of video-induced visual invariances. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13806–13815*, 2020.

[82] Orhan, Emin, Gupta, Vaibhav, Lake, Brenden M. Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems, vol. 33, pp. 9960–9971*, 2020.

[83] Sermanet, Pierre, Lynch, Corey, Chebotar, Yevgen, Hsu, Jasmine, Jang, Eric, Schaal, Stefan, Levine, Sergey, Brain, Google. Time-contrastive networks: Self-supervised learning from video. *2018 IEEE international conference on robotics and automation (ICRA), pp. 1134–1141*, 2018.

[84] Dave, Ishan, Gupta, Rohit, Rizve, Mamshad Nayeem, Shah, Mubarak. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding, vol. 219, pp. 103406*, 2022.

[85] Salehi, Mohammadreza, Gavves, Efstratios, Snoek, Cees GM, Asano, Yuki M. Time does tell: Self-supervised time-tuning of dense image representations. *Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16536–16547*, 2023.

[86] Haresh, Sanjay, Kumar, Sateesh, Coskun, Huseyin, Syed, Shahram N, Konin, Andrey, Zia, Zeeshan, Tran, Quoc-Huy. Learning by aligning videos in time. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5548–5558*, 2021.

[87] Chen, Minghao, Wei, Fangyun, Li, Chong, Cai, Deng. Frame-wise action representations for long videos via sequence contrastive learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13801–13810*, 2022.

[88] Morgado, Pedro, Vasconcelos, Nuno, Misra, Ishan. Audio-visual instance discrimination with cross-modal agreement. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12475–12486*, 2021.

[89] Morgado, Pedro, Misra, Ishan, Vasconcelos, Nuno. Robust Audio-Visual Instance Discrimination. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12934-12945*, 2021.

[90] Assran, Mahmoud, Duval, Quentin, Misra, Ishan, Bojanowski, Piotr, Vincent, Pascal, Rabbat, Michael, LeCun, Yann, Ballas, Nicolas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[91] Bardes, Adrien, Garrido, Quentin, Ponce, Jean, Chen, Xinlei, Rabbat, Michael, LeCun, Yann, Assran, Mahmoud, Ballas, Nicolas. Revisiting Feature Prediction for Learning Visual Representations from Video. *arXiv preprint arXiv:2404.08471*, 2024.

[92] Balestriero, Randall, Ibrahim, Mark, Sobal, Vlad, Morcos, Ari, Shekhar, Shashank, Goldstein, Tom, Bordes, Florian, Bardes, Adrien, Mialon, Gregoire, Tian, Yuandong, others. A Cookbook of Self-Supervised

Learning. *arXiv preprint arXiv:2304.12210*, 2023.

[93] Uselis, Arnas, Dittadi, Andrea, Oh, Seong Joon. Does Data Scaling Lead to Visual Compositional Generalization?. *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[94] Luo, Tiange, Logeswaran, Lajanugen, Kim, Jaekyeom, Johnson, Justin, Lee, Honglak. Towards Minimal Fine-Tuning of VLMs. *arXiv preprint arXiv:2512.19219*, 2025.

[95] Ypsilantis, Nikolaos-Antonios, Chen, Kaifeng, Araujo, André, Chum, Ondřej. Infusing Fine-Grained Visual Knowledge to Vision-Language Models. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2025.

[96] Jiang, Ziyan, Meng, Rui, Yang, Xinyi, Yavuz, Semih, Zhou, Yingbo, Chen, Wenhu. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks. *arXiv preprint arXiv:2410.05160*, 2024.

[97] Meng, Rui, Jiang, Ziyan, Liu, Ye, Su, Mingyi, Yang, Xinyi, Fu, Yuepeng, Qin, Can, Chen, Zeyuan, Xu, Ran, Xiong, Caiming, Zhou, Yingbo, Chen, Wenhu, Yavuz, Semih. VLM2Vec-V2: Advancing Multimodal Embedding for Videos, Images, and Visual Documents. *arXiv preprint arXiv:2507.04590*, 2025.

[98] Zhai, Xiaohua, Wang, Xiao, Mustafa, Basil, Steiner, Andreas, Keysers, Daniel, Kolesnikov, Alexander, Beyer, Lucas. LiT: Zero-Shot Transfer With Locked-Image Text Tuning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18102–18112*, 2022.

[99] Zhai, Xiaohua, Mustafa, Basil, Kolesnikov, Alexander, Beyer, Lucas. Sigmoid Loss for Language Image Pre-Training. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[100] Li, Liunian Harold, Yatskar, Mark, Yin, Da, Hsieh, Cho-Jui, Chang, Kai-Wei. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557*, 2019.

[101] Li, Junnan, Li, Dongxu, Savarese, Silvio, Hoi, Steven. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *Proceedings of the International Conference on Machine Learning (ICML), pp. 19730–19742*, 2023.

[102] Bai, Shuai, Chen, Keqin, Liu, Xuejing, Wang, Jialin, Ge, Wenbin, Song, Sibo, Dang, Kai, others. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

[103] Zhu, Jinguo, others. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025.

[104] Gemma Team. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025.

[105] Liu, Ziwei, Luo, Ping, Qiu, Shi, Wang, Xiaogang, Tang, Xiaoou. DeepFashion: Powering Robust Clothes Recognition and Retrieval With Rich Annotations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1096–1104*, 2016.

[106] Wu, Hui, Gao, Yupeng, Guo, Xiaoxiao, Al-Halah, Ziad, Rennie, Steven, Grauman, Kristen, Feris, Rogério. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11307–11317*, 2021.

[107] Liu, Ziwei, Luo, Ping, Wang, Xiaogang, Tang, Xiaoou. Deep Learning Face Attributes in the Wild. *Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3730–3738*, 2015.

[108] Liu, Zheyuan, Rodriguez-Opazo, Cristian, Teney, Damien, Gould, Stephen. Image Retrieval on Real-life Images With Pre-trained Vision-and-Language Models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[109] Vaze, Sagar, Carion, Nicolas, Misra, Ishan. GeneCIS: A Benchmark for General Conditional Image Similarity. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[110] Zhou, Xingyi, Girdhar, Rohit, Joulin, Armand, Krähenbühl, Philipp, Misra, Ishan. Detecting Twenty-thousand Classes Using Image-level Supervision. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[111] Oquab, Maxime, Darcet, Timothée, Moutakanni, Théo, Vo, Huy V., Szafraniec, Marc, Khalidov, Vasil, Fernandez, Pierre, Haziza, Daniel, Massa, Francisco, El-Nouby, Alaaeldin, others. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*,

2023.

[112] Siméoni, Oriane, Vo, Huy V., Seitzer, Maximilian, Baldassarre, Federico, Oquab, Maxime, others. DINOv3: Self-supervised Large Visual Models for Vision at Unprecedented Scale. *arXiv preprint arXiv:2508.10104*, 2025.

[113] Veit, Andreas, Belongie, Serge, Karaletsos, Theofanis. Conditional Similarity Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1781–1789*, 2017.

[114] Ma, Zhe, Dong, Jianfeng, Long, Zhongzi, Zhang, Yao, He, Yuan, Xue, Hui, Ji, Shouling. Fine-Grained Fashion Similarity Learning by Attribute-Specific Embedding Network. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[115] Jiao, Yang, Gao, Yan, Meng, Jingjing, Shang, Jin, Sun, Yi. Learning Attribute and Class-Specific Representation Duet for Fine-Grained Fashion Analysis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[116] Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick, Ommer, Björn. High-Resolution Image Synthesis With Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[117] Brooks, Tim, Holynski, Aleksander, Efros, Alexei A. InstructPix2Pix: Learning to Follow Image Editing Instructions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*,

2023.

[118] Zhang, Lvmin, Rao, Anyi, Agrawala, Maneesh. Adding Conditional Control to Text-to-Image Diffusion Models. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[119] Lv, Qi, Li, Hao, Deng, Xiang, Shao, Rui, Wang, Michael Yu, Nie, Liqiang. RoboMP^2: A Robotic Multimodal Perception-Planning Framework with Multimodal Large Language Models. *arXiv preprint arXiv:2404.04929*, 2024.