Towards Text-Guided Attribute-Disentangled Multimodal Representation Learning

Yibing Wei[1]*, Sudeep Katakol[2], Manuel Brack[2], Jinhong Lin[1], Haoyue Bai[1], Yu-Teng Li[2], Richard Zhang[2], Eli Shechtman[2], Hareesh Ravi[2], Ajinkya Kale[2]

[1]University of Wisconsin-Madison [2]Adobe

Abstract

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 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.

1. Introduction

Multimodal embeddings, which capture semantic correspondences between vision and language, support many core applications such as cross-modal retrieval [7, 10, 11, 15, 25, 32, 35], medical reporting [6, 24], controllable content creation [2, 5, 11, 13, 26, 27, 37], and robotic perception [20, 38]. With the rise of powerful vision–language models (VLMs), recent work further derives general-purpose multimodal embeddings directly from frozen VLMs [3, 8, 22].

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 [8], 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. 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 [8, 16–18, 30, 33], 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)

*Work done during an internship at Adobe.

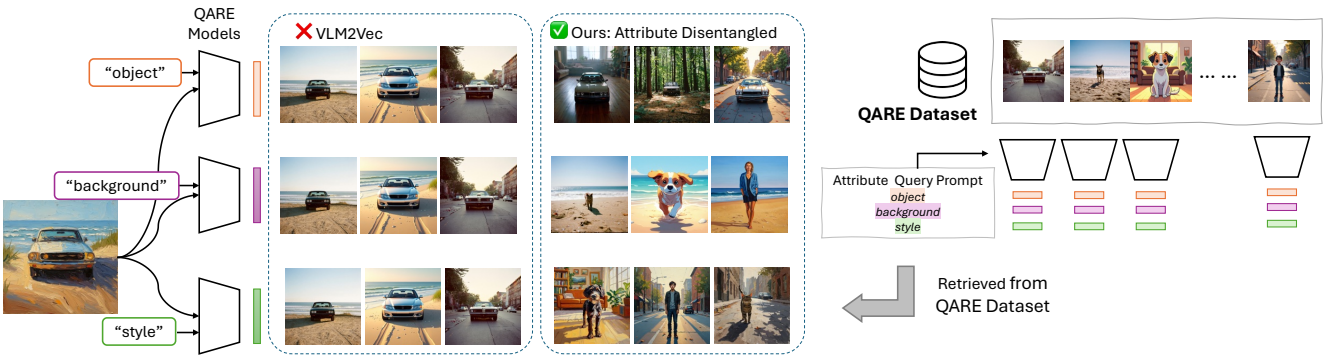


Figure 1. Overview of QARE: From Entangled Global Embeddings to Queryable Attribute-Specific Representations. The left side illustrates the limitation of VLM2Vec [8] (✗): it fails to follow attribute prompts, producing nearly identical retrieval results for an image paired with different attribute queries. 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 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.

2. Related Work

Multimodal embeddings Early multimodal embedding models, such as VisualBERT [14], learned joint vision–language spaces via cross-modal attention. Dual-encoder methods like CLIP, ALIGN, LiT, and SigLIP [7, 25, 35, 36] scaled this paradigm using large image–text corpora, achieving strong global retrieval features. Models such as BLIP-2 [12] further explored modular alignment between frozen vision encoders and LLMs. Recent work shifts from training from scratch to extracting embeddings from pretrained VLMs. VLM2Vec [8] 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 [9, 21, 31], 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 [1], InternVL [40], and Gemma 3 [4], 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 underex-

plored. 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 [8] focus on global tasks such as classification or holistic matching. Meanwhile, Composed Image Retrieval (CIR) benchmarks (e.g., Fashion-IQ [34], CIRR [19], GeneCIS [29]) 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.

3. 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.

3.1. 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:

$$E(I_1, \text{object}) \approx E(I_2, \text{object}). \tag{1}$$

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.

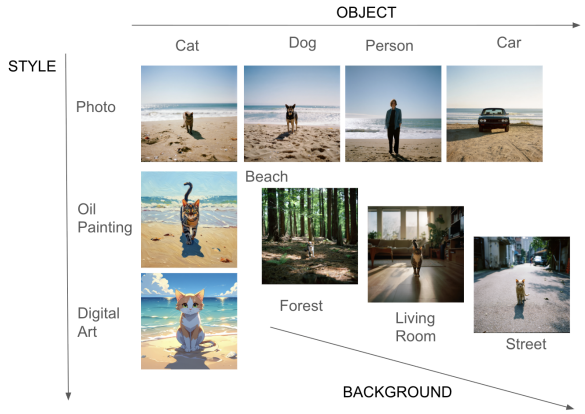
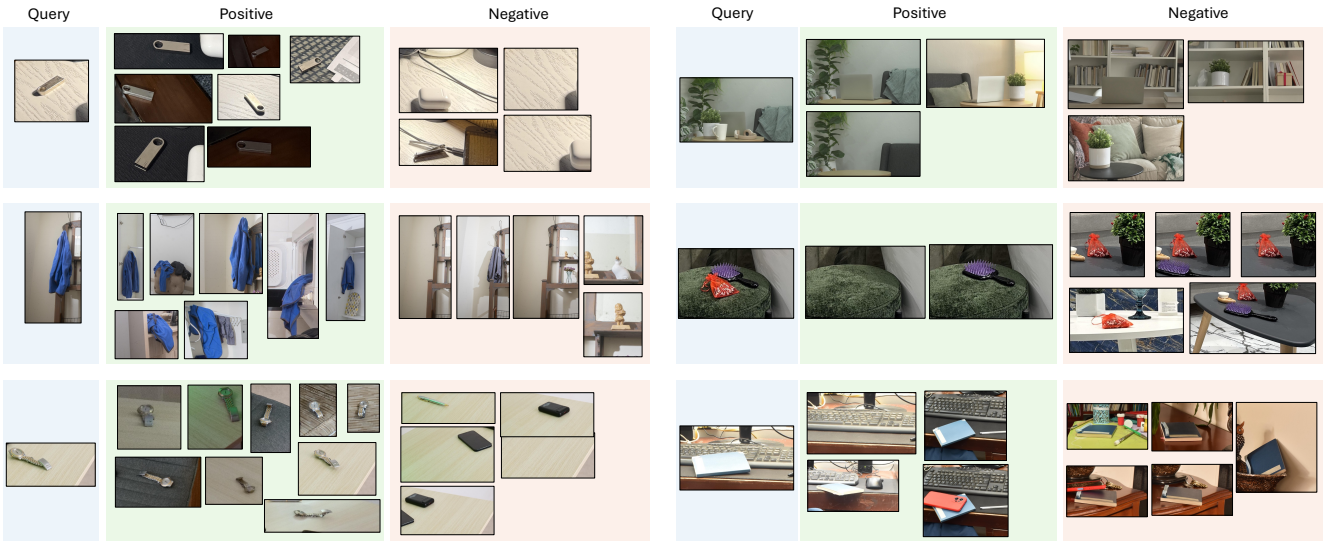


Figure 2. QARE-Bench Synthetic Set. The synthetic set is created by combinatorially composing instances from three orthogonal attribute categories: **Object** (4 types), **Style** (3 types), and **Background** (4 types). The figure illustrates the attribute axes and shows representative examples.

3.2. 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. 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. 2.



(a) 'Object' examples from QARE-BENCH Real. **(b)** 'Background' examples from QARE-BENCH Real.

Figure 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, where positives show the target object across diverse scenes, forcing models to learn an invariant representation.

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 $48 \times 3 = 192$ 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, 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 [39]), 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 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. 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.

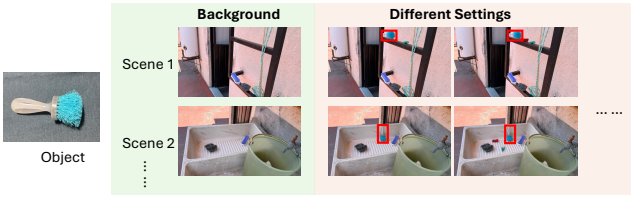


Figure 4. QARE-Bench Real Set Image Source. Each target object instance was photographed across multiple distinct scenes. Within each scene, the object was arranged in various compositions, often alongside other distractor objects. Crucially, we also captured corresponding images of the scene with the target object removed, enabling the creation of verifiably accurate positive and negative examples.

3.3. 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 3.1.

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, i.e. a positive sample. The Average Precision (AP) for a single query is defined as:

$$\mathrm{AP}(q) = \frac{1}{t_q} \sum_{k=1}^{|\mathcal{R}|} P_q(k) \cdot rel_q(k), \tag{2}$$

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 embeddings:

$$S(I) = \frac{1}{\binom{|\mathcal{A}|}{2}} \sum_{1 \le i < j \le |\mathcal{A}|} \mathrm{sim}(E(I, a_i), E(I, a_j)). \tag{3}$$

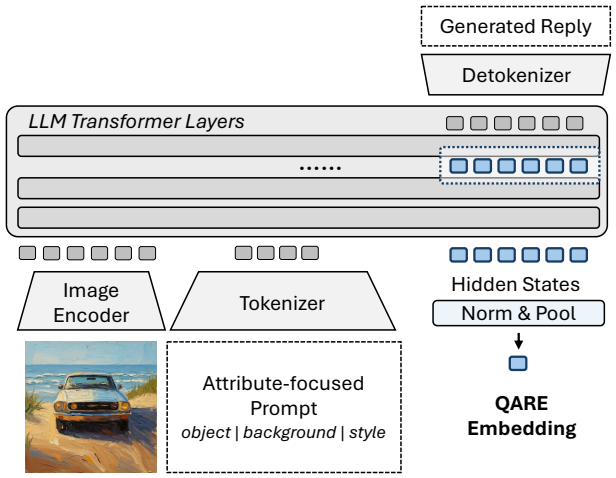
Our reported metric is the **Average Intra-Image Similarity**, which is $S(I)$ averaged over all images in the test set. A low score indicates strong query specificity and effective disentanglement, as the attribute representations lie in different directions in the embedding space. Conversely, a high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

4. Methods

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.

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 image and the specific attribute query, effectively isolating the target attribute's features. We compare average and weighted pooling in Table 2 and adopt average pooling by default.

Attribute-Focused Prompting. We design structured prompts that compel the model to focus exclusively on a single attribute of the given image while



(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 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

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.
```

Figure 6. Attribute-focused prompt format (object).

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. Fig. 6 shows the prompt designed for the *object* attribute. Prompts for the *style* and *background* attributes follow the same structure and are provided in the supplementary material. This combination of positive and negative constraints forces the VLM to generate text that is highly specific to the queried attribute. As shown in Fig. 5b, this results in distinct, disentangled descriptions for the same input image when different attribute prompts are used.

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.

5.1. Comparison Results

Across VLMs and Scales. We first evaluated our TF-QARE approach across diverse VLM backbones and scales (see Panel (3) of Table 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 spe-

Method	Backbone	QARE-Bench Synthetic					QARE-Bench Real			
		mAP (↑)				AIS (↓)	mAP (↑)			AIS (↓)
		obj	sty	bg	all		obj	bg	all	
(1) Post-Trained, Queryable										
VLM2VecV1 [8]	Qwen2-VL-7B	8.9	29.6	11.6	16.7	0.97	35.8	36.6	36.2	0.96
VLM2VecV2 [22]	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 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.

cific 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 1 compares our TF-QARE with post-trained, queryable embeddings from VLM2VecV1 [8] and VLM2VecV2 [22]. 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 6 for further discussion).

Compared with Global Visual Encoders. Panel (2) of Table 1 compares QARE with zero-shot global visual encoders such as CLIP [25], SigLIP [36], DINOv2 [23], and DINOv3 [28]. 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.

5.2. 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 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 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

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 [8, 22] which consistently underperform our zero-shot approach (see Table 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 [1, 4, 40]. 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.

7. Conclusion

In this work, 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 finetuning 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 finetuning.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 8

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[3] ByteDance Seed Team. Built on seed1.6-flash, seed-1.6-embedding launched. https://seed.bytedance.com/en/blog/built-on-seed1-6-flash-seed-1-6-embedding-launched, 2025. Accessed: 2025-11-13. 1

[4] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 2, 8

[5] Xuehai He, Jian Zheng, Jacob Zhiyuan Fang, Robinson Piramuthu, Mohit Bansal, Vicente Ordonez, Gunnar A. Sigurdsson, Nanyun Peng, and Xin Eric Wang. Flexecontrol: Flexible and efficient multimodal control for text-to-image generation. *arXiv preprint arXiv:2405.04834*, 2024. 1

[6] Jia-Hong Huang, Ting-Wei Wu, and Marcel Worring. Contextualized keyword representations for multimodal retinal image captioning. In *Proceedings of the 2021 ACM International Conference on Multimedia Retrieval (ICMR '21)*, 2021. 1

[7] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR, 2021. 1, 2

[8] Zeyu Jiang, Yue Zhang, and Mohit Bansal. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024. 1, 2, 3, 7, 8

[9] Yang Jiao, Yan Gao, Jingjing Meng, Jin Shang, and Yi Sun. Learning attribute and class-specific representation duet for fine-grained fashion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[10] Fanheng Kong, Jingyuan Zhang, Yahui Liu, Hongzhi Zhang, Shi Feng, Xiaocui Yang, Daling Wang, Yu Tian, Victoria W., Fuzheng Zhang, and Guorui Zhou. Modality curation: Building universal embeddings for advanced multimodal information retrieval. *arXiv preprint arXiv:2505.19650*, 2025. 1

[11] Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023. 1

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[13] Leheng Li, Weichao Qiu, Xu Yan, Jing He, Kaiqiang Zhou, Yingjie Cai, Qing Lian, Bingbing Liu, and Ying-Cong Chen. Omnibooth: Learning latent control for image synthesis with multi-modal instruction. *arXiv preprint arXiv:2410.04932*, 2024. 1

[14] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[15] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*, 2024. 1

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 1

[17] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[18] Zhengyuan Liu, Xuewen Lin, Guanglu Song, and Hongsheng Chen. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2175–2185, 2021. 1

[19] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. CIRR dataset. 3

[20] Qi Lv, Hao Li, Xiang Deng, Rui Shao, Michael Yu Wang, and Liqiang Nie. Robomp²: A robotic multimodal perception-planning framework with multimodal large language models. *arXiv preprint arXiv:2404.04929*, 2024. 1

[21] Zhe Ma, Jianfeng Dong, Zhongzi Long, Yao Zhang, Yuan He, Hui Xue, and Shouling Ji. Fine-grained fashion similarity learning by attribute-specific embedding network. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2

[22] Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025. 1, 7, 8

[23] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li,

Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

[24] Yu Pan, Li-Jun Liu, Xiao-Bing Yang, Wei Peng, et al. Chest radiology report generation based on cross-modal multi-scale feature fusion. *Journal of Radiation Research and Applied Sciences*, 17(1):100823, 2024. 1

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR, 2021. 1, 2, 7

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1

[27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS) 35*, 2022. 1

[28] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3: Self-supervised large visual models for vision at unprecedented scale. *arXiv preprint arXiv:2508.10104*, 2025. 7

[29] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[30] Shyamgopal Karthik Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21719–21729, 2023. 1

[31] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1781–1789, 2017. 2

[32] Ziyang Wang, Heba Elfardy, Markus Dreyer, Kevin Small, and Mohit Bansal. Unified embeddings for multimodal retrieval via frozen llms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1537–1547, 2024. 1

[33] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021. 1

[34] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogério Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[35] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18102–18112, 2022. 1, 2

[36] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 7

[37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023. 1

[38] Wenzheng Zhao, Kruthika Gangaraju, and Fengpei Yuan. Multimodal perception-driven decision-making for human-robot interaction: a survey. *Frontiers in Robotics and AI*, 2025. 1

[39] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 4

[40] Jinguo Zhu et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 2, 8