Towards Text-Guided Attribute-Disentangled Multimodal Representation Learning

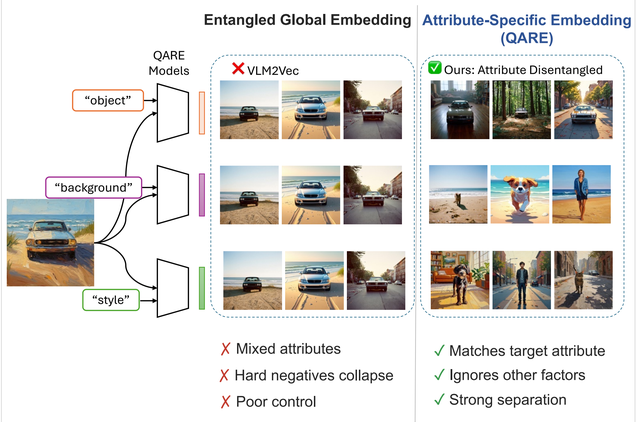
Yibing Wei[1], Sudeep Katakol[1], Manuel Brack[2], Jinhong Lin[1], Haoyue Bai[1], Yu-Teng Li[2], Richard Zhang[2], Eli Shechtman[2], Hareesh Ravi[2], Ajinkya Kale[2]

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

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation & Task

- Existing multimodal embeddings are predominantly **global** and **entangled**
- They mix different attribute (e.g **object, background, and style**), limiting fine-grained retrieval, controllable editing, and attribute-level reasoning.
- We study **Queryable Attribute Representation Extraction (QARE):** given an image I and attribute a, compute $E(I, a)$ that is **sensitive** to the queried attribute and **invariant** to others.
- **Goal: precise, controllable** attribute-specific embeddings.

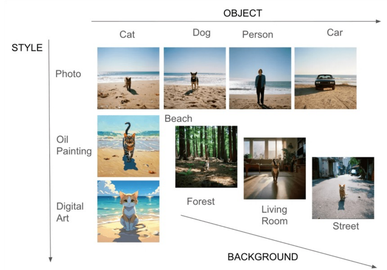


Task: Given image I and attribute a, compute $E(I, a)$ that is **sensitive to a** and **invariant** to others.

QARE-Bench

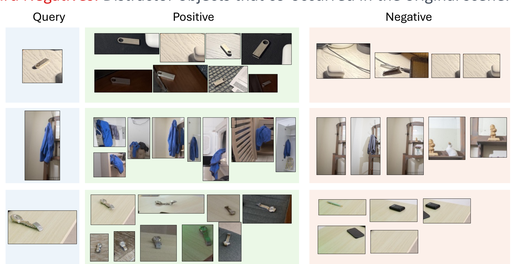
Synthetic Set (Controlled Combinations)

Attribute: object, background, style
192 attribute-conditioned test instances

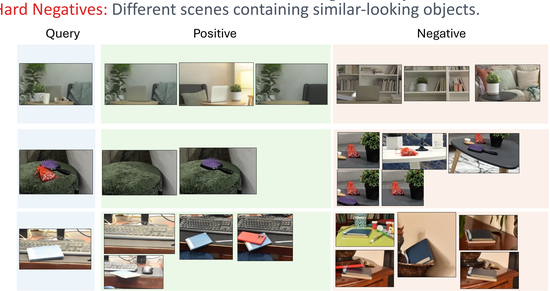


Real Set (Hard by Design)

Object Query Groups: 6,184 crops across 325 unique object instances.
Positives: Identical object instance across varied contexts, poses, and lighting.
Hard Negatives: Distractor objects that co-occurred in the original scene.



Background Query Groups: 2,758 crops across 243 background scenes.
Positives: Exact same scene with different foreground objects.
Hard Negatives: Different scenes containing similar-looking objects.



Evaluation Protocol

mAP (Mean Average Precision)

Attribute-conditioned retrieval. Rank by cosine similarity of $E(I, a)$. Robust to varying numbers of positives.

AIS (Average Intra-Image Similarity)

Average similarity among different attributes of the same image. Lower AIS means stronger disentanglement and query specificity.

Good QARE models: high mAP and low AIS.

TF-QARE (Training-Free Extraction)

✓ Training-free extraction from a frozen VLM. ✓ Attribute-focused prompts isolate object, background, or style.

Main Results

Our framework boosts retrieval: synthetic all mAP up to 83.3, real all mAP up to 65.6; AIS drops to 0.55. We consistently outperform VLM2Vec and global encoders, proving text-guided extraction unlocks latent attribute structure across Qwen, InternVL, and Gemma backbones.
Scaling helps—InternVL3 rises from 45.6 to 77.6 synthetic mAP—yet ultra-large models can pay an alignment tax, trading rigid descriptions for dialogue. Mid-scale VLMs often hit the sweet spot.

Method	Backbone	QARE-Bench Synthetic mAP (↑) obj	sty	bg	all	AIS (↓)	QARE-Bench Real mAP (↑) obj	bg	all	AIS (↓)
(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

Where It Work Best?

Layer ablations reveal **penultimate decoder layer** give peak separation (Qwen2-VL-7B).

	layer	Syn. mAP 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

Key Takeaways

- ✓ Training-free extraction from frozen VLMs outperforms fine-tuned and global baselines.
- ✓ Frozen VLMs encodes rich attribute-specific signals.
- ✓ Prompt-guided extraction is a simple and powerful alternative to fine-tuning.
- ✗ Training-free extraction is effective but insufficient for complete disentanglement.

Future Direction: Efficient VLMs fine-tuning for sharpening attribute disentanglement without forgetting pretrained multimodal knowledge