STUCK ON SHUFFLE: ON ATTRIBUTE–OBJECT BINDING IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

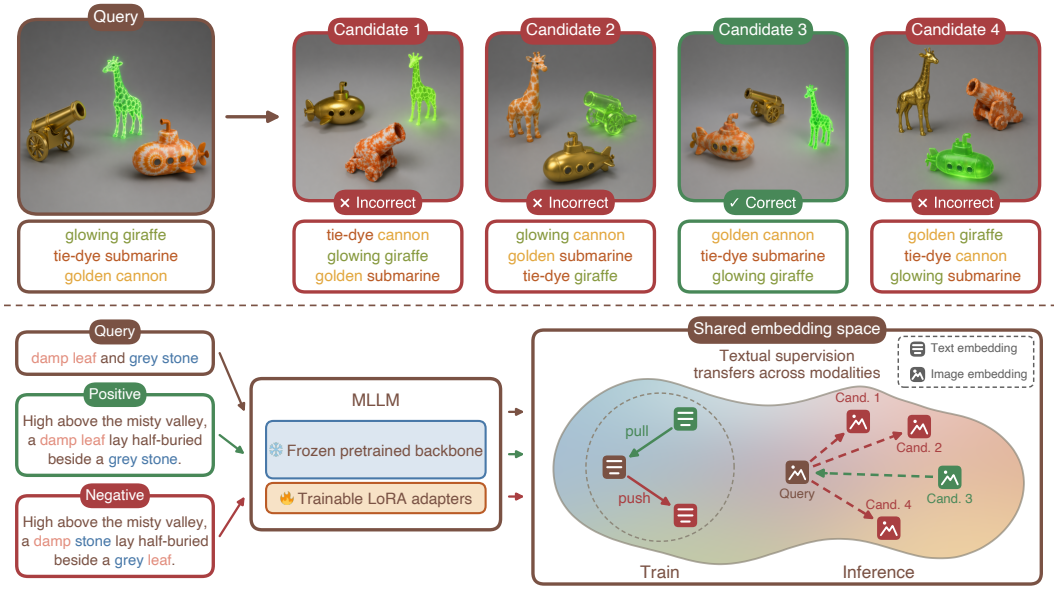

Figure 1: **Overview. Top:** SABER evaluates attribute–object binding by reassigning fixed sets of objects and attributes. **Bottom:** Synthetic text-only finetuning improves binding in text and image representations, boosting unimodal and cross-modal retrieval on SABER and existing benchmarks

ABSTRACT

Retrieval is a fundamental task in multimodal systems: given an image or text query, the goal is to find the most relevant images or texts in a collection, often using a shared embedding space. Reliable retrieval requires these embeddings to capture not only which concepts are present, but also how they are composed. One key compositional challenge is *attribute–object binding*: encoding which attributes belong to which objects. Existing benchmarks provide a limited view of this capability, typically covering a subset of modality pairings, simple two-object swaps, or small attribute sets. Consequently, it remains unclear how robustly current models preserve attribute–object bindings across different modality pairings and more complex settings. We therefore introduce **SABER**, a controlled benchmark for evaluating attribute–object binding across all four image and text modality pairings, spanning diverse attribute types and scenes with up to ten objects. Evaluating models on SABER reveals substantial binding failures in both CLIP-style dual encoders and recent MLLM-based embedders. These findings motivate us to ask how binding can be improved efficiently. While multimodal supervision may be effective, collecting suitable images is difficult and costly. Text, on the other hand, is well suited to binding supervision: it is cheap to generate, supports a broader range of attributes and enables precisely controlled hard negatives. Recent work also shows that capabilities learned through text-only supervision can transfer to visual modalities in MLLMs. We thus finetune an MLLM-based embedder exclusively on synthetic attribute-binding text. Despite the absence of image supervision, the gains transfer to cross-modal and image–image retrieval while preserving general multimodal embedding performance. Our model achieves SOTA results on SABER while also improving on negation and instance-retrieval benchmarks, demonstrating broader benefits of our binding-focused supervision.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Multimodal retrieval aims to identify the most relevant images or texts in a large collection for a given image or text query. Modern systems typically use vision–language models (VLMs) to embed different modalities in a shared space (Radford et al., 2021; Jiang et al., 2025). Retrieval quality depends on these embeddings encoding both which concepts are present and how those concepts are composed. By encoding collection items once and storing their embeddings, this approach can scale to large collections through efficient similarity search (Johnson et al., 2019). In contrast, pairwise models must jointly process the query with each candidate, making them prohibitively expensive at large scale (Reimers & Gurevych, 2019).

One important aspect of composition is attribute–object binding: associating each attribute with the correct object (Thrush et al., 2022; Lewis et al., 2024). As illustrated in Figure 1 (top), given a query containing a *golden cannon*, *tie-dye submarine*, and *glowing giraffe*, a model should retrieve candidates preserving these bindings rather than those with the same objects and attributes paired differently. Yet VLM embeddings often behave like bag-of-words matchers, capturing which concepts are present without reliably preserving their bindings (Yuksekgonul et al., 2023).

However, existing evidence provides only a partial picture of these binding failures, since current benchmarks cover a limited range of settings. For example, many benchmarks evaluate only a single modality pairing, such as image queries with text candidates (Yuksekgonul et al., 2023; Hsieh et al., 2023), while others use a limited attribute vocabulary, often focusing on color (Tang et al., 2023; Burapacheep et al., 2024), and most focus on simple two-object attribute swaps (Dumpala et al., 2024; Koishigarina et al., 2026). This makes it difficult to determine whether these binding failures are general or specific to particular modalities, scene complexity, or attribute types.

To study these questions systematically, we introduce **SABER** (***S**ynthetic **A**ttribute **B**inding **E**valuation for **R**epresentations*), a controlled synthetic benchmark that isolates attribute–object binding by holding the set of objects and attributes fixed while varying their assignments. SABER supports unimodal and cross-modal retrieval over images and text, with broad object and attribute vocabularies and scenes containing up to ten objects. We use SABER to evaluate dual-encoder VLMs, such as CLIP (Radford et al., 2021), as well as newer embedding models derived from multimodal large language models (MLLMs), such as Qwen3-VL-Embedding (Li et al., 2026). Our evaluation on SABER reveals substantial attribute-binding failures across both model families, all four modality pairings, and all levels of scene complexity.

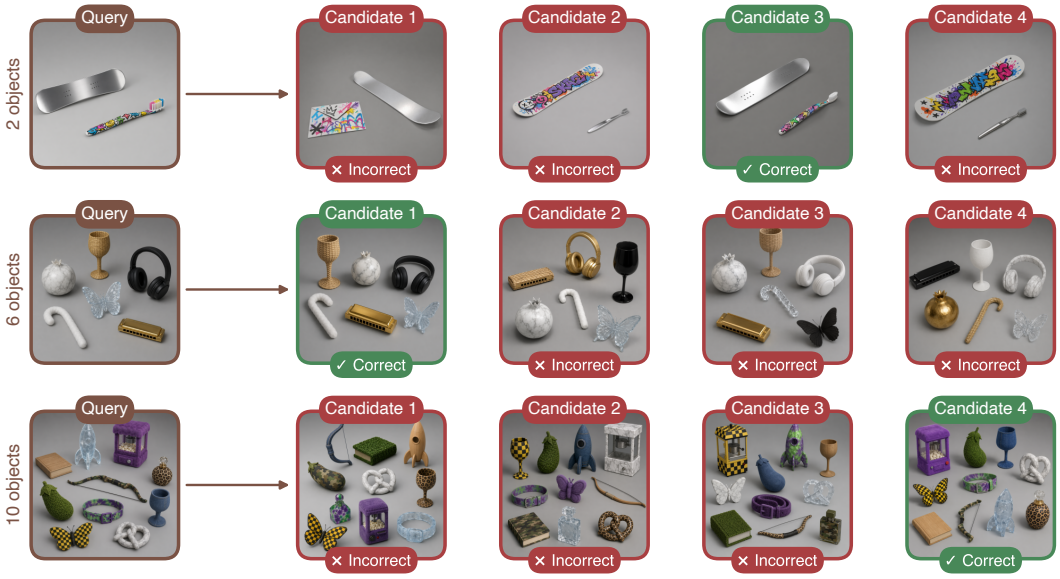

Figure 2: **SABER multiple-choice questions.** Each query has four candidates with 2–10 attribute–object pairs: one preserves the bindings, while the others form hard negatives by reassigning attributes. For two-object examples, one negative instead replaces an object or attribute. Although only images are shown, the accompanying captions enable evaluation within and across modalities.

We next ask whether these failures can be reduced through targeted binding supervision. Direct multimodal supervision is a natural option, but constructing controlled images with the same concepts but different attribute–object assignments is difficult and costly. Text offers a simpler alternative: it is cheap to generate, can express visual and non-visual attributes, makes attribute–object bindings explicit rather than implicit in pixels, and enables precisely controlled hard negatives by reassigning attributes among objects. Furthermore, recent studies on MLLMs show that capabilities acquired through text-only supervision can transfer to other modalities (Jiang et al., 2024; Li et al., 2025; Hu et al., 2025). Together, these observations motivate us to ask whether refining a pretrained MLLM-based multimodal embedder for a specific capability using text alone can transfer improvements across modalities. We test this by finetuning a pretrained multimodal embedder exclusively on synthetic attribute–binding text (Figure 1, bottom).

Despite receiving no image supervision during finetuning, our model achieves the best overall performance on SABER, substantially outperforming all evaluated baselines across all four modality pairings. The gains extend beyond SABER: our model outperforms existing models on all 9 attribute-binding benchmarks tested, spanning natural and synthetic images and diverse text sources, while largely preserving general multimodal embedding performance. Beyond attribute binding, finetuning substantially improves performance on both negation and instance retrieval benchmarks, suggesting that binding-focused supervision can yield broader improvements beyond its original training objective. We will release SABER, our textual finetuning dataset, and our finetuned models to support future research on compositionality and attribute binding.

2 Related Work

2.1 Multimodal Embedding Models

Vision–language embedding models have traditionally relied on dual encoders trained contrastively on web-scale image–caption pairs, as popularized by CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). Subsequent work has advanced this paradigm through greater scale (Cherti et al., 2023), improved data curation and training recipes (Gadre et al., 2023; Sun et al., 2023), and alternative objectives such as SigLIP's sigmoid loss (Zhai et al., 2023).

More recently, pretrained MLLMs have been adapted into task-conditioned embedding models for multimodal retrieval. VLM2Vec (Jiang et al., 2025) uses multitask contrastive training to convert a pretrained MLLM into an embedding model. LamRA-Ret (Liu et al., 2025) adds language-only retrieval pretraining before finetuning on multimodal retrieval tasks, improving retrieval performance. Qwen3-VL-Embedding (Li et al., 2026) combines large-scale contrastive pretraining with distillation from a reranker, which jointly processes each query–candidate pair and assigns it a relevance score. The MMEB benchmark series (Jiang et al., 2025; Meng et al., 2026; Huang et al., 2026) evaluates such multimodal embedders across diverse retrieval settings and input modalities. Our work evaluates attribute–object binding in both CLIP-style dual encoders and MLLM-based embedders.

2.2 Compositionality in Vision–Language Representations

Prior work has repeatedly found that VLM embeddings exhibit bag-of-words behavior, with limited sensitivity to compositional structure. Winoground (Thrush et al., 2022) illustrates this using captions containing the same words but expressing different compositions, such as *"some plants surrounding a lightbulb"* and *"a lightbulb surrounding some plants,"* which models often struggle to match to the correct images. ARO (Yuksekgonul et al., 2023) and SugarCrepe (Hsieh et al., 2023) probe these failures more systematically using hard negatives that swap or replace objects and attributes, modify relations, or alter word order.

Multiple explanations for this bag-of-words behavior have motivated different mitigation strategies. One line of work attributes these failures to insufficient compositional signal or other biases in standard image–text training data (Doveh et al., 2023; Patel et al., 2024; Gurung et al., 2025). Accordingly, proposed remedies include augmenting the training data with compositional hard negatives (Yuksekgonul et al., 2023; Zhang et al., 2024; Patel et al., 2024) and improving the caption density and image–text alignment (Doveh et al., 2023). However, beyond being costly, automatically constructed multimodal supervision can also be noisy, with artifacts in text perturbations, hallucinations in generated captions, and unintended variation in synthetic images.

A second explanation is that a single global embedding may be too blunt a representation for compositional structure, requiring it to encode diverse information, such as content, appearance, spatial layout, and attribute–object bindings, all within the same vector (Kamath et al., 2023; Kang et al., 2025; Weller et al., 2026). Methods addressing this bottleneck either retain token and patch representations for late-interaction scoring (Kang et al., 2025; Assouel et al., 2025) or use pairwise scorers to directly model interactions between inputs (Krojer et al., 2023; Lin et al., 2024a;b).

We instead use synthetic text-only supervision, where attribute–object bindings can be manipulated precisely and at scale to construct controlled hard negatives, without requiring controlled image generation or changing the standard single-vector architecture and inference procedure.

2.3 Attribute-Binding Benchmarks

Existing attribute-binding benchmarks vary considerably in scope and design. VG-Attribution (Yuksekgonul et al., 2023) and SugarCrepe (Hsieh et al., 2023) use natural images but are limited to image–text retrieval. SugarCrepe++ (Dumpala et al., 2024) extends SugarCrepe with text–text retrieval, while ColorSwap (Burapacheep et al., 2024) and BiVLC-Swap-Att. (Miranda et al., 2024) leverage synthetic images to enable bidirectional image–text retrieval.

More controlled benchmarks isolate attribute–object binding through explicitly constructed scenes with limited vocabularies. NCD (Tang et al., 2023; Anwar et al., 2025) evaluates color binding in scenes containing fruits and vegetables. CLEVR-based benchmarks (Johnson et al., 2017; Koishigarina et al., 2026) use rendered geometric objects and scale to scenes with up to ten bindings, while PUG:SPAR (Bordes et al., 2023) and PUG:SPARE (Koishigarina et al., 2026) systematically vary attributes in rendered animal scenes. Despite this greater control, evaluation remains limited to image–text retrieval. Figure 9 in the Appendix provides representative examples from each benchmark.

3 SABER Benchmark

Existing attribute-binding benchmarks are limited in modality coverage, vocabulary, or the number of bindings per sample, and rely on small candidate-set evaluations without assessing large-gallery retrieval with multiple valid positives (Yuksekgonul et al., 2023; Hsieh et al., 2023; Burapacheep et al., 2024; Miranda et al., 2024; Koishigarina et al., 2026). As a result, they provide only a partial view of how well current representations preserve attribute–object bindings.

To address these limitations, we introduce **SABER** (***S**ynthetic **A**ttribute **B**inding **E**valuation for **R**epresentations*), a benchmark structured as 4-way multiple-choice questions (MCQs), whose candidates can also be pooled for large-gallery retrieval. SABER covers all four image and text modality pairings, draws from sets of 232 everyday objects and 50 visual attributes spanning colors, materials, patterns, and effects, and scales to scenes with up to ten objects. Table 1 summarizes the differences between SABER and prior benchmarks.

Table 1: **Existing attribute-binding benchmarks.** *Open* denotes an unrestricted vocabulary; MCQ includes Winoground-style 2×2 matching and SugarCrepe++'s three-caption evaluation.

	Modalities				Evaluation Method		Benchmark Statistics			
Dataset	**T→T**	**T→I**	**I→T**	**I→I**	**MCQ**	**Retrieval**	**# Eval. Samples**	**Object Vocab.**	**Attribute Vocab.**	**# Bindings / Sample**
NCD	×	×	✓	×	✓	×	1,998	19	6	2
VG-Attribution	×	×	✓	×	✓	×	28,748	*open*	*open*	2
ColorSwap	×	✓	✓	×	✓	×	1,000	*open*	*open*	2
SugarCrepe-Swap-Att.	×	×	✓	×	✓	×	666	*open*	*open*	2
SugarCrepe++-Swap-Att.	✓	×	✓	×	✓	×	666	*open*	*open*	2
BiVLC-Swap-Att.	×	✓	✓	×	✓	×	278	*open*	*open*	2
CLEVR	×	×	✓	×	✓	×	43,519	3	8	2–10
PUG:SPAR	×	×	✓	×	✓	×	19,840	32	4	2
PUG:SPARE	×	×	✓	×	✓	×	88,704	12	8	2
SABER (ours)	✓	✓	✓	✓	✓	✓	1,800	232	50	2–10



Figure 3: **Nested SABER chains.** Nine linked MCQs form a chain from two to ten objects, read left to right and top to bottom. Each successive query adds one binding while preserving all earlier ones. Half of SABER's MCQs are organized into such chains, allowing later correct candidates to serve as additional positives for earlier queries in large-gallery retrieval.

Each of our MCQs, illustrated in Figures 1 and 2, contains a query and four candidates, each represented by both an image and a caption. One candidate preserves the query's attribute–object bindings, while the remaining three are hard negatives that alter these assignments. Except for one negative in the two-object setting (Candidate 1 in Figure 2), all candidates contain the same objects and attributes as the query, isolating binding from other challenges such as object recognition.

The benchmark draws from our sets of 232 everyday objects and 50 visual attributes. Sampled objects and attributes are paired in a templated query caption, such as *"a sandy snowboard and a pink cherry."* This fixed phrasing minimizes linguistic variation unrelated to attribute–object binding. After constructing the corresponding candidate captions, we synthesize the query and candidate images using GPT Image 2 (OpenAI, 2026), with instructions specifying neutral CLEVR-style scenes and consistent realizations of each attribute across all five images in an MCQ. This keeps attribute appearance consistent across candidates, so that their visual differences primarily reflect changes in binding. All generated images are manually inspected to ensure that the intended objects and attribute bindings are rendered correctly, yielding unambiguous MCQs with a unique correct answer.

SABER contains 200 MCQs for each scene size from 2 to 10 objects: 100 independent examples and 100 arranged into nested chains, as illustrated in Figure 3. Each successive query in a chain adds one attribute–object binding while preserving earlier ones. Overall, this yields 1,800 queries and 7,200 candidates, or 9,000 image–caption pairs.

We evaluate models on SABER using two complementary protocols: 4-way MCQ and large-gallery retrieval. In the MCQ setting, the model ranks four candidates by similarity to the query and succeeds when the binding-preserving candidate ranks first. In the retrieval setting, each query is ranked against all 7,200 candidates, with any candidate containing all query bindings treated as positive. Independent queries have one positive, while nested-chain queries may have multiple positives from later scenes that preserve earlier bindings. This reflects practical retrieval, where the queried bindings may appear within a larger, more complex scene. Further benchmark-generation details appear in Appendix A.

4 Text-Only Finetuning for Attribute Binding

To assess the binding capabilities of current embedding models, we evaluate a range of CLIP-style dual encoders (Radford et al., 2021; Bolya et al., 2025), including compositional and attribute-

binding variants such as TripletCLIP (Patel et al., 2024), READ-CLIP (Kwon et al., 2025), and LABCLIP (Koishigarina et al., 2026) together with recent MLLM-based embedders (Liu et al., 2025; Li et al., 2026). Table 2 reveals clear binding failures across existing models on SABER, motivating targeted finetuning to improve binding sensitivity.

We therefore turn to text-only finetuning on synthetic attribute-binding data. Text is particularly well suited to this setting because bindings are explicit and easy to manipulate, enabling precisely controlled hard negatives without the incidental variation that can arise in generated images.

Using Gemma 4 (Google DeepMind, 2026), we construct $200{,}000$ finetuning triplets, each containing a query, a positive, and a hard negative. Each query specifies two or three attribute–object pairs, with the positive then contextualizing those bindings. For example, the query *"red banner and torn flag"* may be paired with *"High above the roar of the crowd, ... there waved a red banner and a torn flag."* We construct a hard negative by permuting either the attributes or the objects, so that it differs from the positive only in its attribute–object assignments. To reduce shortcut cues, we choose attributes that remain plausible for all objects in each example and randomize the order of the attribute–object pairs between the query and the positive.

Although the finetuning triplets are entirely textual, our goal is to improve attribute binding across all text and image modality pairings. More formally, let x_q^m denote a query expressing a set of attribute–object bindings in modality $m \in \{\text{Text}, \text{Image}\}$, and let x_+^n and x_-^n denote a binding-preserving positive and a binding-conflicting negative, respectively, in modality $n \in \{\text{Text}, \text{Image}\}$. Ideally, the embedding model should place the positive closer to the query under every modality pairing:

$$D\big(f_{\boldsymbol{\theta}}(x_q^m), f_{\boldsymbol{\theta}}(x_+^n)\big) > D\big(f_{\boldsymbol{\theta}}(x_q^m), f_{\boldsymbol{\theta}}(x_-^n)\big)\,, \qquad \forall (m,n) \in \{\text{Text}, \text{Image}\}^2,$$

where $f_{\boldsymbol{\theta}}$ is the multimodal embedder and D is cosine similarity in the shared embedding space.

We use text-only supervision to encourage this behavior. Given a textual query t_q, a positive t_+, and a hard negative t_-, we minimize the following triplet loss (Schroff et al., 2015)

$$\mathcal{L}_{\text{text}} = \left[D(f_{\boldsymbol{\theta}}(t_q), f_{\boldsymbol{\theta}}(t_-)) - D(f_{\boldsymbol{\theta}}(t_q), f_{\boldsymbol{\theta}}(t_+)) + \gamma\right]_+ ,$$

where γ is the triplet margin and $[z]_+ = \max(z, 0)$ denotes the positive-part function.

Unlike CLIP-style dual encoders with separate image and text encoders, MLLMs contain parameters that are shared across modalities. Although we optimize only the textual loss above, updating these shared parameters from textual triplets also affects the representations produced for images, providing a mechanism for the learned binding behavior to transfer to the other modality. We finetune Qwen3-VL-Embedding-2B and Qwen3-VL-Embedding-8B (Li et al., 2026) with LoRA (Hu et al., 2022) using this objective and subsequently apply LiNeS (Wang et al., 2025) to rescale the layer-wise parameter updates and mitigate forgetting of pretrained capabilities.

Appendix B provides additional details on the finetuning dataset and optimization procedure. Appendix C ablates the finetuning data and LiNeS and analyzes attention changes before and after finetuning. Appendix D shows that our finetuning recipe extends to other MLLM backbones. Appendix E compares finetuning with text versus images on SABER and shows that textual supervision yields stronger cross-modal retrieval performance.

5 Experiments

5.1 Can Text-Only Finetuning Improve Attribute Binding Across Modalities?

Table 2 compares our finetuned models with several baselines, on SABER under both 4-way multiple-choice questions and large-gallery retrieval. For all models, candidates are ranked by cosine similarity between the corresponding query and candidate embeddings. We report results for all four image and text modality pairings, with retrieval evaluated separately on the independent and nested queries. Among the pretrained models, Qwen3-VL-Embedding-8B achieves the strongest mean performance.

Our textual finetuning substantially improves MCQ performance on SABER across all modality pairings. Our finetuned 8B model reaches a mean accuracy of 84.2%, up from 52.5% for Qwen3-VL-Embedding-8B, with near-perfect text–text accuracy of 99.3%.

Table 2: **Performance on SABER.** SABER-MCQ reports mean 4-way accuracy (%) and SABER-Retrieval mean mAP (%), both averaged over 2–10 objects. Retrieval is split into independent and nested subsets over the pooled candidate gallery. Our finetuned 8B model performs best across all four modality pairings in both protocols. Results by number of objects are reported in Tables 17, 18, and 19 in the Appendix. Best and second-best results are shown in bold and underlined, respectively.

	SABER-MCQ					SABER-Retrieval (Independent)					SABER-Retrieval (Nested)				
Model	**T→T**	**T→I**	**I→T**	**I→I**	**Mean**	**T→T**	**T→I**	**I→T**	**I→I**	**Mean**	**T→T**	**T→I**	**I→T**	**I→I**	**Mean**
CLIP ViT-B/32	53.6	23.0	22.3	75.5	43.6	74.0	38.8	33.1	85.1	57.8	38.0	13.7	16.1	49.3	29.3
LABCLIP ViT-B/32	55.2	22.8	24.2	75.3	44.4	74.3	35.0	25.8	85.1	55.0	38.1	13.0	12.1	49.2	28.1
TripletCLIP ViT-B/32	42.9	27.1	22.2	60.2	38.1	65.9	8.8	5.0	63.2	35.7	35.8	5.4	3.5	32.1	19.2
READ-CLIP ViT-B/32	47.3	23.9	23.7	82.3	44.3	68.4	33.5	23.6	89.7	53.8	37.7	10.7	13.1	54.1	28.9
PE-Core-bigG-14-448	66.1	29.7	27.6	67.6	47.7	81.1	56.0	55.5	81.5	68.5	41.5	23.6	32.9	47.1	36.3
LamRA-Ret	49.1	30.6	27.0	64.9	42.9	70.4	56.8	51.5	79.2	64.5	37.3	27.3	28.0	43.7	34.1
Qwen3-VL-Embedding-2B	35.0	27.8	26.4	73.9	40.8	60.1	54.0	55.0	85.5	63.6	34.0	28.1	30.9	47.3	35.1
Qwen3-VL-Embedding-8B	53.7	34.9	36.3	84.9	52.5	73.5	59.6	62.6	91.9	71.9	37.8	33.1	29.9	52.3	38.3
Ours-2B	76.9	34.8	32.9	80.9	56.4	87.8	59.3	58.3	89.8	73.8	46.2	29.4	32.5	49.9	39.5
Ours-8B	**99.3**	**69.0**	**74.1**	**94.3**	**84.2**	**99.6**	**82.3**	**85.5**	**97.1**	**91.1**	**65.5**	**46.3**	**46.0**	**58.6**	**54.1**

Despite using no visual data during finetuning, it gains 34.1, 37.8, and 9.4 points in the text–image, image–text, and image–image settings, respectively. Our 2B model shows the same pattern, improving over its pretrained counterpart across all four pairings, with an average gain of 15.6 points.

In independent retrieval, models must find the binding-preserving match within the full 7,200-candidate gallery, containing both binding-swapped distractors and thousands of unrelated scenes. For all models except the CLIP-based ViT-B/32 variants, the correct candidate and its three binding-swapped distractors occupy the top four retrieved positions in over 90% of queries, so retrieval is largely governed by the same binding comparison as the MCQ task. For the CLIP-based models, this occurs in only 27–53% of queries overall, with the failure concentrated in the cross-modal pairings. Unrelated scenes are therefore frequently retrieved instead, revealing an additional failure mode not captured by MCQ. Our 8B model again performs best, improving mean mAP by 19.2 points over Qwen3-VL-Embedding-8B.

Retrieval performance drops substantially on the nested subset for all models, reflecting the more demanding setting in which queries may match larger scenes containing additional bindings. Our 8B model nevertheless performs best across all four modality pairings, reaching 54.1% mean mAP compared with 38.3% for Qwen3-VL-Embedding-8B. The nested setting exposes a challenge absent from MCQ and independent retrieval: *containment*, or recognizing valid supersets rather than only exact matches. Across models, candidates containing only a subset of the query bindings rank above the exact match in only 3.3–17.1% of comparisons on average across modalities, but above candidates that contain all query bindings plus additional ones in 33–39%. These results highlight containment as a remaining challenge distinct from binding-swap discrimination.

Tables 17, 18, and 19 in the Appendix provide results broken down by the number of attribute–object bindings. Notably, the improvements extend to scenes with up to ten objects, even though our finetuning triplets permute bindings among only two or three attribute–object pairs.

5.2 Do the Gains Generalize Beyond SABER?

We next test whether the improvements observed on SABER transfer to existing attribute-binding benchmarks. Table 3 evaluates 9 existing benchmarks across 12 retrieval settings spanning natural and synthetic images as well as human-written, LLM-generated, and templated text. We report accuracy under each benchmark's native evaluation protocol. Our finetuned 8B model outperforms every baseline across all 12 settings, improving over Qwen3-VL-Embedding-8B by 15.4 points on average. Relative to Qwen3-VL-Embedding-8B, gains reach 22.8 points in the SugarCrepe++ text–text setting, 33.9 points in the ColorSwap text–image setting, and 25.6 points in the CLEVR image–text setting. Our finetuned 2B model likewise improves over its pretrained counterpart by 8.2 points on average. These results indicate that the learned improvements in attribute binding are not limited to SABER's synthetic scenes or templated text, nor to the structure of the finetuning data, where short queries are matched against contextualized positives and hard negatives.

Table 3: **Performance on existing attribute-binding benchmarks.** Accuracy (%) under each benchmark's native multiple-choice protocol. Best and second-best results are shown in bold and underlined, respectively.

Dataset	CLIP B/32	LABCLIP B/32	TripletCLIP B/32	READ-CLIP B/32	PE CORE-G	LamRA-Ret	Qwen3-VL Emb-2B	Qwen3-VL Emb-8B	Ours-2B	Ours-8B
Text→Text										
SugarCrepe++-Swap-Att.	32.7	41.0	34.7	56.8	53.2	66.2	50.3	70.9	<u>78.7</u>	**93.7**
Text→Image										
ColorSwap	17.9	15.5	14.0	24.5	44.6	23.1	22.8	<u>45.9</u>	34.2	**79.8**
BiVLC-Swap-Att.	19.1	24.1	7.2	<u>31.7</u>	19.4	21.2	15.1	21.6	16.2	**32.0**
Image→Text										
NCD	71.1	66.2	62.7	63.2	70.2	48.3	52.5	<u>80.6</u>	60.8	**95.4**
VG-Attribution	62.5	69.3	57.6	<u>80.1</u>	67.9	76.4	67.0	78.0	67.0	**80.4**
ColorSwap	38.2	29.3	39.6	34.2	67.1	46.4	47.6	<u>71.4</u>	56.7	**88.4**
SugarCrepe-Swap-Att.	64.0	74.8	64.3	82.7	78.2	79.7	77.2	<u>85.7</u>	80.3	**90.8**
SugarCrepe++-Swap-Att.	44.9	53.9	42.0	66.2	59.5	62.5	55.1	65.8	<u>67.9</u>	**82.7**
BiVLC-Swap-Att.	47.1	43.9	38.1	43.9	69.1	56.8	55.0	<u>73.7</u>	59.0	**74.8**
CLEVR	50.3	50.6	51.0	51.1	51.4	52.5	52.1	60.6	<u>64.1</u>	**86.2**
PUG:SPAR	47.5	52.5	65.7	43.0	46.8	40.9	61.1	48.9	**70.3**	<u>67.7</u>
PUG:SPARE	50.4	50.4	50.5	51.5	50.8	51.6	<u>53.7</u>	52.8	52.3	**69.1**

5.3 How Much General Multimodal Performance Is Lost After Finetuning?

MLLM-based embedders combine modality-specific input components with a largely shared backbone, mapping inputs from multiple modalities into a common embedding space. This enables a single model to support a broad range of retrieval tasks. A natural concern is that binding-focused finetuning may compromise this broader multimodal capability. We evaluate this trade-off on the non-audio portion of MMEB-v3 (Huang et al., 2026), spanning 182 image, video, visual-document, text, and agent tasks. As shown in Table 4, despite using no MMEB-v3 data during finetuning, aggregate performance changes by only -0.5 points for the 2B model and $+0.5$ points for the 8B model. Performance also remains similar across domains, with modest declines on image, video, and text tasks, mixed changes on visual-document tasks, and gains on agent tasks.

Table 4: **MMEB-v3 results.** Average Hit@1 is reported for image, video, and agent tasks and NDCG@5 for VisDoc and text tasks. Our finetuned models remain within 0.5 points of their pretrained counterparts overall. Best results per pretrained–finetuned pair are shown in bold.

Model	Image	Video	VisDoc	Text	Agent	All
# of Datasets	37	18	27[1]	53	47	182
Qwen3-VL-Embedding-2B	**69.7**	**56.3**	**74.2**	**40.3**	39.8	**52.8**
Ours-2B	67.7	56.1	73.0	39.7	**41.1**	52.3
Qwen3-VL-Embedding-8B	**73.7**	**59.6**	74.3	**43.7**	39.1	54.7
Ours-8B	72.5	58.7	**75.4**	43.1	**42.3**	**55.2**

6 From Attribute Binding to Negation

Attribute binding is only one form of compositional sensitivity. To assess whether binding-focused finetuning improves compositional sensitivity more broadly, we evaluate our models on NegBench (Alhamoud et al., 2025), which tests negation understanding. We compare the base Qwen3-VL-Embedding-8B, our finetuned model, and a variant trained after removing the 7.1% of finetuning examples containing negation terms such as *no*, *without*, and *neither*. This ablation tests whether gains arise from incidental exposure to negation rather than transfer from binding supervision. We also include Qwen3-VL-Reranker-8B, a stronger pairwise model, as an additional reference.

NegBench groups MCQs according to the type of the correct statement: *affirmation* describes present objects, *negation* absent objects, and *hybrid* both. Each MCQ contains the correct answer together with one additional candidate of each type. Table 5 shows that finetuning improves overall accuracy by 9.3 points on the COCO (Lin et al., 2014) subset and 10.5 points on the PASCAL

[1]MMEB-v3 reports 24 VisDoc tasks but enumerates 27; we score all 27 for every model.

VOC 2007 (Everingham et al., 2010) subset, driven mainly by large gains on hybrid cases. In these cases, the model must distinguish candidates such as *"bowl but no cup"* and *"cup but no bowl,"* which share the same bag of words but assign negation to different objects. This closely resembles our attribute-binding setup, where correctness depends on associating the correct property with each object. By contrast, in the affirmation and negation groups, the two same-type candidates contain different bags of words. Notably, the negation-filtered variant retains nearly all of these gains, suggesting transfer from binding supervision rather than direct negation exposure. Qwen3-VL-Reranker-8B remains stronger on pure-negation cases, but is comparable overall to our model.

Table 5: **NegBench results.** MCQ accuracy (%) by correct-statement type on the COCO and VOC 2007 subsets. Our finetuned model improves overall performance, with the largest gains on hybrid statements. Best results are shown in bold.

	NegBench$_{coco}$				NegBench$_{voc}$			
Model	**Affirmation**	**Negation**	**Hybrid**	**All**	**Affirmation**	**Negation**	**Hybrid**	**All**
Qwen3-VL-Embedding-8B	78.3	6.8	36.3	41.4	89.9	4.7	58.0	39.8
Ours-8B (negation-filtered)	82.7	6.6	58.1	50.3	91.5	6.4	78.2	49.7
Ours-8B	**83.4**	6.8	**58.5**	**50.7**	91.8	6.1	**79.6**	50.3
Qwen3-VL-Reranker-8B	80.6	**17.8**	46.9	49.3	**95.0**	**28.1**	58.0	**50.4**

7 Broader Benefits: Instance-Level Retrieval

We next ask whether binding-focused finetuning benefits instance retrieval, where the goal is to retrieve the same object instance rather than its category. Instances are often distinguished by their attributes, such as the color of a shirt, suggesting that improved binding could strengthen instance-level representations. We therefore evaluate Qwen3-VL-Embedding-8B and our finetuned model on ILIAS (Kordopatis-Zilos et al., 2025), INSTRE (Wang & Jiang, 2015), PODS (Sundaram et al., 2025), Paris6K (Philbin et al., 2008), Oxford5K (Philbin et al., 2007), and Sculptures6K (Arandjelović & Zisserman, 2011), covering eight image–image settings and one text–image setting.

Table 6 reports mAP across all settings. Our finetuned model improves over the pretrained model in every setting, with gains ranging from $+2.1$ to $+9.9$ mAP. To examine whether these gains reflect better discrimination among visually similar instances, we evaluate retrieval separately within ILIAS's fine-grained object categories, such as cakes or cards. Within each category, each image is used in turn as a query against the remaining images, with images depicting the same instance treated as relevant. Our model improves over the pretrained model in all 38 fine-grained categories, with a mean mAP increase of 8.5 points. Appendix F provides qualitative instance retrieval examples, as well as the full category-level analysis.

Table 6: **Instance retrieval performance.** Results report mean mAP (%) across image–image and text–image benchmarks for Qwen3-VL-Embedding-8B and our finetuned model. For INSTRE, S and M denote the single- and multi-object subsets, respectively. Best results are shown in bold.

	I→I								T→I
Model	**ILIAS**	**INSTRE S→S**	**INSTRE M→S**	**INSTRE M→S+M**	**PODS**	**Paris6K**	**Oxford5K**	**Sculptures6K**	**ILIAS**
Qwen3-VL-Emb-8B	58.5	92.8	81.3	75.6	85.7	86.8	61.3	50.8	52.3
Ours-8B	**68.4**	**94.9**	**86.7**	**79.0**	**88.1**	**88.9**	**64.3**	**59.6**	**60.8**

8 Discussion

Our results reveal a gap between broad retrieval capability and precise attribute–object binding. We introduce SABER to characterize these failures and show that targeted text-only finetuning improves binding across modality pairings, and that these gains transfer to independent benchmarks, negation, and instance retrieval, while largely preserving general multimodal performance. In Appendix G, we show these gains extend beyond attribute swaps to a partial-description formulation. Appendix H shows that supplying the embedding model with binding-specific instructions helps but does not match the gains from our finetuning. A key limitation is that SABER is synthetic, and generated images may contain unintended variation in appearance, layout, object size, or orientation beyond the intended binding changes. Taken together, our findings suggest that compositional precision is a distinct dimension of embedding quality that warrants explicit evaluation and optimization.

AI USE STATEMENT

Generative AI tools were used for data generation and in several experimental components of this work. Gemma 4 was used to generate synthetic text examples for our finetuning dataset, while GPT-Image-2 was used to generate the synthetic SABER images from author-defined object and attribute sets. Multiple vision–language models were also used to propose and refine candidate instructions for the instruction-sensitivity experiment; the authors reviewed and edited these instructions, selected the final set, and conducted all evaluations.

Generative AI tools were additionally used to support code implementation, figure design, and the drafting and editing of portions of the manuscript. AI tools were also used to assist with literature search and the identification of related work; all referenced works were independently verified and read by the authors. All AI-assisted code, figures, experimental materials, and manuscript text were reviewed and approved by the authors.

These tools did not independently formulate the research questions, design the methodology, interpret the results, or determine the conclusions. The authors take full responsibility for all aspects of this work.

ETHICS STATEMENT

This work does not involve the collection of new human-subject data and uses only existing research benchmarks and synthetically generated data. Some third-party benchmarks contain publicly sourced images, but we do not use identity information or analyze personal attributes. Attribute-binding performance should not be interpreted as evidence of broader fairness, safety, or suitability for deployment. Our results may inherit biases and limitations from the pretrained embedding models, generative models, and datasets used in this work. Licenses and applicable usage terms for third-party datasets and models are reported in Table 20 in the Appendix.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the SABER benchmark construction, textual finetuning data generation, training procedure, and evaluation protocols in the main paper and appendices. We will release SABER, our textual finetuning dataset, and the weights of our finetuned models to facilitate reproducibility and support future research on compositionality and attribute binding.

REFERENCES

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29612–29622, 2025.

Saeed Anwar, Muhammad Tahir, Chongyi Li, Ajmal Mian, Fahad Shahbaz Khan, and Abdul Wahab Muzaffar. Image colorization: A survey and dataset. *Information Fusion*, 114:102720, 2025.

Relja Arandjelović and Andrew Zisserman. Smooth object retrieval using a bag of boundaries. In *2011 International Conference on Computer Vision*, pp. 375–382. IEEE, 2011.

Rim Assouel, Pietro Astolfi, Florian Bordes, Michal Drozdzal, and Adriana Romero-Soriano. Object-centric binding in contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 38:131418–131458, 2025.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv:2504.13181*, 2025.

Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari S. Morcos. Pug: Photorealistic and semantically controllable synthetic data for representation learning, 2023.

Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. Colorswap: A color and word order dataset for multimodal evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1716–1726, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. *Advances in Neural Information Processing Systems*, 36:76137–76150, 2023.

Sri H Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *Advances in Neural Information Processing Systems*, 37:17972–18018, 2024.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Google DeepMind. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Bijay Gurung, David T Hoffmann, and Thomas Brox. Common data properties limit object-attribute binding in clip. In *DAGM German Conference on Pattern Recognition*, pp. 353–368. Springer, 2025.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zhe Hu, Jing Li, Zhongzhu Pu, Hou Pong Ken Chan, and Yu Yin. Praxis-vlm: Vision-grounded decision making via text-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 38:2305–2331, 2025.

Haohang Huang, Xuan Lu, Mingyi Su, Xuan Zhang, Ziyan Jiang, Ping Nie, Kai Zou, Tomas Pfister, Wenhu Chen, Wei Zhang, Xiaoyu Shen, and Rui Meng. MMEB-v3: Measuring the performance gaps of omni-modality embedding models. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=EgNjgS4yOh.

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

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TE0KOzWYAF.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547, 2019.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4933–4944, 2023.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is clip ideal? no. can we fix it? yes! In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22436–22446, 2025.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. Clip behaves like a bag-of-words model cross-modally but not uni-modally. In *International Conference on Learning Representations*, volume 2026, pp. 96106–96127, 2026.

Giorgos Kordopatis-Zilos, Vladan Stojnić, Anna Manko, Pavel Šuma, Nikolaos-Antonios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiří Matas, Ondřej Chum, and Giorgos Tolias. ILIAS: Instance-level image retrieval at scale. In *Computer Vision and Pattern Recognition (CVPR)*, 2025.

Benno Krojer, Elinor Poole-Dayan, Vikram Voleti, Christopher Pal, and Siva Reddy. Are diffusion models vision-and-language reasoners? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jihoon Kwon, Kyle Min, and Jy-yong Sohn. Enhancing compositional reasoning in clip via reconstruction and alignment of text descriptions. *Advances in Neural Information Processing Systems*, 38:13939–13966, 2025.

Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1487–1500, 2024.

Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. Temporal reasoning transfer from text to video. In *International Conference on Learning Representations*, volume 2025, pp. 74070–74102, 2025.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In *Forty-first International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=J5VB1h3Aed.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024b.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4015–4025, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Raghuveer Thirukovalluru, Xuan Zhang, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=TpU38jbKIJ.

Imanol Miranda, Ander Salaberria, Eneko Agirre, and Gorka Azkune. Bivlc: Extending vision-language compositionality evaluation with text-to-image retrieval. *Advances in Neural Information Processing Systems*, 37:101880–101904, 2024.

OpenAI. Introducing ChatGPT Images 2.0, 2026. URL https://openai.com/index/introducing-chatgpt-images-2-0/. Accessed: 2026-07-26.

Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. *Advances in neural information processing systems*, 37:32731–32760, 2024.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8. Ieee, 2007.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8. IEEE, 2008.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Tingyu Song, Mingxin Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Zhijie Nie, Yilun Zhao, and Shu Wu. Uembed: Unified sparse and dense multimodal embeddings, 2026. URL https://arxiv.org/abs/2608.02583.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Shobhita Sundaram, Julia Chae, Yonglong Tian, Sara Beery, and Phillip Isola. Personalized representation from personalized generation. In *International Conference on Learning Representations*, volume 2025, pp. 92572–92606, 2025.

Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and Ilker Yildirim. When are lemons purple? the concept association bias of vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14333–14348, 2023.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Ke Wang, Nikos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, Francois Fleuret, and Pascal Frossard. Lines: Post-training layer scaling prevents forgetting and enhances model merging. In *International Conference on Learning Representations*, volume 2025, pp. 27100–27133, 2025.

Shuang Wang and Shuqiang Jiang. Instre: a new benchmark for instance-level object retrieval and recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(3):1–21, 2015.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=k9CzIvzfaA.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *11th International Conference on Learning Representations, ICLR 2023*, 2023.

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*, pp. 11975–11986, 2023.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13774–13784, 2024.

Junjie Zhou, Ke Mei, Lei Li, Tianyi Wang, Fengyun Rao, and Jing Lyu. Wemm-embedding: Wechat multi-modal embedding technical report. 2026. URL https://arxiv.org/abs/2608.24053.

A SABER BENCHMARK DETAILS

This section provides additional details on SABER's attribute sets, candidate construction, textual realization, image synthesis, and benchmark subsets.

A.1 BUILDING BLOCKS

The benchmark is generated from two lists: a set of 232 everyday objects (e.g., toothbrush, teapot, kayak) and a set of 50 visual attributes (e.g., red, glowing, sandy). The attributes are grouped into 6 semantic categories, summarized in Table 7. In addition to a semantic category, each attribute is assigned a type that determines how it is realized during generation:

- **`standalone`** (27): a fixed appearance with no free slot, such as *camouflage* or *transparent*. Some attributes use a fixed generation form; for example, *soap-bubble* is expressed as *iridescent soap-bubble*.
- **`single-color`** (4): a template with one color slot, e.g., *glowing* $\rightarrow$ $\{color\}$ *glowing*.
- **`two-color`** (6): a template with two distinct color slots, e.g., *striped* $\rightarrow$ $\{color_1\}$ *and* $\{color_2\}$ *striped*.
- **`fixed-variant`** (13): a template instantiated from a small closed set of variants, e.g., *wooden* $\rightarrow$ $\{$*light-brown, dark-brown*$\}$ *wooden*.

Color slots are filled from a palette of nine basic colors: red, orange, yellow, green, blue, purple, pink, white, and black. For two-color attributes, the two sampled colors are always distinct.

Table 7: **Number of SABER attributes in each semantic category.**

Attribute category	#
Basic color	9
Other color	2
Metallic	3
Material	19
Pattern	15
Effect	2
Total	50

A.2 EXAMPLE STRUCTURE

To construct a multiple-choice question (MCQ) consisting of one query and four candidate answers, each with N attribute–object bindings, we sample N distinct objects and N distinct attributes and pair them to form the query. We then derive four candidates, according to the following rules:

- **`candidate 1 (binding-preserving):`** preserves the query bindings while reordering the attribute–object pairs.
- **`candidate 2 (attribute-swap):`** keeps the objects in the same order as candidate 1 but permutes the attributes across them, producing incorrect attribute–object bindings.
- **`candidate 3 (object-swap):`** keeps the attributes in the same order as candidate 1 but permutes the objects across them, producing another incorrect binding assignment.
- **`candidate 4 (joint-swap):`** permutes both the objects and the attributes, producing another incorrect binding assignment.

For $N \geq 3$, the four candidates are required to have pairwise distinct binding assignments; any permutation that duplicates an existing candidate is rejected and resampled. For $N = 2$, where only two one-to-one assignments are possible, candidate 4 instead preserves one query binding and replaces either the object or the attribute in the other binding with a new element not present in the query. For visualization purposes, the candidates are shown in shuffled order in Figures 1 and 2 of the main paper.

A.3 Dual Textual Realization

Each query and candidate is expressed as a conjunction of its attribute–object pairs, such as *"A striped cube and a transparent kite."* Indefinite articles (*a*/*an*) are selected automatically based on pronunciation. For each query and candidate, we produce two parallel textual forms:

- the *text* form, which uses the basic attribute forms (e.g., *a striped cube*) and constitutes the benchmark's text modality;
- the *generation* form, which uses the fully instantiated appearance (e.g., *a white and orange striped cube*) and is used to construct the text-to-image prompt.

Within each example, every attribute is instantiated once and reused across the query and all candidates. For instance, if *striped* is realized as *white and orange striped*, that same realization is used wherever *striped* appears. This keeps each attribute visually consistent across all images and reduces appearance variation unrelated to binding.

A.4 Image Synthesis

Images are synthesized from the generation-form captions of the query and each candidate using GPT-Image-2 at 1024×1024 resolution and medium quality. The rendering prompt requests a CLEVR-style 3D scene on a neutral gray background in which the listed objects are clearly separated and scattered rather than laid out in a row, with each object appearing exactly once. A central requirement of the prompt is *binding locality*: every attribute must apply only to its assigned object. In particular, the prompt specifies that colors, materials, patterns, textures, and visual effects must not transfer, blend, or bleed across objects. To reduce positional shortcuts, each rendered image is horizontally flipped with probability 0.5. Finally, all generated images are manually screened to verify that the depicted attribute–object bindings match their associated captions.

A.5 Benchmark Subsets

For each object count $N \in \{2, \ldots, 10\}$, SABER contains 200 MCQ examples, evenly split between independently sampled examples and nested chains. The independently sampled examples provide broader coverage of object, attribute, and binding combinations. In contrast, each nested chain starts with a two-object example and adds one new attribute–object binding at each subsequent object count, while reshuffling the order of the pairs. Thus, later queries preserve all bindings from earlier queries while introducing additional objects. Their binding-preserving candidates therefore serve as additional positives for earlier queries in large-gallery retrieval.

B Text-Only Finetuning for Attribute Binding

This section provides additional details on the construction of our text-only finetuning dataset, including its seven construction types, diversity controls, and validation procedure, as well as the training split, optimization setup, and evaluation of generalization to held-out binding structures.

B.1 Generation Setup

We generate the finetuning dataset using Gemma 4 31B-IT (Google DeepMind, 2026). For each example, the model selects the objects and attributes and writes a contextualized positive caption expressing the intended attribute–object bindings. From this structured output, we construct the query and two hard negatives. The query lists the attribute–object pairs in randomized order, while the negatives are formed by permuting either the attributes or the objects in the positive caption. This short-query, longer-candidate format reflects practical retrieval, where a concise search query is matched against a gallery of items that may contain the queried content alongside additional information. Thus, the positive and both negatives contain exactly the same words and differ only in their attribute–object bindings. Each example is then automatically validated, with failures discarded and regenerated.

The final dataset contains $100{,}000$ examples distributed across the seven construction types shown in Table 8.

Table 8: **Number of finetuning examples by construction type.**

Construction type	#
pair_swap	20,000
attr_bundle	20,000
predicative	20,000
triple_bind_cycle	10,048
triple_bind_swap	9,952
deferred_bind_anaphoric	9,417
deferred_bind_respective	10,583
Total	100,000

B.2 Construction types

Each example belongs to one of seven construction types, which vary the number of attribute–object pairs and the sentence structure that expresses their bindings. Table 9 shows one example of each type:

- **`pair_swap`**: two objects, each directly preceded by its attribute. The attribute negative exchanges the two attributes, and the object negative exchanges the two nouns.
- **`triple_bind_cycle`**: three objects. The attribute negative rotates the three attributes by one position, and the object negative does the same with the nouns, so all three bindings are wrong in each negative.
- **`triple_bind_swap`**: three objects, but only two pairs are transposed. One of the three pairs, chosen at random, remains correctly bound in both negatives, so the model must locate *which* bindings changed rather than simply detect that something changed.
- **`attr_bundle`**: two objects, each carrying two attributes from different semantic slots (e.g., color and material, or age and condition), which prevents redundant pairs such as *"red crimson."* The attribute negative exchanges the attributes of only *one* slot, so it differs from the positive by a single word per object; the object negative exchanges the nouns.
- **`predicative`**: two objects whose attributes are detached from their nouns through a copula, change-of-state verb, or relative clause (e.g., *"the book was dated"*). The adjacent phrase "⟨attribute⟩ ⟨object⟩" may not appear anywhere in the caption.
- **`deferred_bind_anaphoric`**: both nouns are introduced before either attribute appears, and the attributes are then assigned by back-reference (e.g., *"the former … the latter"* or *"the first … the second"*). In half of the examples the second object is described first, so the reference must actually be resolved. Ambiguous references such as a bare *"it"* or *"they"* are rejected.
- **`deferred_bind_respective`**: both nouns are introduced before either attribute appears, and the attributes are then assigned jointly in noun order (e.g., *"… were new and neat respectively"*), so the binding is carried entirely by word order.

In the **`predicative`** and **`deferred_bind`** types, attributes are not adjacent to their objects, so the correct bindings must be recovered from sentence structure rather than local word co-occurrence.

B.3 Diversity controls

Repeatedly sampling a language model without additional constraints tends to produce a narrow range of vocabulary, styles, and sentence structures. To increase diversity, we therefore sample several properties independently for each example and specify them in the prompt:

- **Context.** Each prompt samples from 1,000 seed scenes, with optional time-of-day and weather modifiers, yielding 81,000 distinct contexts. It additionally samples from 32 writing registers (e.g., news caption, field notebook, auction listing, maintenance log) and 16 sentence-opener styles.

Table 9: **Examples of the seven finetuning construction types.** Each block shows a query, positive caption, and two hard negatives formed by permuting attributes or objects. Bold marks words moved relative to the positive.

`pair_swap`	
Query	wrinkled tunic and cotton shroud
Positive	"Frost is coming," he noted, touching the cotton shroud and wrinkled tunic.
Neg. (attr.)	"Frost is coming," he noted, touching the **wrinkled** shroud and **cotton** tunic.
Neg. (obj.)	"Frost is coming," he noted, touching the cotton **tunic** and wrinkled **shroud**.
`triple_bind_cycle`	
Query	plain chair, decorative lamp and metallic table
Positive	"It's eerie," says the mayor, beside a plain chair, metallic table and decorative lamp.
Neg. (attr.)	"It's eerie," says the mayor, beside a **decorative** chair, **plain** table and **metallic** lamp.
Neg. (obj.)	"It's eerie," says the mayor, beside a plain **lamp**, metallic **chair** and decorative **table**.
`triple_bind_swap`	
Query	shared canvas, inexpensive palette and simple easel
Positive	"What a mess!" cried he, seeing the shared canvas, simple easel and inexpensive palette.
Neg. (attr.)	"What a mess!" cried he, seeing the shared canvas, **inexpensive** easel and **simple** palette.
Neg. (obj.)	"What a mess!" cried he, seeing the shared canvas, simple **palette** and inexpensive **easel**.
`attr_bundle`	
Query	inviting delicate glass and ugly ridged plate
Positive	Sipping tea, I used an inviting delicate glass and ugly ridged plate.
Neg. (attr.)	Sipping tea, I used an **ugly** delicate glass and **inviting** ridged plate.
Neg. (obj.)	Sipping tea, I used an inviting delicate **plate** and ugly ridged **glass**.
`predicative`	
Query	torn page and dated book
Positive	Suddenly, the book was dated and the page was torn per report.
Neg. (attr.)	Suddenly, the book was **torn** and the page was **dated** per report.
Neg. (obj.)	Suddenly, the **page** was dated and the **book** was torn per report.
`deferred_bind_anaphoric`	
Query	hard umbrella and greenish awning
Positive	Under the sun, an awning stood by an umbrella. The first was greenish, and the second felt hard.
Neg. (attr.)	Under the sun, an awning stood by an umbrella. The first was **hard**, and the second felt **greenish**.
Neg. (obj.)	Under the sun, an **umbrella** stood by an **awning**. The first was greenish, and the second felt hard.
`deferred_bind_respective`	
Query	neat fence and new bench
Positive	The bench by the path and the fence near the pinkish petal were new and neat respectively.
Neg. (attr.)	The bench by the path and the fence near the pinkish petal were **neat** and **new** respectively.
Neg. (obj.)	The **fence** by the path and the **bench** near the pinkish petal were new and neat respectively.

- **Length.** Each example is assigned one of four target length bands (10–15, 16–22, 23–30, or 31–42 words), shifted upward by 2–4 words for the three-object constructions. The target length band is chosen so that it is long enough to accommodate the assigned binding-distance range; the **`deferred_bind`** constructions therefore use only the three longer bands. The target band is specified in the prompt and enforced during validation, allowing one word below and four words above the band.
- **Binding distance.** Each example is assigned a target range for the distance over which an attribute–object binding must be resolved, preventing the relevant words from always appearing close together. For constructions in which attributes directly precede their objects, this is the minimum number of words separating consecutive core phrases (windows from 0–2 up to 14–26 words for two objects, and up to 7–13 for three, where the minimum is taken over both gaps); for **`predicative`**, it is the smaller of the two distances between

an object and its attribute; and for **deferred_bind**, it is the span from the final object mention to the first attribute assignment. The target range is specified in the prompt and enforced during validation.

- **Distractors.** For **pair_swap**, **predicative**, and the two **deferred_bind** construction types, the prompt asks for up to three additional attribute–object pairs that are woven into the sentence but do not appear in the query. The positive text therefore describes more than the query asks about, so the model must locate the queried bindings among unrelated ones rather than treat every attribute–object pair in the text as relevant.
- **Vocabulary.** Examples that duplicate the complete set of attributes and objects from an earlier example of the same construction type are rejected. To prevent a small set of words from dominating the dataset, each attribute and object is subject to a running frequency cap over the examples accepted so far. The cap is the larger of ten occurrences and a fixed fraction of accepted examples, with the fraction scaled by the number of attributes per example: 1.5%, 2.25%, and 3% for constructions with two, three, and four attributes, respectively. This keeps the allowance per attribute slot constant while preventing frequent words from becoming overrepresented early in generation. Each prompt also includes an avoid list containing the 100 most recently used and 40 most frequent words, together with a hard ban list for words that have already reached the cap.
- **Attribute pool.** Frequency caps alone were insufficient to diversify the attributes, because the model continued to favor a small set of common adjectives even after they had reached the cap, causing many generated examples to be rejected. We therefore also provide attribute suggestions in the prompt. These are drawn from a curated pool of 772 common adjectives organized into 19 families that apply broadly across object classes, including color, texture, material, condition, size, age, and value. The prompt includes the least-used attributes from each family that have not yet reached the frequency cap, sampled per family so that no single family dominates the suggestions. The pool serves only as a source of suggestions: attributes outside the pool are also accepted if they pass validation.

B.4 Validation

Each generated example passes a structural validation step before being added to the dataset; failures are discarded, logged, and regenerated. In addition to the length, binding-distance, and vocabulary constraints described in Section B.3, validation checks that the example has the intended binding structure.

Attributes are required to be single words, and all objects and attributes must be distinct, each appearing exactly once in the query, positive, and both negatives, so that all queried bindings are specified unambiguously. Each negative must be a word-level permutation of the positive (treating *a* and *an* as equivalent) that breaks every queried binding, introduces all intended incorrect bindings, and remains distinct from both the positive and the other negative. Article-agreement errors created by a permutation (e.g., *"a envelope"*) are repaired automatically, and any remaining article error not already present in the positive causes the example to be rejected.

Finally, validation enforces the structural constraints of each construction: attributes directly precede their objects in **pair_swap**, **triple_bind**, and **attr_bundle**; are never adjacent to their objects in **predicative**; and in **deferred_bind** are never adjacent to their objects and appear only after both objects have been introduced.

B.5 Finetuning and Evaluation

We hold out both **deferred_bind** construction types entirely and randomly split the remaining $80{,}000$ examples into training, validation, and test sets in an $80/10/10$ ratio, yielding $64{,}000$, $8{,}000$, and $8{,}000$ examples, respectively. Each example yields two triplets, one per hard negative, for a total of $200{,}000$ triplets: $128{,}000$ for training, $16{,}000$ for validation, $16{,}000$ for testing, and $40{,}000$ held-out **deferred_bind** triplets.

Each model is finetuned using LoRA (Hu et al., 2022) adapters with rank $r{=}4$, $\alpha{=}4$, and zero dropout, applied to the query, key, value, and output projections in every decoder layer. The vision

tower remains frozen, resulting in 3.8M trainable parameters for the 8B model. All inputs use the model's default instruction.

Training runs for one epoch on a single B200 GPU using AdamW (Loshchilov & Hutter, 2019), with a constant learning rate of 5×10^{-5} and weight decay 0.01. The objective is a hinge triplet loss over cosine similarity with margin γ=0.05. After training, we apply LiNeS (Wang et al., 2025) post hoc to rescale the LoRA updates across decoder layers using the default setting of Wang et al. (2025).

We evaluate attribute-binding performance using triplet accuracy, defined as the fraction of triplets for which the positive is closer to the query than the negative; chance performance is 50%. Table 10 reports this metric across all evaluated models on the in-distribution test set and the two held-out **`deferred_bind`** constructions, which test generalization to binding structures never seen during training.

Table 10: **Triplet accuracy on held-out finetuning examples.** Accuracy (%) on the in-distribution test set and the two held-out deferred-binding construction types. Both of our finetuned models outperform all baselines across all three evaluations. Best results are shown in bold.

Dataset	CLIP B/32	LABCLIP B/32	TripletCLIP B/32	READ-CLIP B/32	PE CORE-G	LamRA-Ret	Qwen3-VL Emb-2B	Qwen3-VL Emb-8B	Ours-2B	Ours-8B
Test set	62.9	62.2	57.6	65.0	65.4	56.2	57.1	60.4	84.0	**93.6**
deferred_bind_anaphoric	50.4	50.1	50.9	51.8	51.1	51.2	52.7	53.8	59.4	**72.9**
deferred_bind_respective	51.0	50.9	52.1	51.5	51.1	51.1	51.4	50.8	53.2	**56.6**

Our finetuned models improve substantially on the held-out test set, with Ours-8B reaching 93.6% triplet accuracy compared with 60.4% for the pretrained Qwen3-VL-Embedding-8B. The gains also generalize to the two **`deferred_bind`** constructions that are never seen during training: Ours-8B improves from 53.8% to 72.9% on **`deferred_bind_anaphoric`** and from 50.8% to 56.6% on **`deferred_bind_respective`**. This indicates that the learned binding improvements are not limited to the specific sentence structures used for finetuning.

C Ablations

C.1 Training Data

Table 11 first tests whether the binding gains in Section 5 reflect a general compositional capability or merely exposure to SABER's vocabulary. We retrain on three size-matched subsets of the training pool (30,900 rows each): *Qwen3-VL-Emb-8B-NoOverlap*, containing only examples with no SABER attributes or objects; *Qwen3-VL-Emb-8B-Random*, a uniform random sample preserving the pool's natural lexical overlap with SABER; and *Qwen3-VL-Emb-8B-OverlapOnly*, containing only examples that mention at least one SABER attribute or object. Even without any vocabulary overlap, finetuning yields large gains across all four SABER pairings, improving text–text accuracy from 53.7 to 99.3, text–image from 34.9 to 65.7, image–text from 36.3 to 71.8, and image–image from 84.9 to 93.1. Increasing lexical overlap with SABER provides only modest additional gains. Relative to *Qwen3-VL-Emb-8B-NoOverlap*, *Qwen3-VL-Emb-8B-Random* improves mean SABER accuracy by 1.6 points, while *Qwen3-VL-Emb-8B-OverlapOnly* improves it by 2.0 points. MMEB-v3 performance remains nearly unchanged across the three training subsets, with the overall score ranging from 55.2 to 55.3 and closely matching our finetuned 8B model. Overall, the results indicate that the binding gains primarily reflect transferable binding supervision, with SABER-specific lexical overlap contributing only modestly.

The final data ablation changes the structure of the training data rather than its lexical overlap with SABER. Our finetuned 8B model is trained only on short-query, long-candidate examples, where a short conjunction of attribute–object pairs is matched against a contextualized positive sentence and negatives formed by swapping attributes or objects. For *Qwen3-VL-Emb-8B-MixedLengths*, we augment this data with two additional formats constructed from the same underlying examples: short queries with short candidates and long queries with short candidates. The resulting training set contains 192,000 examples and exposes the same bindings across all three query–candidate length configurations. *Qwen3-VL-Emb-8B-MixedLengths* improves over our finetuned 8B model across all

Table 11: **Training-data ablation.** SABER-MCQ and MMEB-v3 performance for lexical-overlap and query–candidate length ablations of our 8B model. SABER-MCQ reports 4-way accuracy, while MMEB-v3 reports Hit@1 for Image, Video, and Agent and NDCG@5 for VisDoc and Text. Best results are shown in bold.

	SABER-MCQ				MMEB-v3					
Model	**T→T**	**T→I**	**I→T**	**I→I**	**Image**	**Video**	**VisDoc**	**Text**	**Agent**	**All**
Qwen3-VL-Emb-8B-NoOverlap	99.3	65.7	71.8	93.1	72.6	58.8	75.5	43.1	**42.6**	**55.3**
Qwen3-VL-Emb-8B-Random	99.4	68.7	73.8	94.3	72.5	58.8	75.4	43.2	42.3	**55.3**
Qwen3-VL-Emb-8B-OverlapOnly	99.4	69.8	74.1	94.7	72.5	58.8	75.4	43.3	42.0	55.2
Qwen3-VL-Emb-8B-MixedLengths	**99.6**	**70.6**	**77.1**	**94.8**	72.3	58.8	**75.6**	42.9	**42.6**	55.2
Qwen3-VL-Emb-8B	53.7	34.9	36.3	84.9	**73.7**	**59.6**	74.3	**43.7**	39.1	54.7
Ours-8B	99.3	69.0	74.1	94.3	72.5	58.7	75.4	43.1	42.3	55.2

four SABER pairings, increasing mean accuracy by 1.4 points. MMEB-v3 performance, however, is essentially unchanged, with the overall score remaining at 55.2. Because the mixed-length training set includes a short-query, short-candidate format that more closely matches SABER's templated query–candidate structure, we do not interpret these gains as evidence of better general attribute binding. Instead, the higher SABER scores may partly reflect closer alignment between the training and evaluation formats.

C.2 Effect of LiNeS

We next examine the effect of LiNeS (Wang et al., 2025) on the trade-off between attribute binding and general multimodal embedding performance. We evaluate *Qwen3-VL-Emb-8B-NoLiNeS* after finetuning but before applying LiNeS, with all other settings unchanged. As shown in Table 12, it achieves higher SABER accuracy than our final model on all four modality pairings, improving mean accuracy by 9.4 points. However, MMEB-v3 performance drops substantially, from 55.2 to 47.8, well below the pretrained model's 54.7. Thus, LiNeS sacrifices some binding gains to substantially recover general multimodal embedding performance.

Table 12: **LiNeS ablation.** SABER-MCQ and MMEB-v3 performance for our 8B model before and after applying LiNeS. SABER-MCQ reports 4-way accuracy, while MMEB-v3 reports Hit@1 for Image, Video, and Agent and NDCG@5 for VisDoc and Text. Best results are shown in bold.

	SABER-MCQ				MMEB-v3					
Model	**T→T**	**T→I**	**I→T**	**I→I**	**Image**	**Video**	**VisDoc**	**Text**	**Agent**	**All**
Qwen3-VL-Emb-8B-NoLiNeS	**100.0**	**87.3**	**87.8**	**99.0**	60.8	51.2	68.8	34.6	39.0	47.8
Qwen3-VL-Emb-8B	53.7	34.9	36.3	84.9	**73.7**	**59.6**	74.3	**43.7**	39.1	54.7
Ours-8B	99.3	69.0	74.1	94.3	72.5	58.7	**75.4**	43.1	**42.3**	**55.2**

C.3 Attention-Based Binding Analysis and Direct Supervision

To test whether finetuning changes how attributes and objects interact, we directly analyze text-side attention. Because causal attention flows only from later to earlier tokens, we measure each object–attribute connection in the available direction, from the later word to the earlier one. For words spanning multiple sub-tokens, we sum the attention from each sub-token of the later word to all sub-tokens of the earlier word, then average over the later-word sub-tokens and attention heads to obtain one value per layer and word pair. For each object, we define a binding margin as its mean attention to its own attribute words minus its mean attention to the attributes of the other objects in the same caption, using the same later-to-earlier convention. We average these object-level margins to obtain one margin per caption and layer; zero indicates no preference for the correct attributes. We compute this margin across all 36 text layers for Qwen3-VL-Embedding-8B, our finetuned model, and our finetuned model without LiNeS, on three text sets: 1,000 short queries sampled from the held-out

test split in proportion to the training distribution over construction types, their corresponding longer positive captions, and 1,000 SABER queries spanning two to ten objects.

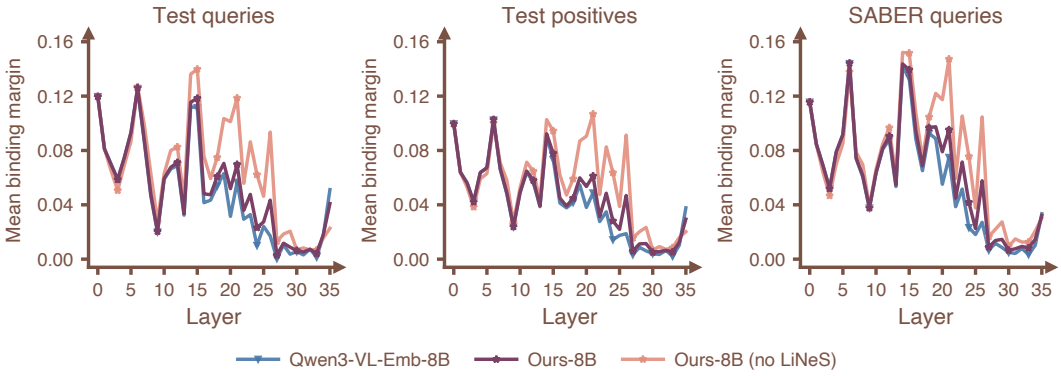

Figure 4: **Layer-wise binding margin.** Mean binding margin across the 36-layer text stack for test queries, their corresponding positive captions, and SABER queries. Finetuning preserves stronger binding in later layers, with a larger effect without LiNeS.

Figure 4 shows a consistent pattern across all three text sets. The base model exhibits positive binding margins in early layers, but this signal decays toward zero in the final third of the network. Finetuning leaves the first ~ 11 layers largely unchanged, while strong differences emerge in layers 18–26, where the base margin collapses. Over this band, Ours-8B retains 1.3–$1.4\times$ the base margin, and the variant without LiNeS retains 1.9–$2.4\times$, across queries, positives, and SABER. This suggests that finetuning helps preserve existing attribute–object structure deeper into the network, where it can influence the pooled embedding. We therefore turn this observation into an attention loss, described next, applied to this same layer range.

For a text input x, let $\mathcal{O}(x)$ denote its object words and $\mathcal{A}_x(o)$ the attribute words associated with object o in input x. Let $\overline{\mathcal{A}}_x(o)$ denote the attribute words associated with the other objects in x. We write $e_x^\ell(u, v)$ for the head-averaged attention between words u and v at layer ℓ, measured in the causally available direction from the later word to the earlier one and aggregated over their sub-tokens.

We define the binding margin of object o at layer ℓ as

$$m_x^\ell(o) = \frac{1}{|\mathcal{A}_x(o)|} \sum_{a \in \mathcal{A}_x(o)} e_x^\ell(o, a) - \frac{1}{|\overline{\mathcal{A}}_x(o)|} \sum_{a \in \overline{\mathcal{A}}_x(o)} e_x^\ell(o, a). \tag{1}$$

Let $\mathcal{B}$ denote a batch of text queries, $\mathcal{L} = \{18, \ldots, 26\}$ the layers on which we apply the attention loss, $[z]_+ = \max(0, z)$ the positive-part function, and $\tau = 0.1$ the target binding margin. We penalize object–layer pairs whose margin falls below τ:

$$\mathcal{L}_{\text{attn}} = \frac{1}{|\mathcal{L}| \sum_{x \in \mathcal{B}} |\mathcal{O}(x)|} \sum_{x \in \mathcal{B}} \sum_{o \in \mathcal{O}(x)} \sum_{\ell \in \mathcal{L}} \left[\tau - m_x^\ell(o)\right]_+. \tag{2}$$

The total training objective combines the triplet and attention losses, with $\lambda_{\text{attn}} = 0.1$ weighting the latter:

$$\mathcal{L} = \mathcal{L}_{\text{triplet}} + \lambda_{\text{attn}} \mathcal{L}_{\text{attn}}. \tag{3}$$

Training with the combined objective yields the results in Table 13. With LiNeS, adding the attention loss provides only a modest improvement, increasing mean SABER accuracy by 0.4 points over Ours-8B while slightly reducing MMEB-v3 from 55.2 to 55.0. Without LiNeS, the model reaches much higher SABER accuracy, but this behavior is already observed when removing LiNeS without attention supervision (Section C.2). In fact, adding the attention loss to the setting without LiNeS slightly decreases mean SABER accuracy from 93.5 to 93.2 and reduces MMEB-v3 from 47.8 to 47.2. Overall, direct attention supervision provides little additional benefit beyond our existing training objective.

Table 13: **Attention loss ablation.** SABER-MCQ and MMEB-v3 performance for our 8B model trained with the attention loss, with and without LiNeS. SABER-MCQ reports 4-way accuracy, while MMEB-v3 reports Hit@1 for Image, Video, and Agent and NDCG@5 for VisDoc and Text. Best results are shown in bold.

	SABER-MCQ				MMEB-v3					
Model	**T→T**	**T→I**	**I→T**	**I→I**	**Image**	**Video**	**VisDoc**	**Text**	**Agent**	**All**
Qwen3-VL-Emb-8B-Attn-LiNeS	99.4	70.0	74.8	94.2	72.4	58.6	75.2	42.8	42.2	55.0
Qwen3-VL-Emb-8B-Attn-NoLiNeS	**100.0**	**86.4**	**87.2**	**99.2**	59.5	50.5	66.3	34.8	39.2	47.2
Qwen3-VL-Emb-8B	53.7	34.9	36.3	84.9	**73.7**	**59.6**	74.3	**43.7**	39.1	54.7
Ours-8B	99.3	69.0	74.1	94.3	72.5	58.7	**75.4**	43.1	**42.3**	**55.2**

D Generalization Across Multimodal Embedders

To test whether our finetuning strategy is specific to Qwen3-VL-Embedding or generalizes to other multimodal embedders, we apply the same recipe to two additional model families, UEmbed (Song et al., 2026) and WeMM-Embedding (Zhou et al., 2026), each at 2B and 9B parameters. Both are built on Qwen3.5 (Qwen Team, 2026), whose hybrid decoder uses softmax attention only every fourth layer, with linear-attention blocks in the remaining layers.

We retain the training setup from our main experiments, using the same data, learning rate, and LiNeS. We apply rank-4 LoRA to the query, key, value, and output projections of the full-attention layers and evaluate the checkpoint after one epoch. Each model uses its official prompting and pooling scheme; for UEmbed, we use the dense embedding rather than its sparse representation. The unmodified public checkpoint serves as the baseline for comparison with its finetuned counterpart on SABER and MMEB-v3.

As shown in Table 14, binding-focused finetuning consistently improves SABER across both model families, model sizes, and all four modality pairings. The largest gains generally occur in text–text retrieval: UEmbed improves by 48.6 and 44.3 points at 2B and 9B, respectively, while WeMM-Embedding improves by 37.5 and 31.9 points. Importantly, these gains transfer beyond the text modality used for training. All four models also improve on both cross-modal settings and on image–image retrieval, with particularly large cross-modal gains for the 9B models. For example, WeMM-Embedding-9B improves from 30.6 to 58.7 in text–image retrieval and from 32.6 to 64.4 in image–text retrieval. This consistent transfer suggests that the effect is not specific to the Qwen3-VL-Embedding backbone and that text-only binding supervision can improve multimodal representations across different embedding architectures.

Table 14: **Generalization to additional embedding backbones.** Performance of UEmbed and WeMM-Embedding on SABER and MMEB-v3 before and after binding-focused finetuning. Finetuning improves all four SABER modality pairings across both model families and scales, while largely preserving MMEB-v3 performance. SABER-MCQ reports 4-way accuracy, while MMEB-v3 reports Hit@1 for Image, Video, and Agent and NDCG@5 for VisDoc and Text. Best results per pretrained–finetuned pair are shown in bold.

	SABER-MCQ				MMEB-v3					
Model	**T→T**	**T→I**	**I→T**	**I→I**	**Image**	**Video**	**VisDoc**	**Text**	**Agent**	**All**
UEmbed-2B	33.2	27.3	24.7	83.4	**63.7**	**50.8**	**71.5**	**35.4**	**35.9**	**48.2**
UEmbed-2B-Finetuned	**81.8**	**39.2**	**39.8**	**88.3**	62.2	48.9	67.8	31.9	34.2	45.6
UEmbed-9B	46.9	36.4	32.2	88.7	**68.6**	**55.6**	**75.2**	**40.9**	**38.1**	**52.3**
UEmbed-9B-Finetuned	**91.2**	**57.3**	**49.7**	**94.6**	67.6	55.1	74.5	40.4	35.9	51.3
WeMM-Embedding-2B	51.4	29.8	28.9	59.6	**78.0**	**65.2**	**78.4**	**44.2**	44.8	**58.4**
WeMM-Embedding-2B-Finetuned	**88.9**	**38.9**	**41.2**	**73.3**	77.1	64.3	77.6	43.9	**45.9**	58.2
WeMM-Embedding-9B	65.4	30.6	32.6	63.7	**80.2**	**68.5**	**81.1**	**46.9**	**50.1**	**61.7**
WeMM-Embedding-9B-Finetuned	**97.3**	**58.7**	**64.4**	**78.6**	79.7	67.9	80.1	46.7	49.4	61.1

These gains come with relatively modest changes in general-purpose retrieval performance. MMEB-v3 decreases by 1.0 point for UEmbed-9B and by only 0.2 and 0.6 points for WeMM-Embedding-2B and 9B, respectively, despite their large SABER improvements. UEmbed-2B shows a larger 2.6-point decrease, indicating a less favorable binding–generalization trade-off at this scale. Overall, however, the same lightweight finetuning recipe substantially improves attribute binding across additional backbones while largely retaining their broader retrieval capabilities.

E Comparing Text and Image Finetuning for Attribute Binding

Our main approach uses text-only supervision because text makes attribute–object bindings explicit and easy to manipulate, enabling precisely controlled hard negatives. However, this high degree of control may also make text supervision artificially clean. Images inevitably introduce variation in appearance, pose, lighting, size, spatial arrangement, and other visual factors. Although such variation makes the learning problem harder, it may also provide useful regularization, forcing the model to learn binding representations that are robust to nuisance variation rather than relying on the clean structure of text. We therefore ask whether text or image supervision is more effective for learning attribute binding. To study this, we split SABER into a $70/10/20$ train/validation/test split and finetune on increasing amounts of either text or image supervision, corresponding to 10%, 20%, 35%, and 70% of the full benchmark. Results are shown in Table 15.

Table 15: **Text vs. image finetuning on SABER.** Performance of Qwen3-VL-Embedding-8B after finetuning with increasing amounts of SABER data using either text–text or image–image supervision. Text supervision yields stronger overall SABER performance and consistently better cross-modal transfer, while image supervision performs best on the matched image–image setting. SABER–MCQ reports 4-way accuracy on the fixed 360-example test split, with all models evaluated at epoch 40. MMEB-v3 reports Hit@1 for Image, Video, and Agent tasks and NDCG@5 for VisDoc and Text tasks. Best results are shown in bold.

Finetuning Modalities	SABER Finetuning Data	SABER-MCQ 360 Test Examples (20%)				MMEB-v3					
		T→T	T→I	I→T	I→I	Image	Video	VisDoc	Text	Agent	All
T→T	180 (10%)	87.2	40.6	46.7	87.5	73.8	59.8	74.8	43.7	39.3	54.9
	360 (20%)	95.6	52.5	55.6	90.0	**73.9**	**59.9**	74.9	43.7	39.5	55.0
	630 (35%)	96.9	54.7	59.4	90.6	73.8	**59.9**	75.0	**43.8**	**39.7**	**55.1**
	1,260 (70%)	**97.8**	**55.6**	**60.3**	90.8	73.8	59.8	75.1	43.7	39.6	55.0
I→I	180 (10%)	56.1	32.8	39.4	95.3	**73.9**	59.6	74.6	43.7	39.2	54.8
	360 (20%)	61.1	36.7	44.4	96.9	**73.9**	59.7	75.0	43.5	39.5	54.9
	630 (35%)	66.1	38.6	51.9	97.8	**73.9**	59.5	75.0	43.4	39.6	54.9
	1,260 (70%)	67.2	42.5	52.8	**99.2**	73.8	59.6	**75.2**	43.5	39.5	54.9

Across both supervision modalities, SABER performance improves as more finetuning data is used. As expected, each modality performs best in its matched evaluation setting: text–text finetuning yields substantially higher text–text accuracy, whereas image–image finetuning is consistently stronger on image–image retrieval. More interestingly, within this controlled SABER setting, text supervision transfers better to the two cross-modal evaluations. At every training-set size, text–text finetuning outperforms image–image finetuning on both text–image and image–text retrieval. With the full training split, the gap is 13.1 points in text–image retrieval and 7.5 points in image–text retrieval.

MMEB-v3 performance remains largely stable across both supervision modalities and training-set sizes, with the overall score varying only from 54.8 to 55.1, possibly reflecting the relatively small amount of finetuning data used here. While the SABER results suggest that text supervision may provide a more effective signal for cross-modal binding transfer, this comparison is limited to a relatively small benchmark with a single family of synthetic images and templated texts, and should not be taken as evidence that text supervision is generally superior to image supervision.

F Evaluation and Analysis on Instance Retrieval

F.1 Datasets and Evaluation Protocols

We evaluate on six instance-retrieval benchmarks spanning landmarks, sculptures, household objects, and diverse object instances. Oxford5K (Philbin et al., 2007) and Paris6K (Philbin et al., 2008) evaluate landmark retrieval using cropped query regions. Images labeled *good* or *ok* are treated as relevant, while *junk* images are ignored during evaluation. Sculptures6K (Arandjelović & Zisserman, 2011) similarly uses cropped queries, with benchmark-specific relevance annotations. PODS (Sundaram et al., 2025) contains household objects photographed under several distribution shifts and evaluates retrieval across reference views of each object. ILIAS (Kordopatis-Zilos et al., 2025) contains diverse object instances and supports both image–image and text–image retrieval over a shared image database. INSTRE (Wang & Jiang, 2015) contains single-object images (INSTRE-S) and images depicting pairs of objects (INSTRE-M). Following the original protocol, we evaluate queries from INSTRE-S against INSTRE-S, and queries from INSTRE-M against either INSTRE-S or the combined INSTRE-S+M gallery. In the combined setting, relevance depends on whether a retrieved image contains one or both query instances, with images containing both receiving higher relevance.

We evaluate Qwen3-VL-Embedding-8B and our finetuned model under identical conditions on each benchmark, following the benchmark-specific protocols described above. Images are embedded in full unless a benchmark provides a ground-truth query crop, in which case we use the specified crop. We do not include auxiliary distractor sets. Accordingly, comparisons between the pretrained and finetuned models are directly controlled, while absolute performance may differ from published results that use additional distractors.

Quantitative results across all benchmarks and evaluation settings are reported in Table 6 of the main paper. Figure 5 complements these results with qualitative image–image retrieval examples on ILIAS. We show queries for which either the pretrained or finetuned model has a large relative advantage, together with their top-5 retrieved images. Across all $1{,}232$ ILIAS image queries, the finetuned model performs better on 923, the pretrained model on 108, and the two models tie on 201.

F.2 Why Might Binding-Focused Finetuning Improve Instance Retrieval?

To examine the instance-retrieval gains more closely, we use ILIAS's hierarchical taxonomy to evaluate retrieval within fine-grained object classes. We consider its 38 named fine-grained classes, spanning 491 object instances across seven top-level categories. We exclude leaves labeled *other*, while the Media category contains no named class at this level. Retrieval is evaluated independently within each class using the protocol described in the main paper.

Table 16 summarizes performance by top-level category, averaging mAP over the fine-grained classes within each category; *All* averages over all 38 classes. The finetuned model improves across every top-level category, with gains ranging from $+6.1$ to $+11.2$ points and the largest improvements on Toy ($+11.2$) and Fashion ($+10.0$). Figure 6 provides qualitative examples from this fine-grained evaluation, comparing the top-5 retrievals of the pretrained and finetuned models within individual object classes.

Table 16: **ILIAS fine-grained instance retrieval.** Image–image retrieval mAP (%) averaged over the fine-grained classes within each top-level category; *All* averages over all 38 classes. Best results are shown in bold.

Model	Art	Fashion	Household	Landmark	Product	Technology	Toy	All
# Classes	2	7	7	4	5	5	8	38
# Instances	33	92	50	136	36	49	95	491
Qwen3-VL-Emb-8B	75.4	62.4	72.4	76.1	66.0	76.6	52.5	66.6
Ours-8B	**82.4**	**72.4**	**80.2**	**82.2**	**72.5**	**84.4**	**63.7**	**75.2**



Figure 5: **Qualitative image–image retrieval on ILIAS.** Top-5 retrievals for two queries favoring Qwen3-VL-Embedding-8B (top) and two favoring Ours-8B (bottom). Green positives denote images of the same instance as the query, while red negatives denote other instances.



Figure 6: **Qualitative fine-grained instance retrieval on ILIAS.** Examples from three fine-grained classes under the category-restricted evaluation. For each query, we show the top-5 retrievals from Qwen3-VL-Embedding-8B and Ours-8B. Green positives depict the same instance as the query, while red negatives depict different instances from the same fine-grained class.

G Attribute Binding Beyond Attribute Swaps

SABER follows a common MCQ setup in which queries and candidates contain the same objects and attributes, with the correct candidate preserving the query's bindings and incorrect candidates reassigning attributes across objects (Yuksekgonul et al., 2023; Hsieh et al., 2023; Bordes et al., 2023). We next ask whether the resulting gains reflect broader improvements in attribute binding or specialization to this swap-based setting.

We adopt a complementary formulation from prior work (Lewis et al., 2024) in which candidates describe only part of the query. Given an image containing two attributed objects, the model retrieves a text naming a single attribute–object pair; the correct candidate preserves one of the image's bindings, whereas the hard negative candidates pair each object with the other object's attribute. Because

both concepts in each candidate appear in the image, concept presence alone cannot distinguish the correct binding, and learned concept associations can even favor the mismatch (Tang et al., 2023). We evaluate this formulation on the full CLEVR-based image–text benchmark introduced by Lewis et al. (2024), combining all of its splits into a single benchmark that we denote CLEVR-bind.

Since CLEVR-bind uses only two-object scenes, Figure 7 compares model performance on CLEVR-bind with SABER in the corresponding two-object image–text setting. Across the ten evaluated models, the scores are positively correlated (Spearman $\rho = 0.89$; Kendall $\tau = 0.78$). Our finetuned 8B model improves over its pretrained counterpart and achieves the highest accuracy in both settings. Together, these results suggest that its gains extend beyond attribute swaps to this partial-description formulation.

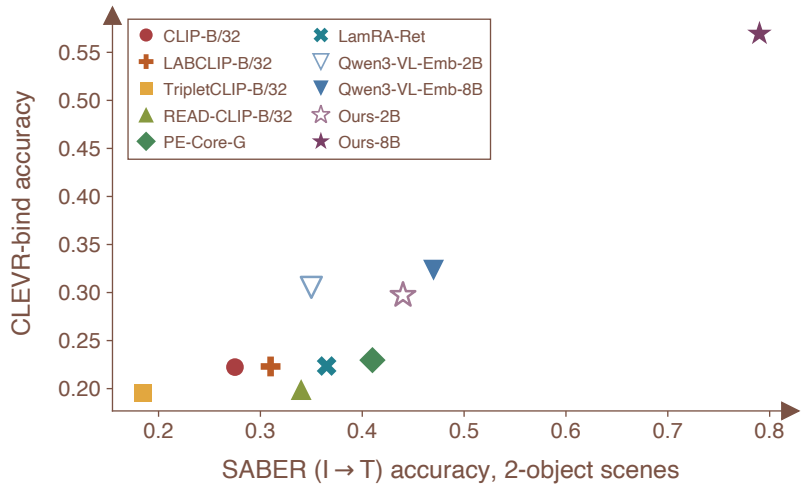

Figure 7: **Agreement across attribute-binding formulations.** Each point represents a model, comparing SABER accuracy on two-object image–text scenes with CLEVR-Bind accuracy. Performance is positively correlated across the ten models (Spearman $\rho = 0.89$; Kendall $\tau = 0.78$), with Ours-8B achieving the highest accuracy in both settings.

H Can Binding-Specific Instructions Replace Finetuning?

Qwen3-VL-Embedding conditions its embeddings on a task instruction, and its developers recommend tailoring these instructions to the target task to improve performance (Li et al., 2026). This raises the question of whether attribute binding can be improved through instruction choice alone, without modifying the model weights. We evaluate the default instruction (*"Represent the user's input."*) alongside 40 binding-specific alternatives on SABER and the MMEB-v3 image tasks, using both the pretrained and finetuned 8B models. Figure 8 shows the resulting trade-off between SABER and MMEB-v3 performance across instruction choices.

For the pretrained Qwen3-VL-Embedding-8B model, binding-specific instructions improve SABER accuracy at the expense of MMEB-v3 image performance, revealing a trade-off between the two. Although our model is finetuned only with the default instruction, it remains instruction-sensitive and exhibits a similar trade-off under the same alternatives. Crucially, with the default instruction alone, our finetuned model already surpasses the highest SABER accuracy achieved by the pretrained model under any of the 40 binding-specific instructions. Thus, tailored instructions can further improve binding performance after finetuning, but do not replace the gains from finetuning itself.

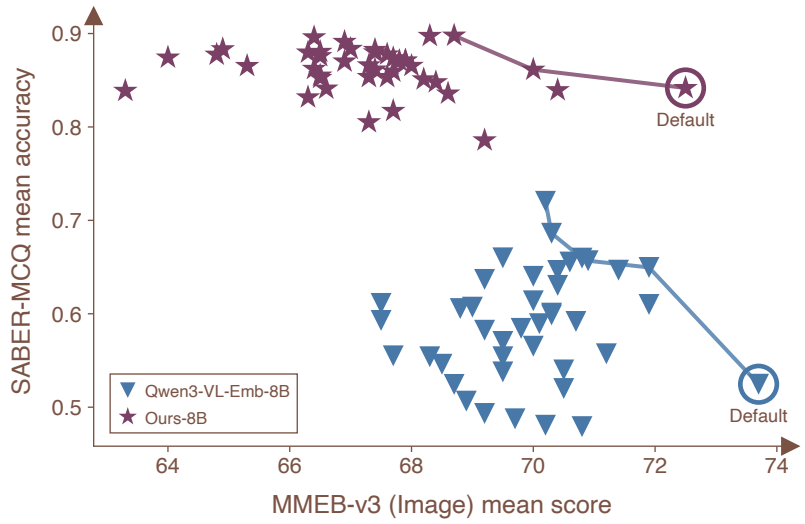

Figure 8: **Instruction trade-offs.** Mean SABER-MCQ accuracy versus mean MMEB-v3 score on image tasks across 41 instruction choices for Qwen3-VL-Embedding-8B and our finetuned model (Ours-8B). Lines trace each model's Pareto frontier over instructions, and circles indicate the default instruction.

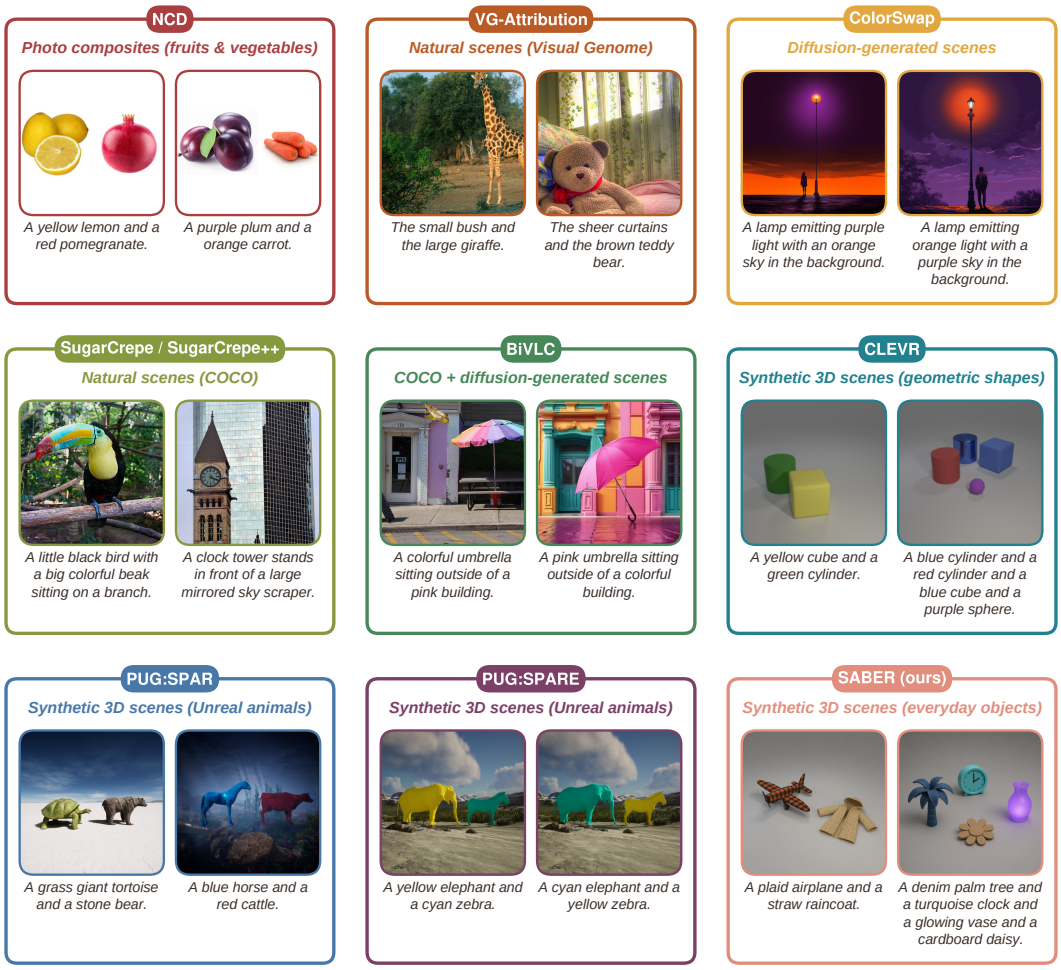

Figure 9: **Examples from attribute-binding benchmarks.** Existing benchmarks span natural images, photo composites, diffusion-generated scenes, and synthetic 3D environments, with varying vocabulary sizes and numbers of bindings per sample.

Table 17: **SABER-MCQ accuracy across scene complexity.** Accuracy (%) for 4-way MCQs across scenes with 2–10 objects, reported for all four modality pairings. *Mean* averages over object counts. Best and second-best results within each modality pairing and column are shown in bold and underlined, respectively.

Modalities	Model	Number of objects									Mean
		2	3	4	5	6	7	8	9	10	
T→T	CLIP ViT-B/32	55.5	57.5	61.5	60.0	55.5	56.0	54.0	41.5	40.5	53.6
	LABCLIP ViT-B/32	58.0	57.5	64.0	62.5	55.0	57.0	55.0	46.5	41.5	55.2
	TripletCLIP ViT-B/32	12.0	36.5	43.5	48.0	50.5	49.0	54.0	43.0	49.5	42.9
	READ-CLIP ViT-B/32	41.0	49.5	52.5	55.5	50.0	50.5	51.0	39.5	36.5	47.3
	PE-Core-bigG-14-448	60.0	64.5	74.0	79.5	70.0	71.0	65.0	57.0	54.0	66.1
	LamRA-Ret	43.5	53.0	53.0	54.0	53.0	53.0	52.0	41.5	38.5	49.1
	Qwen3-VL-Embedding-2B	5.0	30.0	40.5	44.5	40.0	39.5	47.5	36.0	32.0	35.0
	Qwen3-VL-Embedding-8B	51.5	56.5	58.5	62.5	58.0	53.5	58.0	41.5	43.0	53.7
	Ours-2B	63.0	74.5	81.0	86.5	84.5	80.0	80.5	72.5	70.0	76.9
	Ours-8B	**99.0**	**98.0**	**98.5**	**99.0**	**100.0**	**99.5**	**100.0**	**100.0**	**100.0**	**99.3**
T→I	CLIP ViT-B/32	36.5	14.0	24.5	25.0	17.5	23.0	24.5	18.0	24.0	23.0
	LABCLIP ViT-B/32	33.0	17.0	22.0	25.5	19.0	20.0	25.0	21.0	22.5	22.8
	TripletCLIP ViT-B/32	36.0	30.5	27.5	28.5	24.0	23.0	27.5	22.5	24.0	27.1
	READ-CLIP ViT-B/32	34.0	22.0	21.0	21.5	22.5	21.5	27.5	22.0	23.0	23.9
	PE-Core-bigG-14-448	46.0	29.5	27.5	27.5	25.5	26.5	29.0	26.5	29.5	29.7
	LamRA-Ret	45.5	33.0	27.0	27.5	29.0	30.5	25.0	28.5	29.5	30.6
	Qwen3-VL-Embedding-2B	38.0	28.5	28.5	22.0	25.5	29.5	27.0	27.5	24.0	27.8
	Qwen3-VL-Embedding-8B	42.5	38.0	33.0	32.0	35.5	34.0	33.0	34.5	32.0	34.9
	Ours-2B	42.0	34.5	36.0	34.5	26.5	37.0	37.5	36.5	28.5	34.8
	Ours-8B	**74.0**	**66.0**	**65.0**	**68.0**	**66.5**	**71.5**	**74.0**	**69.0**	**67.0**	**69.0**
I→T	CLIP ViT-B/32	27.5	21.5	24.0	17.5	23.5	18.5	23.5	19.0	26.0	22.3
	LABCLIP ViT-B/32	31.0	27.0	27.5	19.5	24.5	23.0	25.0	17.5	22.5	24.2
	TripletCLIP ViT-B/32	18.5	24.5	20.0	27.5	25.0	17.5	18.0	23.0	25.5	22.2
	READ-CLIP ViT-B/32	34.0	20.5	22.5	16.5	25.5	18.5	26.5	23.5	25.5	23.7
	PE-Core-bigG-14-448	41.0	26.0	30.0	21.0	24.5	23.0	30.5	24.5	27.5	27.6
	LamRA-Ret	36.5	25.0	27.5	26.0	24.5	28.0	26.5	26.0	23.0	27.0
	Qwen3-VL-Embedding-2B	35.0	31.0	25.5	27.0	23.0	22.0	27.5	25.0	22.0	26.4
	Qwen3-VL-Embedding-8B	47.0	36.5	37.5	33.0	30.5	33.5	42.5	34.0	32.5	36.3
	Ours-2B	44.0	38.5	30.5	33.0	26.0	26.5	35.5	34.0	28.0	32.9
	Ours-8B	**79.0**	**69.5**	**79.0**	**75.0**	**69.0**	**69.5**	**78.0**	**75.0**	**72.5**	**74.1**
I→I	CLIP ViT-B/32	73.0	70.0	72.5	81.5	71.5	78.0	79.0	79.0	75.0	75.5
	LABCLIP ViT-B/32	73.0	69.5	73.0	81.0	71.5	78.0	78.0	79.0	74.5	75.3
	TripletCLIP ViT-B/32	73.5	61.5	60.0	64.0	54.5	52.5	58.0	58.5	59.5	60.2
	READ-CLIP ViT-B/32	83.5	75.5	79.5	82.0	81.5	81.5	87.5	83.0	86.5	82.3
	PE-Core-bigG-14-448	71.0	67.0	63.0	67.0	69.5	68.0	67.0	69.5	66.0	67.6
	LamRA-Ret	74.5	60.5	61.5	65.0	59.5	71.0	67.0	63.5	62.0	64.9
	Qwen3-VL-Embedding-2B	82.5	67.0	71.0	77.5	70.0	76.0	77.5	71.5	72.5	73.9
	Qwen3-VL-Embedding-8B	87.0	81.5	79.0	82.0	83.0	85.5	89.0	87.0	90.0	84.9
	Ours-2B	86.0	77.0	77.0	84.5	74.0	85.0	79.0	81.0	84.5	80.9
	Ours-8B	**94.5**	**94.5**	**87.5**	**91.5**	**91.0**	**97.0**	**97.0**	**98.5**	**97.0**	**94.3**

Table 18: **SABER large-gallery retrieval on independent scenes.** mAP (%) across scenes with 2–10 objects for the 100 independently generated queries at each scene size, reported for all four modality pairings. *Mean* averages over object counts. Best and second-best results within each modality pairing and column are shown in bold and underlined, respectively.

		Number of objects									
Modalities	**Model**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Mean**
T→T	CLIP ViT-B/32	75.7	77.4	80.4	76.2	74.1	74.7	73.7	70.2	63.3	74.0
	LABCLIP ViT-B/32	77.0	76.1	81.3	77.0	73.0	72.8	73.3	74.2	63.7	74.3
	TripletCLIP ViT-B/32	47.7	59.2	65.5	67.3	71.8	67.9	76.5	67.8	69.7	65.9
	READ-CLIP ViT-B/32	66.0	67.1	71.6	75.9	68.8	71.1	73.4	64.3	57.7	68.4
	PE-Core-bigG-14-448	78.2	79.2	85.8	89.7	82.3	82.2	84.5	76.0	72.4	81.1
	LamRA-Ret	70.8	71.9	73.8	74.7	72.9	71.6	73.3	65.0	59.2	70.4
	Qwen3-VL-Embedding-2B	41.6	52.8	61.3	66.9	65.2	63.3	72.0	59.5	58.1	60.1
	Qwen3-VL-Embedding-8B	72.0	74.3	75.8	78.6	76.5	74.4	77.7	65.7	66.4	73.5
	Ours-2B	81.3	84.6	89.2	94.8	92.4	88.3	90.4	85.2	83.3	87.8
	Ours-8B	**99.0**	**98.5**	**99.5**	**99.5**	**100.0**	**99.5**	**100.0**	**100.0**	**100.0**	**99.6**
T→I	CLIP ViT-B/32	62.5	43.9	44.5	44.6	41.1	35.0	33.7	24.9	19.1	38.8
	LABCLIP ViT-B/32	55.6	39.1	37.6	44.0	35.7	31.9	28.1	23.8	19.2	35.0
	TripletCLIP ViT-B/32	26.8	14.2	11.6	8.3	5.0	3.3	3.9	2.1	3.7	8.8
	READ-CLIP ViT-B/32	58.7	45.3	41.5	41.3	33.3	29.1	21.9	16.5	13.7	33.5
	PE-Core-bigG-14-448	69.7	55.2	56.3	52.6	53.2	54.0	54.9	52.7	55.3	56.0
	LamRA-Ret	70.0	53.8	56.2	58.1	54.2	54.9	51.4	54.4	57.8	56.8
	Qwen3-VL-Embedding-2B	62.7	54.6	53.2	52.2	54.3	56.2	51.8	54.9	46.0	54.0
	Qwen3-VL-Embedding-8B	64.0	59.6	56.3	60.9	58.3	63.5	57.2	57.5	59.2	59.6
	Ours-2B	67.4	58.9	58.2	60.9	54.0	61.4	60.1	60.1	52.8	59.3
	Ours-8B	**83.0**	**81.3**	**76.8**	**82.1**	**81.9**	**83.9**	**89.0**	**81.6**	**80.7**	**82.3**
I→T	CLIP ViT-B/32	47.5	39.4	42.7	32.5	28.7	36.6	29.6	22.1	19.1	33.1
	LABCLIP ViT-B/32	46.1	39.8	35.1	29.0	21.5	22.8	17.1	9.5	10.9	25.8
	TripletCLIP ViT-B/32	14.0	8.4	5.6	4.0	3.8	2.2	2.0	2.1	3.1	5.0
	READ-CLIP ViT-B/32	45.3	31.0	22.1	15.3	19.5	23.2	20.2	17.3	18.3	23.6
	PE-Core-bigG-14-448	65.2	55.3	59.3	49.9	54.4	53.8	54.5	52.5	54.7	55.5
	LamRA-Ret	66.0	51.7	57.2	51.6	45.4	53.7	44.9	51.1	41.7	51.5
	Qwen3-VL-Embedding-2B	62.7	59.1	57.6	48.3	53.9	53.4	55.7	54.1	49.9	55.0
	Qwen3-VL-Embedding-8B	71.9	66.6	66.7	59.7	58.2	59.0	61.4	59.2	61.1	62.6
	Ours-2B	65.8	61.7	62.2	54.2	52.6	54.4	62.3	61.0	50.7	58.3
	Ours-8B	**84.7**	**82.8**	**87.9**	**85.3**	**79.9**	**84.2**	**89.1**	**88.8**	**87.0**	**85.5**
I→I	CLIP ViT-B/32	84.4	82.2	84.2	87.9	83.8	85.8	85.7	86.9	85.2	85.1
	LABCLIP ViT-B/32	84.4	82.1	84.2	87.9	83.8	85.8	85.2	87.1	85.2	85.1
	TripletCLIP ViT-B/32	79.9	74.5	71.7	68.6	56.1	54.4	58.5	55.3	49.7	63.2
	READ-CLIP ViT-B/32	90.7	89.5	89.4	88.5	88.9	87.7	91.8	89.8	90.9	89.7
	PE-Core-bigG-14-448	84.3	81.5	83.1	78.2	81.6	82.1	79.1	82.6	81.2	81.5
	LamRA-Ret	83.7	76.5	79.8	79.1	73.6	84.2	79.2	78.7	78.3	79.2
	Qwen3-VL-Embedding-2B	92.9	79.8	83.3	87.8	85.1	84.1	86.5	84.2	85.7	85.5
	Qwen3-VL-Embedding-8B	93.8	90.7	91.1	90.7	88.1	92.4	92.0	92.7	95.9	91.9
	Ours-2B	94.9	87.2	87.5	92.4	88.5	89.3	86.2	90.1	92.4	89.8
	Ours-8B	**96.3**	**97.2**	**94.9**	**94.2**	**94.4**	**98.5**	**98.5**	**100.0**	**99.5**	**97.1**

Table 19: **SABER large-gallery retrieval on nested chains.** mAP (%) across scenes with 2–10 objects for the 100 nested-chain queries at each scene size, reported for all four modality pairings. *Mean* averages over object counts. Best and second-best results within each modality pairing and column are shown in bold and underlined, respectively.

		Number of objects									
Modalities	**Model**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Mean**
T→T	CLIP ViT-B/32	25.0	28.6	30.7	33.1	38.8	39.6	42.2	46.6	57.4	38.0
	LABCLIP ViT-B/32	27.2	30.0	32.1	34.1	37.6	39.0	42.4	43.7	57.0	38.1
	TripletCLIP ViT-B/32	18.8	25.5	28.3	31.1	33.3	36.2	41.8	45.3	62.1	35.8
	READ-CLIP ViT-B/32	26.4	30.0	31.2	33.1	36.9	38.7	41.0	44.1	57.8	37.7
	PE-Core-bigG-14-448	30.1	32.9	34.5	36.3	38.6	43.0	44.6	50.2	63.7	41.5
	LamRA-Ret	25.0	30.3	31.5	34.5	37.0	39.3	39.4	43.0	56.0	37.3
	Qwen3-VL-Embedding-2B	24.7	29.0	31.0	32.5	33.6	34.2	37.0	38.6	45.5	34.0
	Qwen3-VL-Embedding-8B	29.1	31.8	33.5	36.5	39.4	39.4	41.6	39.8	49.3	37.8
	Ours-2B	31.8	35.4	38.7	42.2	45.0	47.4	50.5	58.3	66.4	46.2
	Ours-8B	**44.5**	**48.7**	**53.5**	**59.6**	**65.7**	**72.4**	**77.3**	**81.0**	**86.4**	**65.5**
T→I	CLIP ViT-B/32	13.6	15.0	15.3	14.5	13.8	13.6	12.5	11.4	13.2	13.7
	LABCLIP ViT-B/32	12.1	14.1	14.0	13.6	12.9	11.5	13.4	11.7	13.9	13.0
	TripletCLIP ViT-B/32	5.4	6.2	6.0	5.7	5.9	5.6	5.4	3.5	5.1	5.4
	READ-CLIP ViT-B/32	11.5	12.8	12.2	12.6	11.6	9.5	10.7	7.7	7.6	10.7
	PE-Core-bigG-14-448	26.7	26.1	24.4	22.8	21.8	21.1	22.9	21.7	24.9	23.6
	LamRA-Ret	19.5	24.0	23.3	24.3	26.0	26.7	28.6	34.5	39.3	27.3
	Qwen3-VL-Embedding-2B	23.3	26.4	27.1	26.6	25.4	26.9	30.6	31.2	35.4	28.1
	Qwen3-VL-Embedding-8B	28.3	30.6	30.1	30.8	32.8	30.9	37.0	37.4	40.2	33.1
	Ours-2B	26.3	28.4	28.9	28.3	25.9	26.6	30.0	34.8	35.5	29.4
	Ours-8B	**35.1**	**36.3**	**37.7**	**39.6**	**43.6**	**46.0**	**54.3**	**61.7**	**62.9**	**46.3**
I→T	CLIP ViT-B/32	15.7	16.3	17.0	17.2	17.5	16.3	16.3	14.5	14.3	16.1
	LABCLIP ViT-B/32	13.0	12.8	13.9	13.3	11.9	11.4	12.5	11.2	9.1	12.1
	TripletCLIP ViT-B/32	4.4	4.8	3.6	3.7	3.5	3.3	3.9	2.2	2.3	3.5
	READ-CLIP ViT-B/32	10.3	12.0	12.7	13.8	14.8	13.2	13.6	14.4	13.5	13.1
	PE-Core-bigG-14-448	29.7	28.9	29.5	29.3	30.9	32.1	37.9	36.7	40.9	32.9
	LamRA-Ret	22.1	25.0	25.5	26.7	28.0	28.7	31.2	30.1	34.6	28.0
	Qwen3-VL-Embedding-2B	24.6	27.2	27.6	30.6	30.2	29.9	34.2	35.4	38.1	30.9
	Qwen3-VL-Embedding-8B	26.2	27.5	28.2	28.3	28.9	27.9	34.3	31.2	36.6	29.9
	Ours-2B	27.5	28.6	28.6	30.0	30.1	31.3	33.1	40.7	42.8	32.5
	Ours-8B	**34.1**	**34.9**	**37.6**	**39.3**	**43.2**	**44.8**	**52.3**	**59.1**	**68.2**	**46.0**
I→I	CLIP ViT-B/32	21.6	29.3	37.8	47.7	51.6	58.4	65.7	66.4	65.0	49.3
	LABCLIP ViT-B/32	21.6	29.2	37.8	47.7	51.7	58.4	65.5	66.4	64.5	49.2
	TripletCLIP ViT-B/32	17.1	20.6	24.8	31.4	30.8	36.5	39.2	46.3	42.0	32.1
	READ-CLIP ViT-B/32	24.3	31.6	41.1	**50.8**	**56.4**	62.2	69.9	73.6	76.6	54.1
	PE-Core-bigG-14-448	30.0	33.2	34.1	39.5	43.8	49.3	58.1	65.3	70.5	47.1
	LamRA-Ret	26.0	30.6	32.9	36.4	39.5	44.3	51.6	60.6	71.2	43.7
	Qwen3-VL-Embedding-2B	28.5	32.7	35.6	37.8	42.6	48.8	56.7	66.8	76.2	47.3
	Qwen3-VL-Embedding-8B	32.4	35.5	36.5	40.2	46.4	52.0	62.9	76.3	88.7	52.3
	Ours-2B	31.4	34.2	36.7	39.3	42.1	51.2	59.6	70.5	83.9	49.9
	Ours-8B	**37.2**	**39.6**	**41.4**	45.9	54.0	**63.1**	**72.1**	**83.5**	**90.4**	**58.6**

Table 20: **Licenses and usage terms of datasets and models used in this work.**

Dataset	License	Model	License
NCD	MIT	CLIP ViT-B/32	MIT
VG-Attribution	MIT	LABCLIP ViT-B/32	MIT
ColorSwap	MIT	TripletCLIP ViT-B/32	Open
SugarCrepe-Swap-Att.	MIT	READ-CLIP ViT-B/32	MIT
SugarCrepe++-Swap-Att.	MIT	PE-Core bigG-14-448	Apache 2.0
BiVLC-Swap-Att.	MIT	LamRA-Ret	MIT
CLEVR (Koishigarina et al., 2026)	CC BY 4.0	Qwen3-VL-Embedding	Apache 2.0
PUG:SPAR	CC BY-NC 4.0	Qwen3-VL-Reranker	Apache 2.0
PUG:SPARE	CC BY-NC 4.0	Gemma 4 31B-IT	Apache 2.0
MMEB-v3	MIT	GPT-Image-2	Proprietary
CLEVR (Lewis et al., 2024)	BSD-3-Clause	WeMM-Embedding	Apache-2.0
NegBench	MIT	UEmbed	CC BY 4.0
ILIAS	MIT		
INSTRE	Flickr T&C		
PODS	MIT		
Paris6K	Flickr T&C + NC		
Oxford5K	Flickr T&C + NC		
Sculptures6K	Flickr T&C + NC		