Hidden in Top-K: Probing Vision-Language Model Understanding Through Post-Retrieval Analysis

Anonymous ACL submission

Abstract

Vision-language models (VLMs) are often evaluated for their linguistic understanding using handcrafted datasets with contrastive image-caption pairs. However, these pairs represent only a narrow range of potential semantic violations: they do not account for the most challenging or all possible minimal alternations related to each concept. We propose a post-retrieval probing framework that examines a model's top-K retrieval to reveal its nuanced weaknesses in semantic comprehension. We evaluate five VLMs—CLIP, BLIP-2, FLAVA, SigLIP2, and fine-tuned Qwen2.5-VL on the SVO-Probes, targeting subject-verb-object role comprehension. Our innovative approach combines traditional metrics with semantic-similarity checks, human evaluations, and o3 reasoning model assessments, providing a clearer picture of model performance. Our results indicate that while VLMs achieve high pairwise accuracy and image-text matching accuracy, they struggle in top-K retrieval, particularly with verb and object understanding (human evaluation success rate at $1 \approx 70\%$). o3 assessments closely align with human judgments, enabling scalable evaluation. Overall, our post-retrieval probing offers a more robust and interpretable view of model performance.

1 Introduction

Recent studies have employed various probing methods, such as image-text matching (ITM), visual question answering, and guided masking, to assess the understanding of linguistic constructs by vision-language models (VLMs). These probing studies have explored specific capabilities, including spatial relations (Liu et al., 2023), verb understanding (Hendricks and Nematzadeh, 2021; Benova et al., 2025), counting (Parcalabescu et al., 2020, 2021), or word order (Thrush et al., 2022; Ma et al., 2023a), leading to intriguing conclusions. For instance, claims have emerged that modern VLMs, such as BLIP-2, have effectively mastered verb understanding based on pairwise accuracy results (Bugliarello et al., 2023). However, these evaluations typically rely on manually constructed positive and negative image-caption pairs designed to study specific linguistic features (see Figure 1).

Figure 1: An example from the SVO-Probes dataset (Hendricks and Nematzadeh, 2021) illustrates the limitations of the dataset annotations; an image labeled with the positive caption: *"A girl runs on a bridge."*, and with (subject ad object) negative captions: *"The man will run on the bridge.", "A girl runs up a hill."* . There are other captions in the dataset that, while semantically accurate, would be deemed incorrect by standard retrieval metrics, resulting in underestimation of the model's comprehension abilities, e.g.: *" A woman running on the street."* or *"A person runs on the road."*

This reliance on hand-crafted contrastive pairs presents two key limitations: (1) they fail to capture the full semantic space of incorrect interpretations – it is possible to violate a concept in many ways that these datasets do not test for, and (2) they do not necessarily surface the most challenging ways to test for each concept in a model. As a result, it is perfectly possible for a model to both (1) perform well on the specific contrastive pairs formed by human annotators and (2) fail on a large number of other pairs (crucially, even on pairs composed of images and captions from the same dataset). To create a comprehensive dataset of contrastive pairs, annotators would either need to assess all possi-

ble combinations or possess some prior knowledge regarding which pairs are likely to be found challenging.

To address this, we propose post-retrieval analysis. To our knowledge, this is the first study that, instead of evaluating a model's understanding ability with an ITM over a dataset of fixed contrastive pairs, examines the top-K retrieved samples by the model and uses the retrieved samples as a new pool for ITM. This setup highlights near misses—false positives that the studied model identified as highly similar to the true positives—and offers a richer, more diagnostic view of a model's linguistic understanding and robustness. The novelty of this approach lies in reusing the retrieved examples as a dataset of contrastive pairs that are more challenging for ITM for the studied model.

However, using a retrieval-based evaluation presents its own principal challenge as observed, e.g., by Hu et al. (2019); Chun et al. (2022): incomplete annotations. Since most datasets label only a few "correct" captions per image (or vice versa), valid but unlabeled matches are mistakenly treated as errors under standard metrics. As an example, for Figure 1, correct retrieval is considered *"A girl runs on a bridge."* . If the model retrieves *"A woman running on the street."*, which is semantically correct, it would be evaluated as an error. To address this, we complement retrieval with three alternative evaluation strategies: semantic similarity success (an alternative to success@K), human annotation, and o3 reasoning model annotation—to provide a more accurate assessment.

Our main contributions are as follows:

1. We propose a retrieval-based probing framework that uncovers nuanced errors in VLMs' behavior, beyond relying on human priors.

2. We introduce three evaluation strategies—semantic similarity, human annotation, and model-based scoring—to address annotation incompleteness.

3. We apply this framework to evaluate five VLMs— CLIP (Ilharco et al., 2021), BLIP-2 (Li et al., 2023), FLAVA (Singh et al., 2022) and SigLIP2 (Tschannen et al., 2025) and fine tunned Qwen2.5-VL (Bai et al., 2025)—on SVO-Probes dataset (Hendricks and Nematzadeh, 2021), demonstrating that retrieval-based probing provides a more diagnostic view of model performance.

Our results from post-retrieval analysis indicate that there is still substantial room for improvement. While the selected models perform well on pairwise accuracy on SVO-Probes, achieving an accuracy of over 80%, they exhibit poor performance when evaluated using retrieval with standard success@K. The results improve with semantic-similarity success evaluation; however, the task remains challenging, even when evaluated by a human on a subset of the datasets, with $\mathrm{S}_h@1$ ranging $60 - 80\%$. Additionally, our analysis of incorrectly retrieved examples suggests that understanding the verb is by far the most challenging part, even for models like SigLIP2 that account for compositionality.

For the model-based evaluation, we conducted experiments with different SOTA large vision-language models, but using the o3 reasoning model (OpenAI, 2025) as a judge led to highest agreement between the model and human annotations. With an F1 score of 0.8368, we consider o3 to be benefitial for automatic evaluation on SVO-Probes.

Finally, using the top 10 retrieved samples for a given query as contrastive samples leads to a significant decrease in ITM performance, supporting our hypothesis about the limitations of hand-designed contrastive image-caption probing datasets.

2 Related Work

Vision-Language Models (VLMs) align visual and textual modalities and are widely used for tasks such as image-text retrieval, captioning, and visual question answering. Prominent architectures include dual encoders like CLIP (Radford et al., 2021), FLAVA (Singh et al., 2022), and SigLIP2 (Tschannen et al., 2025), as well as multi-stage models such as BLIP-2 (Li et al., 2023), LlaVA-NeXT (Liu et al., 2024), or GPT-4 (OpenAI, 2024), which integrate frozen vision encoders with LLMs. While these models achieve strong performance on various downstream tasks, there are ongoing concerns about their true linguistic competence.

Fine-grained benchmarks Probing methods typically rely on handcrafted datasets composed of contrastive image-caption pairs designed to isolate specific linguistic features by foiling (Shekhar et al., 2017). For instance, SVO-Probes tests verb-role comprehension through subject, verb, and object negatives (Hendricks and Nematzadeh, 2021), while VALSE-counting focuses on numerical reasoning (Parcalabescu et al., 2021).

Beyond foiling, benchmarks like WinoGround (Thrush et al., 2022) and CREPE (Ma et al., 2023b) explore compositional grounding, or ViLMA (Kesen et al., 2024) and CV-Probes (Beňová et al., 2024) target verb phrase understanding through contrastive captions. Despite their value, these datasets often capture only narrow, synthetic variations and fail to represent the full diversity—or ambiguity—of real-world inputs.

Unlike prior work, our approach leverages post-retrieval analysis to assess model comprehension. It was previously observed by (Hu et al., 2019; Chun et al., 2022) that the retrieval evaluation has principal challenges, which we address by proposing alternative evaluation techniques. We uncover systematic errors by analyzing the top-K retrieved samples. Additionally, large language models (LLMs) like GPT-4o have been explored as scalable evaluators for retrieval (Alaofi et al., 2024; Wang et al., 2024; Vykopal et al., 2025).

The Retrieval Task and Its Evaluation Image and text retrieval are fundamental tasks in multimodal learning, where, given a query (an image or text), the goal is to retrieve the most relevant corresponding items (text or image) from a database. The output of a retrieval method is a list of samples ranked by relevance, from which the top K samples are selected for further use.

Evaluation of retrieval is commonly performed using success@K (S@K), which represents the proportion of cases in which the top-K ranked samples include at least one sample relevant to the query. Other evaluation metrics are also commonly used, e.g., recall@K or precision@K.

3 Probing through Post-Retrieval Analysis

Existing probing approaches, such as ITM or visual question answering, rely on curated datasets designed to isolate specific linguistic constructs—e.g., verbs, spatial relations, or counting. These typically use contrastive examples: a positive image-caption pair and a minimal negative pair that alters or violates one key component (example in Figure 1). While useful, such datasets cover only a narrow slice of possible errors, and their construction reflects human priors about what is challenging.

The clear limitation of this approach is that while such contrastive pairs are specific (i.e., they only differ in the probed concept, so one will not see "A flamingo flying over a lake." as a negative for Figure 1), they do not exhaustively delineate the extent of each concept. A model may succeed on curated contrastive pairs but still fail on many semantically valid variants not captured by the dataset. So, negative examples may not represent the most difficult edge cases.

We propose an alternative: *probing through post-retrieval analysis*. We analyze the top-K retrieved items for each query. These include both correct matches and close distractors. This allows us to inspect not just whether a model "gets it right," but also the kinds of near-misses it tends to make. The procedure for text retrieval is described in (Algorithm 1). The same algorithm holds for image retrieval.

Algorithm 1: Probing through post-retrieval analysis

```
Data: dataset: an image-caption dataset
foreach image in dataset do
    Rank captions by relevance to the image
    Collect the top K samples
    Compute evaluation metrics for the top K;      // in sec. 3.2.1, 3.2.2, 3.2.3
    Perform exploratory analysis on top K samples;
```

Post-retrieval probing provides a window into the model's representational confusion. Are mismatches mainly due to subject swaps or verb misuse? One can conduct an exploratory analysis on the top K samples and perform a more detailed assessment of the model's capabilities, such as its ability to correctly recognize the subject and object of a relation, recognize activities, and count objects.

3.1 Fine-tuning of Causal LLM for Sentence and Image Encoding

To be able to use causal LLM such as Qwen2.5-VL-7B-Instruct model (Bai et al., 2025) for encoding and later retrieval task of images and captions we leverage a prompt-based method to represent images and captions. We follow the methodology presented in (Jiang et al., 2023). We use template:

This caption: "[text]" means in one word: "

to prompt Qwen to generate next token and extract the hidden vectors of the final token as sentence embeddings. We use the same template for encoding images. We exploit contrastive learning on Qwen, which has been demonstrated as an efficient way to learn sentence embeddings (Gao et al., 2021). We use random subset of MS-COCO dataset where each image x_i has corresponding a positive caption x_i^+ and a random negative caption x_i^- from dataset.

We then use last token hidden state as embedding of image or caption h_i. The training objective is following:

$$l_i = -\log \frac{e^{\cos(h_i, h_i^+)/\tau}}{\sum_{j=1}^{N}(e^{cos(h_i, h_j^+)/\tau} + e^{cos(h_i, h_j^-)/\tau})}$$

where N is the batch size and τ is the temperature hyperparameter. All the hyperparameters of the fine-tunning and stated in Appendix D.

3.2 Evaluation Metrics of Retrieval Tasks

Standard retrieval metrics such as success@K or precision@K assume fully annotated datasets, where all correct matches are labeled. In practice, many semantically valid matches are unlabeled, leading these metrics to underestimate model performance. To address this, we introduce three complementary evaluation strategies tailored for post-retrieval probing.

3.2.1 Semantic-Similarity Success

This metric enables soft matching by assessing the similarity between retrieved candidates and the gold reference. It is most helpful in evaluating the understanding of relations. Under semantic-similarity success, match correctness is assessed by comparing the caption (query/candidate) against the gold reference caption corresponding to the (candidate/query) image. Both captions are decomposed into (subject, relation, object) triplets. We first check if the verbs match exactly. If yes, we compute cosine similarity between subject and object in candidate-query pairs using E5-large embeddings. If both exceed a threshold (we use 0.9), the candidate is considered a match - the similarity score between both subjects and objects (woman vs. girl, road vs. street) in Figure 1 surpasses a defined threshold.

3.2.2 Human Judgment Success

We conduct human evaluation on a sample of top-K text retrievals to establish ground-truth correctness. Given a query (image or caption), annotators rate whether each retrieved item is a valid match. This captures real-world semantic plausibility beyond what is labelled in the dataset.

Each item was annotated by three volunteers, non-native English speakers with tertiary education. Final correctness is determined by majority vote. This serves both as a gold reference and as a benchmark to evaluate automated metrics.

3.2.3 Model-based Evaluation

To scale the evaluation process, we use the o3 reasoning model to classify whether a retrieved item matches the query. We tested the performance of different models, such as GPT-4o, o3-mini, and o4-mini; however, o3 had the best alignment with human annotators. The model is prompted with both the query and the top-K retrieved items and asked to label each as correct or incorrect. It can also specify the type of mismatch (subject, verb, object) for more detailed error analysis.

4 Analysis of Role Understanding

We evaluate our post-retrieval probing method on five vision-language models—CLIP, FLAVA, and SigLIP2 (from the Transformers library), BLIP-2 (from the LAVIS library), and Qwen2.5-VL (on HuggingFace)—which represent a range of multimodal architectures. Our experiments utilize SVO-Probes (Hendricks and Nematzadeh, 2021), a publicly available dataset designed to assess understanding of subject-verb-object structure.

4.1 SVO-Probes: Probing on Subjects, Verbs, and Objects

The SVO-Probes dataset (Hendricks and Nematzadeh, 2021) evaluates whether VLMs comprehend the roles of subjects, verbs, and objects in visual scenes. Each example in the dataset consists of a caption, a positive image, and a contrastive negative image. The negative image in each triplet corresponds to one of three types: subject negative, verb negative, or object negative. This structured format allows for fine-grained probing of role comprehension. An example is shown in Figure 1. SVO-Probes is commonly used for ITM. We were able to download 13855 unique images, the dataset includes 10978 unique captions, organized into 33686 triplets of positive image, negative image, and caption.

4.2 Pairwise Matching and Image-Text Matching

Understanding verbs appears to be largely solved, with all models achieving over 80% on pairwise accuracy (see Table 1). The highest accuracy was observed for object negatives, followed by subjects, with verb negatives proving most difficult. BLIP-2 and SigLIP2 outperformed CLIP by more than 3%, even though they utilize CLIP as a backbone. Overall, BLIP-2, FLAVA, and SigLIP2

exhibit comparable performance. Qwen, evaluated through question-answering tasks involving two images and one caption, achieved an impressive performance of over 90%. However, when examining ITM accuracy in Table 2 (evaluated only on models that have a pretrained binary ITM classification head or visual question answering), we observed a significant performance drop.

	Overall	Subject	Verb	Object
CLIP	84.15	86.51	81.98	88.98
BLIP-2	87.58	89.98	85.27	92.80
FLAVA	87.07	88.44	84.58	93.66
SigLIP2	87.45	86.40	86.07	92.50
Qwen2.5-VL	91.69	93.09	90.43	94.50

Table 1: Pairwise ranking accuracy on SVO-Probes. The classification is based on which image embedding (positive or negative image) has a higher similarity score with the caption.

	Overall	Subject	Verb	Object
BLIP-2	63.59	63.91	63.68	63.08
FLAVA	62.40	63.76	60.79	66.30
Qwen2.5-VL	82.85	84.97	81.74	84.67

Table 2: Accuracy of image-text matching on SVO-Probes. The classification is based on binary classification head of the model.

4.3 Experiments with Retrieval

To test the model's general ability to perform retrieval, we propose the initial experiment on the SVO-Probes dataset. We generated query captions using the template: *This is the image of {subject}.*, where {subject} was substituted by one of 207 subjects in the dataset vocabulary. The model's performance was poor when assessed using traditional metrics, as shown in Table 3. However, when we applied semantic-similarity success@1, the performance improved to $72-85\%$. The best-performing models were BLIP2 and fine-tuned Qwen2.5-VL, both of which are among the latest models we tested. Interestingly, Qwen initially appeared to be the worst-performing model according to the success evaluation, it was actually the top performer when evaluated with semantic similarity. This indicates that fine-tuning of the model was effective, allowing it to create meaningful embeddings.

In the full image retrieval task, which requires a joint understanding of subjects, verbs, and objects, models performed poorly under the standard metric as well as the semantic-similarity success metric. S@1 falls below 10% and even

success@	1	5	10	20
CLIP	34.93	54.07	59.81	64.11
BLIP-2	42.58	63.64	70.33	77.51
FLAVA	36.36	58.85	65.07	72.73
SigLIP2	44.98	63.64	72.73	77.99
Qwen2.5-VL	34.45	59.33	65.07	74.64
semantic-similarity success@				
CLIP	72.73	93.78	98.09	100
BLIP-2	84.69	97.13	99.52	99.52
FLAVA	74.16	98.09	100	100
SigLIP2	76.56	99.52	100	100
Qwen2.5-VL	83.73	97.13	99.04	100

Table 3: Success@K and semantic-similarity success@K evaluation with threshold 0.9 of image retrieval on SVO-Probes dataset, using only captions created with following template: "This is the image of {subject}."

though using semantic-similarity success (Ss@1) improves performance by roughly $7-12\%$, absolute scores remain very low, hovering around 20% in Table 4. This indicates a significant challenge in comprehending fine-grained roles. Among the models, SigLIP2 performs best overall and notably outperforms CLIP, while FLAVA shows modest gains when evaluated semantically. Fine-tuned Qwen, on the other hand, lags behind the other models.

success@	1	5	10	20
CLIP	8.67	23.77	32.84	42.59
BLIP-2	9.76	27.04	37.91	49.84
FLAVA	8.61	24.35	35.65	46.74
SigLIP2	10.43	28.56	39.11	50.61
Qwen 2.5-VL	7.58	19.48	27.34	36.86
semantic-similarity success@				
CLIP	20.28	45.14	56.41	66.73
BLIP-2	22.61	52.40	65.22	75.66
FLAVA	24.90	54.60	66.93	77.39
SigLIP2	22.48	50.33	62.73	73.17
Qwen2.5-VL	16.77	38.02	49.44	61.29

Table 4: Success@K and semantic-similarity success@K evaluation with threshold 0.9 of image retrieval on SVO-Probes dataset.

A breakdown of retrieval by role revealed that subjects are the easiest to retrieve, followed by verbs, with objects being the most difficult. Analyzing retrieved samples by their ground-truth captions Table 5, the percentage of retrieved examples with the same subject as in the query was highest across models, followed by verbs and then by objects. This pattern contrasts with ITM, where object negatives were the easiest to detect. The only exception is Qwen. Although the occurrence of correct subjects in retrieved examples for Qwen was comparable to that of the other models, the analysis revealed that Qwen struggles significantly with objects and even more with verbs.

	K=1	K=5	K=10	K=20
Subject				
CLIP	53.17%	47.75%	44.56%	41.28%
BLIP-2	53.32%	48.89%	46.28%	43.19%
FLAVA	50.88%	47.10%	44.42%	41.62%
SigLIP2	53.75%	48.84%	45.87%	42.56%
Qwen2.5-VL	55.64%	51.12%	48.35%	45.97%
Verb				
CLIP	46.78%	41.01%	37.93%	34.69%
BLIP-2	45.14%	39.76%	36.68%	33.31%
FLAVA	43.64%	37.38%	34.17%	30.72%
SigLIP2	49.80%	43.21%	39.14%	35.07%
Qwen2.5-VL	39.77%	32.85%	29.85%	26.72%
Object				
CLIP	43.25%	37.79%	33.73%	29.47%
BLIP-2	41.85%	36.31%	33.17%	29.27%
FLAVA	42.69%	36.87%	33.78%	30.02%
SigLIP2	48.73%	41.00%	36.47%	31.81%
Qwen2.5-VL	42.30%	32.73%	30.79%	26.42%

Table 5: Subject, verb and object success@K on the SVO-Probes dataset – i.e. how many of the top K retrieved images correctly match the query subject, verb and object in their caption.

	K=1	K=5	K=10	K=20
Subject				
CLIP	46.18%	43.91%	41.92%	39.39%
BLIP-2	56.85%	52.01%	48.96%	45.50%
FLAVA	50.93%	46.88%	44.10%	41.10%
SigLIP2	57.05%	52.27%	48.83%	45.00%
Qwen2.5-VL	55.40%	52.23%	49.39%	45.98%
Verb				
CLIP	34.54%	31.04%	28.95%	26.58%
BLIP-2	49.85%	43.76%	39.62%	35.37%
FLAVA	37.19%	32.28%	29.65%	26.66%
SigLIP2	51.54%	45.02%	40.71%	36.02%
Qwen2.5-VL	41.67%	35.61%	31.99%	28.23%
Object				
CLIP	31.23%	28.94%	26.94%	24.32%
BLIP-2	44.30%	39.34%	35.88%	31.47%
FLAVA	35.71%	31.94%	29.25%	26.02%
SigLIP2	45.68%	40.66%	36.62%	31.77%
Qwen2.5-VL	38.98%	34.66%	31.42%	27.71%

Table 7: Subject, verb and object success at K across SVO-Probes dataset – how many of the top K retrieved captions have the query subject, query verb and query object in them.

Text retrieval was even more challenging than image retrieval. This difficulty likely arised from the nuanced nature of the captions in SVO-Probes, which often differ only in a single word. SigLIP2 and BLIP-2 model outperformed the rest in Table 6. Surprisingly, fine-tuned Qwen ranked as the second-best when evaluated with semantic similarity, even though this was not the case for image retrieval. Further analysis of the retrieved captions showed the trend holds for text retrieval as well: subjects are consistently easier to retrieve than objects. CLIP performed the worst across all roles —suggesting broader difficulty in grounding fine-grained linguistic elements in visual input.

success@	1	5	10	20
CLIP	6.47	20.46	30.29	42.16
BLIP-2	13.23	35.32	47.11	59.51
FLAVA	5.63	17.22	24.94	34.79
SigLIP2	13.42	35.95	48.14	61.16
Qwen2.5-VL	10.36	27.41	37.39	47.56
semantic-similarity success@				
CLIP	11.82	29.87	40.93	52.37
BLIP-2	24.99	47.39	58.12	67.62
FLAVA	17.42	38.19	50.00	60.56
SigLIP2	48.29	74.04	82.08	88.20
Qwen2.5-VL	39.16	63.72	73.19	80.71

Table 6: Success@K and semantic-similarity success@K evaluation with threshold 0.9 of text retrieval on SVO-Probes dataset.

4.4 Human and Model-based Evaluation

Human evaluation suggested that text retrieval performance is significantly better than success@K showed; however, the subject-verb-object understanding is still a challenging task. To assess how well retrieval results align with human expectations, we randomly sampled 100 queries from the dataset. For each image query, the top-10 retrieved captions by each model were annotated by three human annotators, who judged whether each caption was a valid match. Final labels were determined via majority vote. We would like to point out that over 3000 image-caption pairs had to be evaluated, each by 3 annotators just to obtain the results in Table 8. A significant increase, in some cases, a 60% difference, can be seen for all models once a human annotator evaluated them. SigLIP2 was the best-performing model for subject-verb-object understanding, with $S_h@1 = 80.81\%$. The Fleiss' Kappa between three human annotators was 0.641, which suggests moderate agreement between the human annotators.

O3 model can very well approximate the human judgment process for subject-verb-object understanding. It is important to note that some examples could not be evaluated by o3 as the model assumed they were violating the rules of use. We report the performance in Table 8 only for examples that could also be evaluated by o3. For each query, we used a prompt that can be seen in Appendix A to evaluate the models. The results show that o3 slightly underestimated the performance of models on SVO-Probes. We then measured the o3 agreement with human annotations using the F1 score, with the majority human vote serving as the ground truth. The F1 score was 0.8368, which suggested high alignment between the human majority vote and the o3 model.

An example of retrieved examples and their human and o3 evaluation can be seen in the Appendix in E. The evaluations were also carried out for the image retrieval task, the results can be seen in Appendix B. However, for the sake of time and resources, only one annotator evaluated results for image retrieval.

		SVO-Probes		
		K=1	K=5	K=10
CLIP	S@K	8.08	18.18	28.28
	S_h@K	68.69	96.97	100.00
	S_{o3}@K	46.46	94.95	98.99
	P@K	8.08	3.64	2.83
	P_h@K	68.69	62.02	57.98
	P_{o3}@K	46.46	49.29	47.78
BLIP-2	S@K	16.16	41.41	52.53
	S_h@K	77.78	94.79	96.97
	S_{o3}@K	76.77	96.97	100.00
	P@K	16.16	8.28	5.35
	P_h@K	77.78	69.29	63.33
	P_{o3}@K	76.77	61.21	54.85
FLAVA	S@K	6.06	23.23	31.31
	S_h@K	62.63	90.91	95.96
	S_{o3}@K	59.60	93.94	95.96
	P@K	6.06	4.65	3.13
	P_h@K	62.63	54.34	48.99
	P_{o3}@K	59.60	51.31	43.54
SigLIP2	S@K	20.20	44.44	59.60
	S_h@K	80.81	96.97	96.97
	S_{o3}@K	71.72	97.98	97.98
	P@K	20.20	8.89	5.96
	P_h@K	80.81	70.30	66.46
	P_{o3}@K	71.72	58.89	49.89

Table 8: Results of text retrieval on a subset of 100 samples from SVO-Probes. Standard success (S@K), success with human evaluation (S_h@K) and success with o3 model evaluation (S_{o3}@K). The F1 score between human majority vote and o3 evaluation is 0.8368. Cohen's kappa between human majority vote and o3 evaluation is 0.6457.

The same conclusions can be made about fine-tuned Qwen evaluation in Table 9. As only one human annotator evaluated results for Qwen, we provide the results separately. The performance of Qwen was very similar to CLIP with $S_h@1 = 69.70\%$. The agreement between the human annotator and the o3 model was slightly lower than for the rest of the models; however, it still showed significant agreement with an F1 score of 0.79.

4.5 Post-Retrieval Error Analysis

To better understand the types of errors made by VLMs, we performed a qualitative analysis of the top-K retrievals on the SVO-Probes dataset. Given the moderate agreement between human raters and o3 $F1 = 0.84$, we used o3 to label retrievals as correct, or incorrect due to one of three reasons: subject (the subject in the caption does not match the image), verb (the activity described in the caption does not match the image), or object (other

		SVO-Probes		
		K=1	K=5	K=10
Qwen	S@K	13.13	30.30	38.38
	S_h@K	69.70	93.94	94.95
	S_{o3}@K	62.63	83.84	92.93
	P@K	13.13	6.06	3.84
	P_h@K	69.70	61.41	58.08
	P_{o3}@K	62.63	49.09	46.46

Table 9: Results of text retrieval on a subset of 100 samples from SVO-Probes. Standard success (S@K), success with human evaluation (S_h@K) and success with o3 model evaluation (S_{o3}@K). The F1 score between human annotation and o3 evaluation is 0.7942. Cohen's kappa between human annotation and o3 evaluation is 0.5746.

details in the caption do not align with the image) mismatch.

Post-retrieval analysis supported that verb comprehension is still challenging for models, and confirmed that strong ITM or pairwise scores do not guarantee robust role comprehension. We analyzed the top 10 retrieved samples for 100 randomly sampled queries across the models. For each retrieval, o3 identifies the type of error that occurs when a mismatch is detected. Our findings in Table 10 reveal consistent patterns: Subject mismatches were the least frequent across all models, suggesting that models grasp subject identity more reliably than actions or objects. Verb errors were the most frequent, particularly in Qwen for image retrieval, and in FLAVA for text retrieval. The ability to retrieve similar but incorrect items (e.g., matching subject and activity but not the object) highlights semantic ambiguity and model brittleness—critical aspects not captured by standard evaluation. A fascinating insight was a significant difference in performance in image and text retrieval for FLAVA and Qwen (the difference in correctly retrieved examples in image and text retrieval is bigger than 10%). This suggested a disbalance in the model's ability to encode different modalities.

Using retrieved examples for repeated ITM led to significant drop in performance, which illustrates that standard ITM evaluation can mask important model failures. To further test the models abilities, we treated the top-K retrieved examples as a comprehensive contrastive dataset and re-run ITM on these examples using BLIP-2, FLAVA, and the Qwen2.5-VL model. We evaluated these matches in Table 11. While BLIP-2 and FLAVA previously achieved over 60% ITM accuracy and over 80% pairwise accuracy, their performance droped on the retrieved examples. Cru-

		Image retrieval		Text retrieval	
		Human	o3	Human	o3
CLIP	Correct	51.63%	46.84%	57.98%	47.78%
	Subject	10.92%	11.22%	9.80%	13.13%
	Verb	**19.80%**	**25.41%**	**16.67%**	**30.10%**
	Object	17.65%	16.53%	15.56%	8.99%
BLIP2	Correct	61.38%	54.68%	63.33%	54.85%
	Subject	9.26%	10.85%	6.06%	10.51%
	Verb	**19.79%**	**21.60%**	**17.88%**	**20.51%**
	Object	9.57%	12.45%	12.63%	14.14%
FLAVA	Correct	58.49%	48.92%	48.99%	43.53%
	Subject	12.04%	12.80%	10.00%	14.34%
	Verb	**20.00%**	**26.02%**	**26.67%**	**29.09%**
	Object	9.46%	12.26%	14.34%	13.03%
SigLIP2	Correct	62.42%	56.97%	66.46%	58.89%
	Subject	11.31%	11.82%	9.90%	13.43%
	Verb	**16.16%**	**18.69%**	**13.84%**	**14.85%**
	Object	10.10%	12.53%	9.70%	12.83%
Qwen	Correct	43.26%	38.91%	58.08%	46.46%
	Subject	8.59%	9.24%	4.75%	6.77%
	Verb	**31.63%**	**33.04%**	**25.35%**	**29.39%**
	Object	16.52%	18.80%	11.82%	17.37%

Table 10: Analysis of correct and incorrect samples, retrieved by different models. The classification was done using o3 model. Cohens kappa for image retrieval is 0.6323 and the weighted F1 score is 0.7689 and for text retrieval Cohen's kappa is 0.5529 and the weighted F1 score is 0.7276.

cially, retrieved examples include: (i) true positives; and (ii) false positives that are the most likely to be confused with true positives by a specific model. This striking drop illustrates the diagnostic value of post-retrieval probing. The performance of Qwen slightly improved for image retrieval examples compared to the overall SVO-Probes dataset. We argue that the reason for that was that the image retrieval examples by Qwen were not ideal. We can see that for text retrieval in Table 10, Qwen demonstrated better performance, with significantly more correct retrieved samples, compared to image retrieval. Therefore, new image-caption pairs obtained by text retrieval were more challenging compared to image retrieval pairs.

	Image retrieval	Text retrieval
BLIP2	59.70%	49.60%
FLAVA	56.50%	52.90%
Qwen	84.00%	74.40%

Table 11: Image-text matching accuracy on examples retrieved by specific model in top 10. The human annotation is considered as ground truth

5 Conclusion

In this work, we introduce a probing framework based on post-retrieval analysis to more effectively assess the linguistic understanding of vision-language models. Unlike standard methods, our approach does not rely on handcrafted contrastive pairs. Instead, it leverages the retrieval task, recognizing that the top-K retrieved items include both true positives and close false positives—offering a more diagnostic view of model behavior.

Our evaluation of the frequently used SVO-Probes dataset revealed significant gaps in model performance that are obscured by image-text matching scores. While all models - CLIP, BLIP2, FLAVA, SigLIP2, and Qwen2.5-VL - achieved over 80% accuracy in the pairwise accuracy task, their success@1 dropped below 10% on the retrieval task. Semantic similarity metrics partially recovered this performance, suggesting that some"errors" are due to incomplete annotation rather than model failure, however the performance still remained low.

To fully address annotation incompleteness, we also evaluated retrieval with human judgment and the o3 reasoning model scoring, which better captures the range of valid retrievals. o3 achieved moderate agreement with humans ($F1 = 0.84$) for SVO-Probes, demonstrating potential as a scalable evaluation proxy. However, even after using human annotations for evaluation, the task of subject-verb-object comprehension remains challenging with S_h@1 around $62 - 80\%$.

By subjecting top-K retrievals to further analysis, we uncovered systematic model weaknesses in role understanding. Analysis showed that models consistently retrieved subjects more accurately than verbs or objects, contradicting pairwise matching patterns in which object negatives were found to be the easiest for models to classify. Also rerunning the image-text matching on retrieved examples leads to significant drop of performance. These insights underscore the need for this in-depth, retrieval-based probing methods for more robust analysis of multimodal models.

Computational Resources

For training, inference and evaluation, our experiments were run on NVIDIA GeForce RTX 3060 Laptop GPU (6GB VRAM). Total compute was approximately 150 GPU hours (mostly from running retrieval across the dataset, multiple times per model). In addition, we utilized the `o3` reasoning model accessed via Microsoft Azure for evaluation task.

Limitations

While our post-retrieval probing framework offers new insights into the linguistic behavior of vision-language models, it also has several limitations.

First, our analysis is limited to the SVO-Probes dataset, which is synthetic and has a constrained structure. While it targets key linguistic phenomena, it may not fully capture the diversity or ambiguity found in real-world multimodal data.

Second, although we introduced alternative evaluation strategies—semantic similarity, human annotation, and o3 as a judge—they each have their trade-offs. Semantic similarity depends on thresholding and does not help evaluate aspects such as counting; the o3 model, while scalable, does not always align perfectly with human judgment.

Additionally, we evaluated only pre-trained models, with architectures suitable for encoding captions and images, and therefore capable of performing retrieval tasks using cosine similarity as the similarity score. We fine-tuned only one LVLM - Qwen2.5-VL-7B - to also evaluate the capabilities of a state-of-the-art model with LLM. Finally, all experiments were conducted in English, which limits the generalizability of the findings to multilingual settings.

We leave these avenues for future work.

References

Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. 2024. Generative information retrieval evaluation. In *Information Access in the Era of Generative AI*, pages 135–159. Springer.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ivana Beňová, Michal Gregor, and Albert Gatt. 2024. Cv-probes: Studying the interplay of lexical and world knowledge in visually grounded verb understanding. *arXiv preprint arXiv:2409.01389*.

Ivana Benova, Jana Kosecka, Michal Gregor, Martin Tamajka, Marcel Vesely, and Marian Simko. 2025. Beyond image-text matching: Verb understanding in multimodal transformers using guided masking. In *International Conference on Current Trends in Theory and Practice of Computer Science*, pages 80–93. Springer.

Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. 2023. Measuring progress in fine-grained vision-and-language understanding. *arXiv preprint arXiv:2305.07558*.

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. 2022. Eccv caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In *European Conference on Computer Vision*, pages 1–19. Springer.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*.

Hexiang Hu, Ishan Misra, and Laurens Van Der Maaten. 2019. Evaluating text-to-image matching using binary image selection (bison). In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Open clip.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2023. Scaling sentence embeddings with large language models. *arXiv preprint arXiv:2307.16645*.

Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and Erkut Erdem. 2024. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, Vienna, Austria.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023a. Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. 2023b. CREPE: Can Vision-Language Foundation Models Reason Compositionally? pages 10910–10921.

OpenAI. 2024. Gpt-4o. Accessed: 2025-03-03.

OpenAI. 2025. OpenAI o3. https://openai.com/index/introducing-o3-and-o4-mini/. [Large language model].

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2020. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. *arXiv preprint arXiv:2012.12352*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650.

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

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, Tatiana Anikina, Michal Gregor, and Marián Šimko. 2025. Large language models for multilingual previously fact-checked claim detection. *arXiv preprint arXiv:2503.02737*.

Yang Wang, Alberto Garcia Hernandez, Roman Kyslyi, and Nicholas Kersting. 2024. Evaluating quality of answers for retrieval-augmented generation: A strong llm is all you need. *arXiv preprint arXiv:2406.18064*.

Appendix

A Prompting o3 Model for Retrieval Evaluation

To scale up retrieval evaluation, we use o3 reasoning model to assess whether each top-K retrieved item (image or caption) is a correct match for a given query. For SVO-Probes and VALSE-counting datasets, we prompt o3 with the query and each retrieved item, asking it to classify them as correct or incorrect. For incorrect matches, o3 is instructed to specify the error type.

The full prompt used for caption-based queries is shown below. A similar structure is applied for image-based queries in both datasets.

Assess if given images are correct retrievals for the text query provided caption.
*For each image, evaluate if it is correct. If it is incorrect, mention the specific category that best describes the error: - **Subject incorrect**: The subject in the caption does not match the image. - **Verb incorrect**: The activity described in the caption does not match the image. - **Object incorrect**: Other details (e.g., objects or contextual elements) in the caption do not align with the image.*
 # Steps
*1. For each image, evaluate its correctness based on text query. 2. If the image aligns well with the caption, classify it as 'correct'. 3. If it is incorrect, determine the category of the error: - **Subject incorrect** - **Verb incorrect** - **Object incorrect** 4. Output results using a structured, simple list.*
 # Output Format
The results should be listed in this format: - $\langle image_number \rangle$: $\langle classification \rangle$
For example: '1: correct' '2: verb incorrect' '3: object incorrect'.

Table 12: Prompt for o3.

B Image retrieval results

This section presents success@K results for image-to-text retrieval on SVO-Probes , using standard evaluation, human annotation, and o3 assessment. The results are based on 100 randomly sampled image queries for each dataset.

As shown in Table 14, human evaluation consistently yields much higher success@K than standard metrics, revealing the impact of incomplete dataset annotation and o3 estimates closely align with human scores.

We also report precision@K for completeness.

C Instructions for Annotators

Following instructions were provided for each annotator before the annotation process.

Your task is to annotate the relation between the image and captions. Could the caption conceivably describe the image? Does the caption match the image? More than one caption can be correct for a given image. The annotation schema has three options (Yes - No - ?), and "?" with a meaning that you can not tell. Try to use "?" only in cases when you are really not sure about the relationship between image and caption.

- Yes (Relevant) - The caption could match the image fully.
- No (Irrelevant) - The caption contains some part that does not match the image. The subject, verb, object or the number is incorrect (if the image shows man and the caption mentions "woman", the pair is incorrect, if the image shows a man on a red carpet and the caption mentions "actor", assume general knowledge that the man is in fact an actor.)
- ? (Can not tell) You can not tell. Please choose this option only when you are really in the

		SVO-Probes		
		K=1	K=5	K=10
CLIP	S@K	10.20	25.51	35.71
	S_h@K	63.27	88.78	96.94
	S_{o3}@K	65.31	88.78	93.88
	P@K	10.20	5.51	4.18
	P_h@K	63.27	57.55	51.63
	P_{o3}@K	65.31	53.06	46.84
BLIP-2	S@K	10.64	29.79	39.36
	S_h@K	74.47	94.68	95.74
	S_{o3}@K	70.21	93.62	95.74
	P@K	10.64	5.96	4.04
	P_h@K	74.47	68.94	61.38
	P_{o3}@K	70.21	61.70	54.68
FLAVA	S@K	9.68	26.88	36.56
	S_h@K	74.19	95.70	95.70
	S_{o3}@K	68.82	93.55	95.70
	P@K	9.68	5.38	3.66
	P_h@K	74.19	62.37	58.49
	P_{o3}@K	68.82	52.69	48.92
SigLIP2	S@K	13.13	31.31	43.43
	S_h@K	74.75	93.94	95.96
	S_{o3}@K	75.76	92.93	95.96
	P@K	13.13	6.26	4.55
	P_h@K	74.75	68.69	62.42
	P_{o3}@K	75.76	63.84	56.97

Table 13: Results of image retrieval on a subset of 100 samples from SVO-Probes. Standard success (S@K), success with human evaluation (S_h@K) and success with o3 evaluation (S_{o3}@K). The F1 score between human annotator and o3 evaluation is 0.8459.

		SVO-Probes		
		K=1	K=5	K=10
Qwen	S@K	7.61	16.30	26.09
	S_h@K	64.13	85.87	93.48
	S_{o3}@K	60.87	85.87	89.13
	P@K	7.61	3.26	2.61
	P_h@K	64.13	46.74	43.26
	P_{o3}@K	60.87	43.48	38.91

Table 14: Results of text retrieval on a subset of 100 samples from SVO-Probes. Standard success (S@K), success with human evaluation (S_h@K) and success with o3 model evaluation (S_{o3}@K). The F1 score between human majority vote and o3 evaluation is 0.8201. Cohen's kappa between human annotation and o3 evaluation is 0.6952.

middle of yes and no. Always try to choose "yes" or "no".

The expected time for evaluating one pair of an image and caption is approximately 5 seconds, while each document contains multiple pairs of the same image and different captions. Your task is to evaluate the relevance between each caption provided and the image, which is at the top of the web page.

Registration and annotating data

1. First of all, it is necessary to register into the annotation tool, which can be accessed at this URL. To register, you need to input an email address and password. Then click on "Register". It is necessary to remember the email and password because these credentials will be used to log into the system using the same user interface.

2. Login - To log into the system, provide your email address and password and click on the "Login" button.

3. After logging into the system, you will see the interface below, where you should click on "List of documents 2" to get into the list of all social media posts that should be annotated

4. The list of all images for annotating is shown below

5. You are supposed to annotate the data based on the information provided.

Process

- Checking the image - In the annotation tool, you will see the image. Please check it carefully. You can enlarge it by clicking on the image.

- Read a caption - For each image, there will be multiple captions, but the number of captions differs for each image. Please, read carefully the caption and based on the image, annotate the relevance between caption and image as mentioned above. By selecting one of the options (Yes, No, ?) you will annotate the specific pair of caption and image. After providing annotations for each caption (take each caption as an individual task, multiple captions can be relevant for one image), click the SUBMIT button at the bottom of the page, other.

Four examples of retrieval were also shown as part of instructions, together with correct annotations and rational.

D Details for Qwen2.5-VL Fine-Tuning and Evaluation

D.1 The Hyper Parameters

For the fine-tuning of Qwen2.5-VL-7B-Instruct model, we leveraged current efficient fine-tuning method QLoRA with contrastive learning. Following (Jiang et al., 2023), we used $r = 64$, $\alpha = 16$, $dropout = 0.05$. We fine-tuned Qwen on one epoch, temperature $\tau = 0.07$ and learning rate $2e-4$. Due to hardware limitations, we only conducted our experiments with batch size 4. We used 500 pairs of image-caption pairs from MS-COCO (Lin et al., 2014) training split.

D.2 Prompts

The full prompt used for evaluation of Qwen2.5-VL-7B are shown below: image-text matching in 15 and pairwise ranking in 16.

You are an image caption expert. Does this image match the caption: 'caption'? Answer yes only if everything from the caption - subject, verb and object - matches the image, otherwise answer no.

Table 15: Prompt for image-text matching with Qwen2.5-VL.

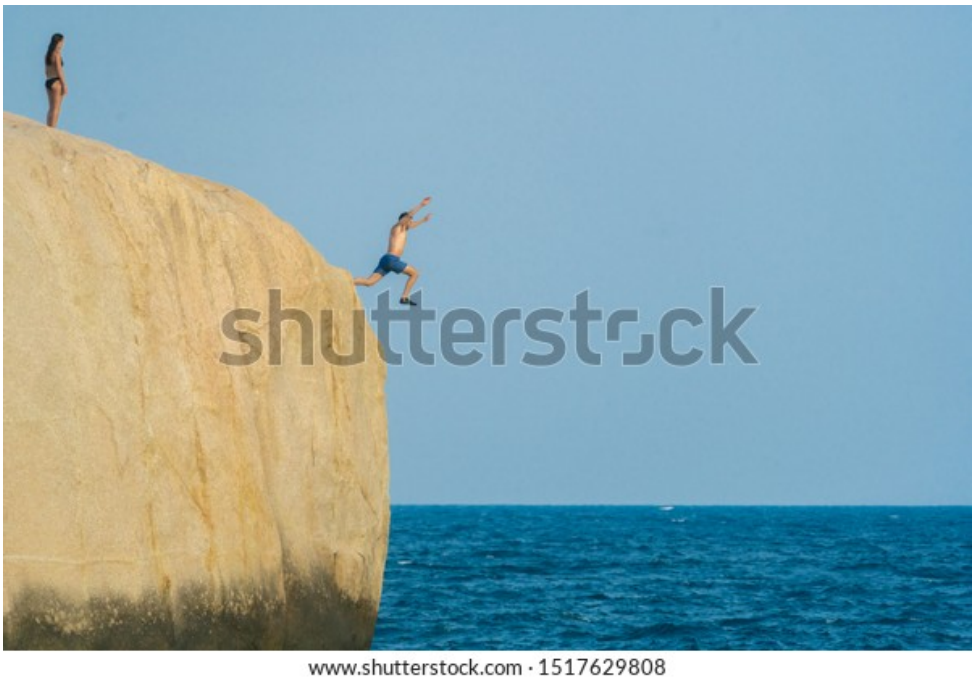
You are an image caption expert. Two images are given above. The first image is IMAGE_A, the second image is IMAGE_B. Caption: 'caption'. Question: Which image matches the caption better? Answer with just 'A' if IMAGE_A matches better, or just 'B' if IMAGE_B matches better.

Table 16: Prompt for pairwise ranking with Qwen2.5-VL.

E Examples of Human and o3 evaluation of retrieval

We provide examples of top-10 caption retrievals from the SVO-Probes dataset, along with correctness labels from human annotators and o3. These examples illustrate the kinds of captions retrieved and how human and LLM judgments agree or differ.

Each row shows a retrieved caption and binary labels from both evaluators. ✔indicates a match between the image and caption; ✖indicates a mismatch. These examples highlight o3's general alignment with human intuition in SVO-Probes, and also showcase challenging cases.



	CLIP			BLIP-2			FLAVA		
Order	Caption	Human	o3	Caption	Human	o3	Caption	Human	o3
1	A man is jumping off a cliff.	✔	✔	A person jumps near the sea.	✔	✔	A man is jumping off a cliff.	✔	✔
2	A man jumps off a rock	✔	✔	Boys jumping off of a rock	✖	✖	Boy is jumping from a rock.	✔	✖
3	a man walking on an edge	✖	✖	couple jump on the beach	✖	✖	The man sits on the rock.	✖	✖
4	Boy is jumping from a rock.	✔	✔	The woman jumps off the rock.	✖	✖	A man stands on the rock.	✖	✖
5	A man is climbing a cliff.	✖	✖	A couple jumps on the beach.	✖	✖	the person rests on the rock	✖	✖
6	The man is standing at the edge of a cliff.	✖	✖	A person jumps at the sea.	✔	✖	A man is climbing a rock.	✖	✖
7	a man is about to jump into the water	✔	✔	A man takes a jump into the sea.	✔	✔	a man jumping in the background	✔	✔
8	A man jumps off a rock.	✔	✔	The woman jumps the cliff.	✖	✖	The girl is climbing a rock.	✖	✖
9	A person climbing to the top of a cliff	✖	✖	The couple jumps on the beach.	✖	✖	A person sitting on a cliff	✖	✖
10	a person takes a jump	✔	✔	Boy is jumping from a rock.	✔	✔	The woman jumps off the rock.	✖	✖

Table 17: CLIP, BLIP-2, and FLAVA retrieved the top 10 captions for image from the SVO-Probes dataset. The columns Human and GPT contain human and o3 annotations for specific captions, whether or not they match the image.

	CLIP			BLIP-2			FLAVA		
Order	Caption	Human	o3	Caption	Human	o3	Caption	Human	o3
1	this person makes the image	✗	✗	A couple kiss in a meadow.	✓	✓	A couple standing on the meadow	✓	✓
2	A couple expecting a baby	✓	✓	A couple standing in a meadow.	✓	✓	The couple is standing in the field kissing.	✓	✓
3	the person looks nice	✓	✓	A couple standing on the meadow	✓	✓	A couple are sitting at a field.	✗	✗
4	The woman is using a camera.	✗	✗	A couple walking in the meadow.	✗	✗	A couple is embracing each other.	✓	✓
5	A woman is expecting a baby.	✓	✓	A couple walking on the meadow	✗	✗	A man and woman are kissing.	✓	✓
6	The woman has a look on her face.	✓	✓	A couple expecting a baby	✓	✓	A man kissing a woman.	✓	✓
7	The woman will lie in bed with her child.	✗	✗	A couple sits in a meadow.	✗	✗	A couple is laying in the grass.	✗	✗
8	The woman will take a photo with her camera.	✗	✗	A couple walks through a meadow.	✗	✗	The woman stood in the grass.	✓	✗
9	the child runs in the field	✗	✗	A couple lies in the meadow.	✗	✗	The man stands in the field.	✓	✗
10	The people look like a happy pair.	✓	✓	The couple is standing in the field kissing.	✓	✓	A couple kisses the other.	✓	✓

Table 18: CLIP, BLIP-2, and FLAVA retrieved the top 10 captions for image from the SVO-Probes dataset. The columns Human and GPT contain human and o3 annotations for specific captions, whether or not they match the image.