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

Anonymous ACL submission

Abstract

Vision-language models (VLMs) are often evaluated on linguistic understanding—such as verb recognition or object counting—using handcrafted datasets with contrastive image-caption pairs. However, these datasets rarely capture the full complexity of real-world language use. We propose a probing framework based on post-retrieval analysis, which evaluates a model's top-K retrievals, which reveals finer-grained weaknesses in model behavior. We evaluate four VLMs—CLIP, BLIP-2, FLAVA, and SigLIP2—on two datasets: SVO-Probes (probing subject-verb-object role understanding) and VALSE-counting (probing numerical comprehension). To mitigate the issue of incomplete retrieval dataset annotations, we complement traditional metrics with three strategies: semantic-similarity success@K, human evaluation, and GPT-4o-based assessment. Our findings show that while VLMs achieve high image-text matching accuracy ($> 80\%$), they struggle in top-K retrieval settings—verb and object understanding (success@1 $\approx$ 70%), but especially for counting (success@1 $\approx$ 35%). Furthermore, GPT-4o aligns moderately with human judgment on verb but fails on counting tasks. We conclude that standard evaluation methods may underestimate VLM capabilities, and post-retrieval probing offers a more robust and nuanced view of their linguistic understanding.

1 Introduction

Recent studies have used various probing methods (like uch as image-text matching, visual question answering or guided masking) to evaluate vision-language models' (VLMs) understanding of linguistic constructs. Probing studies have investigated 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), drawing intriguing conclusions. However, these studies often rely on handcrafted datasets that may not fully capture the complexity or breadth of the targeted phenomena.

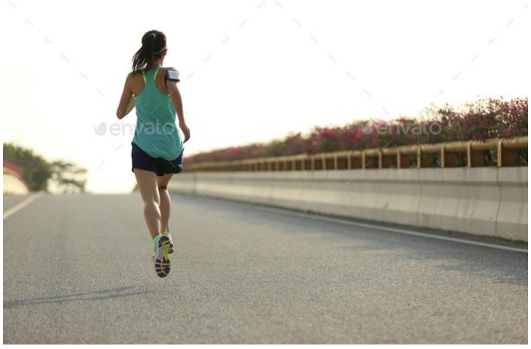


Figure 1: An example from the SVO-Probes dataset; an image labeled with the positive caption: *"A girl runs on a bridge."*, and with negative captions: *"The man will run on the bridge."*, *"A girl runs up a hill."* . Other captions in the dataset, which are semantically correct but would be evaluated as incorrect in the standard retrieval setup, e.g.: *" A woman running on the street."* or *"A person runs on the road."*

For example, image-text matching (ITM) results have led to claims that modern VLMs, such as BLIP-2, have effectively mastered verb understanding (Bugliarello et al., 2023). Yet these evaluations typically use manually constructed positive and negative image-caption pairs designed to highlight specific linguistic features (see Figure 1).

This reliance on hand-designed contrastive pairs raises 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, high performance on image-text matching (ITM) tasks may obscure significant weaknesses in fine-grained linguistic understanding. A comprehensive dataset of contrastive pairs would require annotators to assess all possible combinations or possess prior knowledge regarding which pairs are challenging for a particular model.

To address this, we propose post-retrieval analysis: instead of evaluating a model's binary predictions over fixed contrastive pairs, we examine the top-K items retrieved by the model for a given query. This setup surfaces near misses—false positives that the model finds highly similar to the true positives—and provides a richer, more diagnostic view of a model's linguistic understanding and robustness.

However, retrieval-based evaluation introduces 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" caption per image (or vice versa), valid but unlabeled matches are treated as errors under standard metrics like success@K. As an example, for Figure 1 correct retrieval is considered *"A girl runs on a bridge."* . If the model retrieved *"A woman running on the street."*, which is semantically correct, it would be evaluated as an error. To mitigate this, we complement retrieval with three alternative evaluation strategies: semantic similarity success (an alternative to success@K), human annotation, and GPT-4o-based analysis—to better account for unlabeled but valid matches and provide a more accurate assessment.

We apply this framework to evaluate four VLMs—CLIP (Ilharco et al., 2021), BLIP-2 (Li et al., 2023), FLAVA (Singh et al., 2022) and SigLIP2 (Tschannen et al., 2025)—on two probing datasets: SVO-Probes (Hendricks and Nematzadeh, 2021), for testing verb-role comprehension, and VALSE-counting (Parcalabescu et al., 2021), for numerical understanding.

While the selected models perform well on image-text matching on SVO-Probes (accuracy over 80%) and not so well on VALSE-Counting (accuracy 60%), they exhibit poor performance when evaluated using retrieval with standard success@K. The results improve when evaluating with semantic-similarity success, however, the task proves to be challenging even when evaluated by a human on a subset of 100 samples from the datasets: for SVO-Probes $\mathrm{S}_h@1$ is $60-80\%$, and the task is even more challenging for VALSE-counting with $\mathrm{S}_h@1$ $30-50\%$). Also we argue that not just verb understanding make the task challenging, but also object understanding.

While using GPT4-o as evaluator, we found moderate agreement with human annotations for the SVO-Probes dataset (F1 = 0.7252) but low agreement for counting (F1 = 0.4495). Therefore, GPT may be helpful for SVO-Probes, however, it seems unlikely that it would be helpful for counting.

Our main contributions are as follows:

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

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

3. We benchmark four VLMs across two probing tasks, showing that retrieval-based probing provides a more diagnostic view of model competence.

Results of our post-retrieval analysis indicate that there is still substantial room for improvement. Post-retrieval evaluation by humans measured success@1 on SVO-Probes around 70% – contrary to the findings with image-text matching. Moreover, the analysis by GPT-4o of incorrectly retrieved examples suggests that the object understanding is a similar if not more challenging as verb understanding. The results are even more sobering for counting, with human evaluation success@1 of only around 35%.

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) that combine frozen vision encoders with large language models. While these models perform well on downstream tasks, there has been 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 a text), the goal is to retrieve the most relevant corresponding items (texts or images) from a database. The output of a retrieval method is a list of samples ranked by relevance, from which the top K samples are used further.

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 image-text matching (ITM) or visual question answering, rely on curated datasets designed to isolate specific linguistic constructs—e.g., verbs (Hendricks and Nematzadeh, 2021; Benova et al., 2025), spatial relations (Liu et al., 2023), or counting (Parcalabescu et al., 2020, 2021). These typically use contrastive examples: a "positive" image-caption pair and a minimal "negative" pair that alters one key component (example can be seen in Figure 1). While useful, such datasets cover only a narrow slice of possible errors, and their construction reflects human priors about what should be 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. Furthermore, negative examples may not represent the most difficult edge cases.

As a result, it is perfectly possible for a model to both (i) perform well on the specific contrastive pairs formed by human annotators and (ii) 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 possible combinations or possess some prior knowledge regarding which pairs are likely to be found challenging.

We propose an alternative: *probing through post-retrieval analysis*. Instead of evaluating models only on binary image-text match judgments, we analyze the top-K retrieved items for each query. These include both correct matches and close distractors. This lets us inspect not just whether a model "gets it right," but what kinds of near-misses it tends to make. The procedure for text retrieval is described in (Algorithm 1). The same holds 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. 3.1.1, 3.1.2, 3.1.3
    Perform exploratory analysis on top K samples
```

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

3.1 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.

3.1.1 Semantic-Similarity Success

This metric allows for soft matching by evaluating how similar retrieved candidates are to 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.

3.1.2 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 labeled in the dataset.

Each item was annotated by three volunteers, non-native English speakers with tertiary eduction. Final correctness is determined by majority vote. This serves both as a gold reference and as a benchmark to evaluate automated metrics (e.g., GPT-4o agreement).

3.1.3 GPT-4o Evaluation

To scale the evaluation process, we use GPT-4o (OpenAI, 2024) to classify whether a retrieved item matches the query. GPT-4o 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 Experiments with Retrieval

We evaluate our post-retrieval probing method on four vision-language models— CLIP FLAVA and SigLIP2 (from transformers library) and BLIP-2 (from lavis library)—which represent a range of multimodal architectures. We conduct experiments on two public datasets: SVO-Probes, designed to assess understanding of subject-verb-object structure, and VALSE-counting, which focuses on numerical reasoning. We consider both image-to-text and text-to-image retrieval, as well as standard image-text matching.

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

The SVO-Probes dataset evaluates whether VLMs understand the roles of subjects, verbs, and objects in visual scenes. Each example 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 fine-grained probing of role comprehension. An example is shown in Figure 1. SVO-Probes is commonly used to benchmark image-text matching accuracy.

Verb understanding appears largely solved with all models achieving over 80% accuracy on image-text matching (see Table 1). Accuracy is highest for object negatives, followed by subject negatives, with verb negatives proving most difficult. BLIP-2 and SigLIP2 outperform CLIP by more than 3%, despite using CLIP as a backbone. Overall, BLIP-2, FLAVA, and SigLIP2 perform comparably well.

	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

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

On a simple retrieval task, models perform poorly when evaluated with traditional metrics, however, performance improves when semantic-similarity success is used. We begin with a simple retrieval task to assess how challenging retrieval evaluation is. We generate a query captions using the template: *This is the image of {subject}.*, where {subject} is substituted from the dataset vocabulary. Despite the simplicity of this setup, models in Table 2 achieve only $35-45\%$ success@1 (S@1). However, when evaluated with semantic-similarity success@1 (S_s@1), scores rise dramatically to $72-84\%$.

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
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

Table 2: 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}."

On the full image retrieval task, which involves understanding subjects, verbs, and objects jointly, models perform poorly under standard metrics. The performance increases using semantic-similarity success, however, the task is still very challenging. S@1 falls below 10% across the board in Table 3. When using semantic-similarity success (Ss@1), performance improves by roughly 12 percentage points. Still, absolute scores remain low—Ss@1 hovers around 20%—indicating substantial difficulty with fine-grained role comprehension. Among the models, SigLIP2 performs best overall, particularly outperforming CLIP, while FLAVA shows modest gains when evaluated semantically.

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
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

Table 3: Success@K and semantic-similarity success@K evaluation with threshold 0.9 of image retrieval on SVO-Probes dataset.

A breakdown of retrieval performance by role reveals 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 4, the percentage of retrieved examples with the same subject than in query is highest across models, followed by verbs and then by objects. This pattern contrasts with image-text matching, where object negatives were the easiest to detect. Among the models, FLAVA performs worst on subject and verb retrieval, while its object retrieval is on par with BLIP-2.

	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%
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%
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%

Table 4: 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.

Text retrieval proves even more difficult than image retrieval. This is likely due to the fine-grained nature of the captions in SVO-Probes, which often differ only in single word. SigLIP2 with BLIP-2 model outperform the rest in Table 5. Further analysis of retrieved captions confirms the trend holds for text retrieval as well: subjects are consistently easier to retrieve than objects. CLIP performs worst across all roles in Table 6 —suggesting broader difficulty in grounding fine-grained linguistic elements in visual input.

4.2 VALSE: Probing on Counting Ability

VALSE-Counting The VALSE-Counting dataset (Parcalabescu et al., 2021) focuses on numerical reasoning by testing whether models can correctly count objects in images. Each image is paired with a caption specifying a quantity (e.g., "There are exactly 3 giraffes"). This task assesses both object

	1	5	10	20
success@				
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
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

Table 5: Success@K and semantic-similarity success@K evaluation with threshold 0.9 of text retrieval on SVO-Probes dataset.

	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%
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%
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%

Table 6: 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.

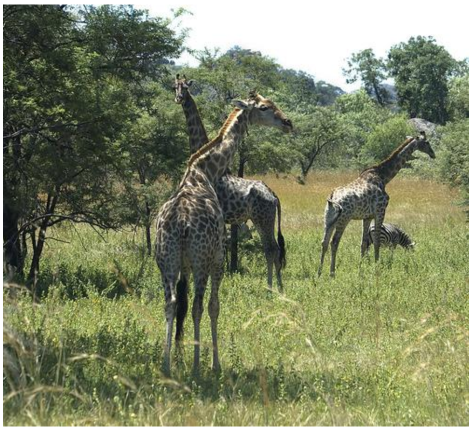


Figure 2: An example from the VALSE-Counting dataset; an image labeled with the positive caption: ”There are exactly 3 of the animals giraffes.”; and with a negative caption: ”There are exactly 14 of the animals giraffes.”. Other semantically correct captions in the dataset that would be evaluated as incorrect in the standard retrieval setup include e.g.: ”There is exactly 1 zebra.”

Model	Overall
CLIP	61.90%
BLIP-2	63.80%
FLAVA	63.80%
SigLIP2	63.10%

Table 7: Image-Text Matching on VALSE dataset.

recognition and quantitative understanding. An example is shown in Figure 2. VALSE has been widely used to benchmark image-text matching.

Compared to verb understanding, counting poses a significantly greater challenge. Image-text matching accuracy on counting is consistently lower than on SVO-Probes, with all models scoring around 62–64% accuracy in Table 7. As image-text matching is a binary classification, these results suggest that VLMs struggle with numerical reasoning more than verb comprehension.

Models perform poorly on image retrieval and even poorer on text retrieval for counting understanding. Image retrieval results in Table 8 show S@1 around 10-14%. Surprisingly, it is slightly better than understanding in SVO-Probes. Unlike SVO-Probes, semantic similarity success is not meaningful for counting. The task requires precise numerical accuracy, not conceptual overlap. This limits the usefulness of approximate matching methods like $S_s@K$ for evaluating counting performance.

Model	S@1	S@5	S@10	S@20
CLIP	11.02	30.15	42.86	55.81
BLIP-2	10.29	29.42	40.80	54.24
FLAVA	11.14	34.02	47.22	62.83
SigLIP2	14.65	35.35	47.58	61.74

Table 8: Success@K for image retrieval on VALSE-counting.

Model	S@1	S@5	S@10	S@20
CLIP	6.20	20.20	28.50	38.70
BLIP-2	9.30	24.20	35.50	46.50
FLAVA	6.30	19.50	31.00	42.00
SigLIP2	9.20	23.80	34.40	49.40

Table 9: Success@K for text retrieval on VALSE-counting.

4.3 Human and GPT Evaluation

Human evaluation suggest that text retrieval performance is better than success@K showed; however, the subject-verb-object understanding and counting comprehension are still very challenging tasks. To assess how well retrieval results align with human expectations, we randomly sampled 100 queries from both datasets. For each image query, the top-10 retrieved items were annotated by three human raters, who judged whether each caption was a valid match. Final labels were determined via majority vote. The results for both datasets are shown in Table 12. A significant increase (for SVO-Probes in some cases, a 60% point difference, for VALSE-counting a $20 - 35\%$) in performance can be seen for all models once a human annotator evaluated them. SigLIP2 is the best-performing model for subject-verb-object understanding, with $\mathrm{S}_h@1$ = 79.79%. while FLAVA is the best for counting with $\mathrm{S}_h@1$ = 49%. The Fleiss' Kappa between three human annotators was 0.641 for SVO-Probes and for VALSE-counting it was 0.743, which suggest that counting is an easier task for humans and they agree more on the answer.

GPT can only approximate the human judgment process for subject-verb-object understanding, as the counting task also represents a substantial challenge for GPT model. It is important to note that some examples could not be evaluated by GPT as the model assumed they where violating the rules of use. We report the performance in Table 12 only for examples that could also be evaluated by GPT. For each query, we used prompt that can be seen in Appendix A, to evalute the models. The Table 12 shows that GPT-4o slightly underestimates the performance of models on SVO-Probes, while it strongly overestimates it for VALSE-counting. We then measured GPT-4o's agreement with human annotations using the F1 score with majority human vote as ground truth. The F1 score for SVO-Probes is 0.73, however, the F1 score for VALSE-counting is 0.45, suggesting that for GPT the verb understanding task is easier than counting.

An example of retrieved examples and their human and GPT evaluation can be seen in the Appendix in F. The evaluations were also carried out for 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			VALSE-counting		
		K=1	K=5	K=10	K=1	K=5	K=10
CLIP	S@K	8.70	19.57	29.35	4.00	12.00	23.00
	S_h@K	68.48	96.74	100.00	33.00	65.00	82.00
	S_g@K	48.91	92.39	97.83	67.00	92.00	96.00
	P@K	8.70	3.91	2.93	4.00	2.60	2.50
	P_h@K	68.48	61.74	57.50	33.00	20.60	19.90
	P_g@K	48.91	46.52	43.48	67.00	47.00	38.00
BLIP-2	S@K	15.62	41.67	52.12	10.00	24.00	31.00
	S_h@K	77.08	94.79	96.88	31.00	71.00	80.00
	S_g@K	69.79	92.71	95.83	63.00	87.00	98.00
	P@K	15.62	8.33	5.42	10.00	4.80	3.20
	P_h@K	77.08	68.33	62.29	31.00	24.60	19.20
	P_g@K	69.79	53.75	46.15	63.00	49.00	41.80
FLAVA	S@K	6.25	22.92	31.25	1.00	16.00	33.00
	S_h@K	61.46	90.62	95.83	49.00	80.00	88.00
	S_g@K	58.33	92.71	100.00	83.00	97.00	99.00
	P@K	6.25	4.58	3.12	1.00	3.40	4.00
	P_h@K	61.46	53.33	48.02	49.00	30.60	27.80
	P_g@K	58.33	45.21	38.02	83.00	58.00	50.30
SigLIP2	S@K	18.09	43.62	58.51	9.00	20.00	30.00
	S_h@K	79.79	96.81	96.81	36.00	65.00	83.00
	S_g@K	75.53	95.74	98.94	71.00	89.00	96.00
	P@K	18.09	8.72	5.85	9.00	4.00	3.20
	P_h@K	79.79	70.43	66.91	36.00	20.80	19.20
	P_g@K	75.53	58.51	49.89	71.00	45.00	38.30

Table 10: Results of text retrieval on a subset of 100 samples from SVO-Probes and VALSE-count. Standard success (S@K), success with human evaluation (S_h@K) and success with GPT4-o evaluation (S_g@K). The F1 score between human majority vote and GPT evaluation is 0.7252 for SVO-Probes and 0.4495 for VALSE-counting. Cohen's kappa between human majority vote and GPT evaluation is 0.4543 for SVO-Probes and 0.2302 for VALSE-counting.

4.3.1 Post-Retrieval Error Analysis of SVO-Probes

To better understand the types of errors made by VLMs, we perform a qualitative analysis of the top-K retrievals on the SVO-Probes dataset. Given the moderate agreement between human raters and GPT-4o $F1 = 0.73$, we use GPT-4o 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 details in the caption do not align with the image) mismatch.

We analyze the top 10 retrieved captions for 100 randomly sampled image queries across the models. For each retrieval, GPT-4o identifies the type of error when a mismatch occurs. Our findings in Appendix C reveal consistent patterns: Subject mismatches are the least frequent across all models, suggesting that models grasp subject identity more reliably than actions or objects. Object errors are the most frequent among incorrect retrievals, particularly for CLIP and in text retrieval for BLIP-2 and SigLIP2. Verb errors are also common, especially for FLAVA. This aligns with our earlier quantitative results and confirms that strong image-text matching scores do not guarantee robust role comprehension. The ability to retrieve topically 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.

To further test this, we treat the top-K retrievals themselves as a comprehensive contrastive dataset and re-run ITM on these examples using BLIP-2 and FLAVA—both of which have dedicated ITM heads. We evaluate these matches using human annotations as ground truth (in Appendix D). While BLIP-2 and FLAVA previously achieved over 80% ITM accuracy on the full SVO-Probes dataset, their performance drops to 50% on the retrieved examples. Crucially, these 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 that standard ITM evaluation can mask important model failures, particularly in subject-verb-object grounding, and highlights the diagnostic value of post-retrieval probing.

5 Conclusion

In this work, we introduced a probing framework based on post-retrieval analysis to better 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 across two targeted datasets, SVO-Probes and VALSE-Counting, revealed significant gaps in model performance that are obscured by image-text matching scores. While all models achieved over 80% accuracy in image-text matching tasks, their success@1 dropped below 10% on both datasets. Semantic similarity metrics partially recovered this performance, suggesting that many "errors" are due to incomplete annotation rather than model failure.

To address annotation incompleteness, we also evaluted retrieval with human judgment and GPT-4o scoring—which better capture the range of valid retrievals. GPT-4o achieved moderate agreement with humans ($F1 = 0.73$) for SVO-Probes, demonstrating potential as a scalable evaluation proxy. However, for VALSE-counting the F1 score was 0.45, indicating counting task is a substantial challenge for the GPT model and can not be used as evaluation proxy.

By subjecting top-K retrievals to further analysis, we uncovered systematic model weaknesses in role understanding and numeric reasoning. These insights highlight the need for deeper, retrieval-based probing methods to build more robust multimodal AI systems. Analysis also showed that models consistently retrieved subjects more accurately than verbs or objects, contradicting image-text matching patterns where object negatives were easiest for models to classify. On counting tasks, all models struggled.

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 two datasets—SVO-Probes and VALSE-Counting—that are synthetic and constrained in structure. While they target key linguistic phenomena, they 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 GPT-4o—they each have their trade-offs. Semantic similarity depends on thresholding and fails in tasks like counting; GPT-4o, while scalable, does not always align perfectly with human judgment.

Additionally, we evaluated only pretrained models, with architecture suitable for encoding of captions and images, and therefore capable of doing retrieval task using cosine similarity as similarity score. Finally, all experiments were conducted in English, limiting the generalizability of findings to multilingual settings.

We leave these avenues for future work.

Computational Resources

For inference and evaluation, our experiments were run on NVIDIA GeForce RTX 3060 Laptop GPU (6GB VRAM). Total compute was approximately 50 GPU hours (mostly from running retrieval across two datasets, multiple times per model).

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.

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.

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.

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 pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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.

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 GPT-4o for Retrieval Evaluation

To scale up retrieval evaluation, we use GPT-4o 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 GPT-4o with the query and each retrieved item, asking it to classify them as correct or incorrect. For incorrect matches, GPT-4o 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 11: Prompt for GPT-4o.

B Image retrieval results

This section presents success@K results for image-to-text retrieval on both SVO-Probes and VALSE-counting, using standard evaluation, human annotation, and GPT-4o assessment. The results are based on 100 randomly sampled image queries for each dataset.

As shown in Table 12, human evaluation consistently yields much higher success@K than standard metrics, revealing the impact of incomplete dataset annotation. GPT-4o estimates closely align with human scores in SVO-Probes but deviate significantly in counting tasks.

We also report precision@K for completeness.

		SVO-Probes			VALSE-counting		
		K=1	K=5	K=10	K=1	K=5	K=10
CLIP	S@K	10.87	26.09	35.87	9.20	25.29	35.63
	S_h@K	64.13	88.04	96.74	24.14	67.82	78.16
	S_g@K	65.22	96.74	100.00	52.87	72.41	79.31
	P@K	10.87	5.43	4.13	9.20	5.06	3.79
	P_h@K	64.13	57.83	51.74	24.14	25.98	22.64
	P_g@K	65.22	55.65	46.96	52.87	33.33	20.80
BLIP-2	S@K	11.24	30.34	39.33	9.30	27.91	38.37
	S_h@K	74.16	94.38	95.51	25.58	60.47	74.42
	S_g@K	69.66	97.75	98.88	52.33	70.93	72.09
	P@K	11.24	6.07	4.04	9.30	5.58	4.07
	P_h@K	74.16	67.64	59.89	25.58	20.23	17.91
	P_g@K	69.66	57.98	48.20	52.33	30.70	18.14
FLAVA	S@K	8.79	25.27	35.16	10.47	29.07	37.21
	S_h@K	72.53	94.51	94.51	29.07	69.77	81.40
	S_g@K	67.03	94.41	98.80	54.65	74.42	80.23
	P@K	8.79	5.05	3.52	10.47	6.05	4.07
	P_h@K	72.53	60.88	56.92	29.07	24.88	22.21
	P_g@K	67.03	54.07	47.14	54.65	37.44	21.98
SigLIP2	S@K	13.40	30.93	42.27	11.63	30.23	44.19
	S_h@K	74.23	93.81	95.88	26.74	52.33	61.63
	S_g@K	79.38	91.75	95.88	59.30	74.42	77.91
	P@K	13.40	6.19	4.43	11.63	6.05	4.77
	P_h@K	74.23	68.45	62.16	26.74	19.07	14.53
	P_g@K	79.38	63.71	57.63	59.30	36.05	19.53

Table 12: Results of image retrieval on a subset of 100 samples from SVO-Probes and VALSE-count. Standard success (S@K), success with human evaluation (S_h@K) and success with GPT4-o evaluation (S_g@K). The F1 score between human annotator and GPT evaluation for SVO-Probes is 0.7628 and for VALSE-counting it is 0.4056.

C Post-retrieval analysis on SVO-Probes dataset

Using GPT-4o, we categorize errors in top 10 text and image retrievals from SVO-Probes into three types: subject, verb, and object incorrect. Table 13 shows the percentage of retrievals falling into each category across models.

This breakdown offers deeper insight into model behavior. Across all models, object mismatches dominate the errors, followed by verbs, with subjects being the most reliably retrieved. This confirms that even when models retrieve semantically close items, fine-grained role comprehension remains a challenge.

D Image-Text Matching with BLIP2 and FLAVA Using Retrieved Samples

To evaluate how models perform on their own retrieval outputs, we apply the image-text matching (ITM) heads of BLIP-2 and FLAVA to the top-K retrieved items. Human annotations are used to judge correctness.

Table 16 shows that while both models score over 80% ITM accuracy on the full dataset, performance drops to 50% on retrieved examples. Since these contain both true positives and hard false positives, this result underscores that ITM accuracy

		Image retrieval	Text retrieval
CLIP	Correct	46.96%	43.48%
	Subject incorrect	7.5%	15.98%
	Verb incorrect	18.59%	19.35%
	Object incorrect	**19.02%**	**21.20%**
BLIP2	Correct	48.20%	46.05%
	Subject incorrect	6.40%	11.04%
	Verb incorrect	**20.34%**	21.36%
	Object incorrect	15.17%	**21.55%**
FLAVA	Correct	48.82%	38.02%
	Subject incorrect	7.42%	17.60%
	Verb incorrect	**22.80%**	**24.69%**
	Object incorrect	11.18%	19.58%
SigLIP2	Correct	57.63%	49.89%
	Subject incorrect	10.93%	15.64%
	Verb incorrect	**19.28%**	16.38%
	Object incorrect	11.96%	**18.09%**

Table 13: Analysis of correct and incorrect samples, retrieved by different models. The classification was done using GPT-4o model.

alone does not reflect deeper linguistic understanding.

	Image retrieval	Text retrieval
BLIP2	59.70%	49.60%
FLAVA	56.50%	52.90%

Table 14: Image-text matching accuracy on examples retrieved by specific model.

E 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 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

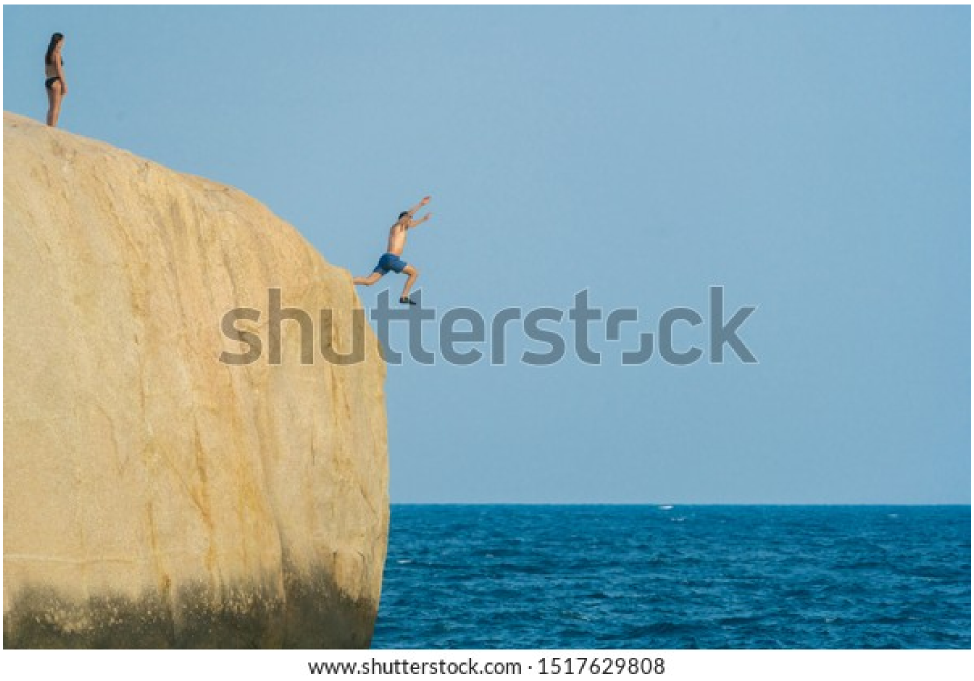
caption as an individual task, multiple captions can be relevant for one image), click the SUBMIT button at the bottom of the page, other.

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

F Examples of Human and GPT evaluation of retrieval

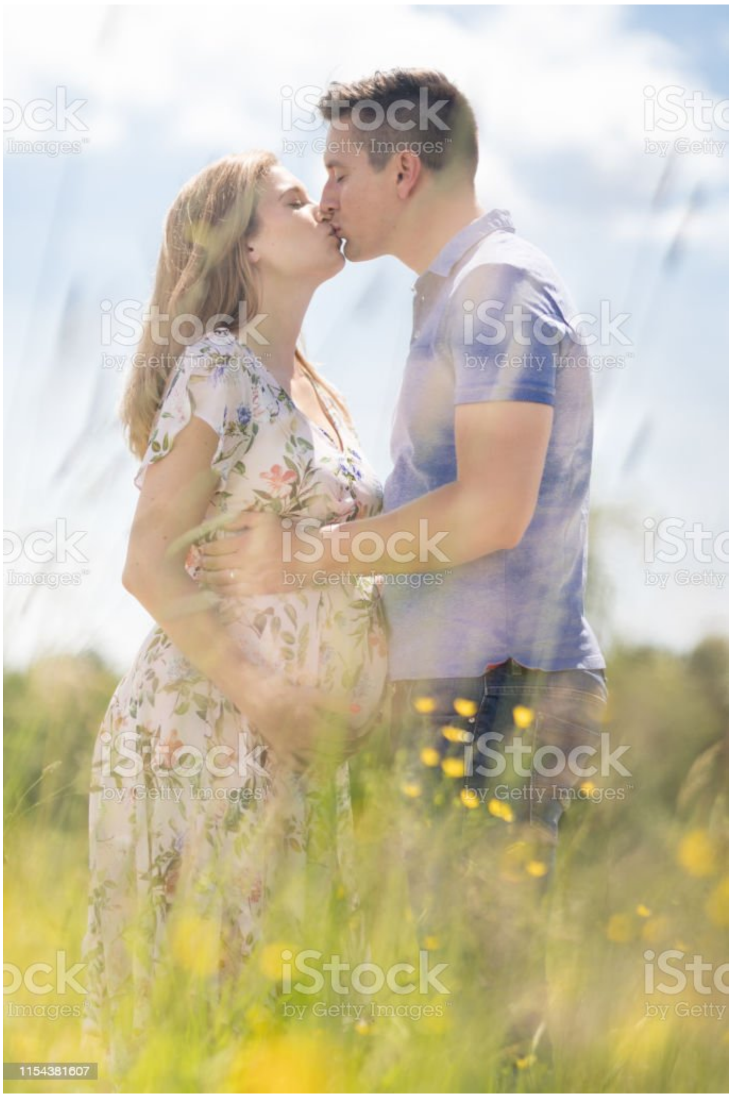
We provide examples of top-10 caption retrievals from the SVO-Probes dataset, along with correctness labels from human annotators and GPT-4o. These examples illustrate the kinds of captions retrieved and how human and LLM judgments agree or differ.

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 GPT-4o's general alignment with human intuition in SVO-Probes, and also showcase challenging cases.



	CLIP			BLIP-2			FLAVA		
Order	Caption	Human	GPT	Caption	Human	GPT	Caption	Human	GPT
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 15: 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 GPT-4o annotations for specific captions, whether or not they match the image.



	CLIP			BLIP-2			FLAVA		
Order	Caption	Human	GPT	Caption	Human	GPT	Caption	Human	GPT
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 16: 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 GPT-4o annotations for specific captions, whether or not they match the image.