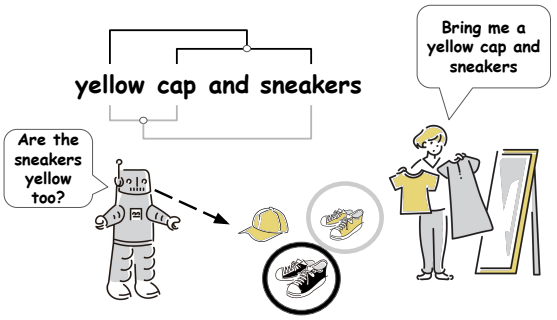
ViLStrUB: A Benchmark for Vision and Language Alignment under Structural Ambiguity

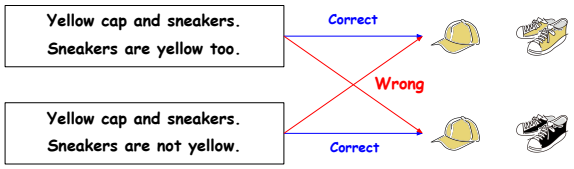
Anonymous ACL submission

Abstract

Structural ambiguity arises when a single sentence admits multiple valid interpretations due to its syntactic structure, posing a fundamental challenge for language understanding. While visual scenes can provide useful cues for resolving such ambiguity, this requires Vision and Language Models (VLMs) to reliably align each possible interpretation with the corresponding visual scene. We introduce **Vi**sion and **L**anguage **Str**uctural **U**nderstanding **B**enchmark (ViLStrUB), a benchmark designed to evaluate vision and language alignment under structural ambiguity, consisting of ambiguous captions, their disambiguated interpretations, and corresponding images across seven ambiguity categories. Using classification-based evaluation settings, we assess a diverse set of contrastive and LLM-based generative VLMs and compare their performance. Our results show that most models perform near chance level and exhibit large gaps from human performance, revealing persistent limitations in aligning structurally distinct interpretations with visual scenes.



(a) The system grounds an ambiguous instruction in visual scene and asks for clarification when needed.



(b) Successful disambiguation requires correctly aligning each interpretation with its visual scene.

Figure 1: Use case of a task-oriented dialogue system capable of disambiguation with visual scenes.

1 Introduction

Structural ambiguity arises when a sentence supports multiple interpretations due to its syntactic structure. Even in a simple phrase such as "yellow cap and sneakers" in Figure 1, the situation referred to by the phrase differs depending on whether yellow modifies only the cap or both the cap and sneakers. Resolving such structural ambiguity is crucial for task-oriented dialogue systems operating in the real world, as user instructions must be accurately interpreted by the systems for their efficient assisting (Tellex et al., 2011, 2014; Shridhar et al., 2020; Bodonhelyi et al., 2024).

Structural ambiguity can sometimes be reduced by introspection (Chomsky, 1965), rewriting an utterance into a less ambiguous form (Duan et al., 2016; Stengel-Eskin et al., 2023). However, in actual dialogue and spoken interaction in the real world, ambiguities remain in speech which can be resolved using additional contextual information, such as dialogue history, prosody, or visual scene (DeVault and Stone, 2009; Widiaputri et al., 2023; Kuribayashi and Baldwin, 2025). Among these, visual scene is frequently regarded as a particularly important cue in real-world interaction (Barnard and Johnson, 2005; Roy, 2005; Roy and Reiter, 2005; Reiter et al., 2005; Hutmacher, 2019). In the example in Figure 1, if the scene contains only the yellow sneakers, no disambiguation is required. In contrast, when the scene also includes sneakers of other colours, the phrase "yellow cap and sneakers" remains structurally ambiguous, and clarification becomes necessary to interpret the instruction correctly. In other words, dialogue system components that perform scene understanding, such as Vision and Language Models (VLMs),

must possess the capability to reason over such ambiguity and leverage visual evidence to support appropriate actions.

A large body of prior work on the capabilities of VLMs has primarily focused on compositionality. These studies have evaluated whether models can correctly ground sentences whose meanings differ due to changes in word order (Thrush et al., 2022; Yuksekgonul et al., 2022; Yamada et al., 2023; Chung et al., 2025). Various types of ambiguity under multimodal settings have also been discussed (Berzak et al., 2015; Mehrabi et al., 2023; Stengel-Eskin et al., 2023; Kuribayashi and Baldwin, 2025; Wang et al., 2025; Chung et al., 2024; Inadumi et al., 2025). However, much of this prior work does not systematically address the combination of candidate interpretations induced by syntactic structure and real-world situations on a sufficient scale. A large-scale and comprehensive benchmark is required to assess whether existing VLMs can exhibit such fine-grained understanding, because small differences in sentence structure can correspond to substantially different real-world events.

To address these issues, we propose the **Vi**sion and **L**anguage **Str**uctural **U**nderstanding **B**enchmark (ViLStrUB). The dataset consists of structurally ambiguous captions paired with their disambiguated interpretations and corresponding images. For instance, in Figure 1, "yellow cap and sneakers" serves as an ambiguous caption. By associating it with interpretive sentences such as "sneakers are yellow too" or "sneakers are not yellow," the ambiguity in the original caption is explicitly resolved. Furthermore, for each disambiguated interpretation, we provide corresponding images reflecting the resolved ambiguity. Based on this dataset, we define both Image-to-Text and Text-to-Image tasks as a novel VLM benchmark. We formulate the problem as selecting the correct alignment between an image and a disambiguated interpretation. Human evaluation of the constructed benchmark confirms that sufficiently trained human annotators can solve these tasks with high accuracy.

Using this benchmark, we conduct a comprehensive evaluation of a wide range of VLMs, from contrastive models represented as CLIP (Radford et al., 2021) to LLM-based generative models (OpenAI, 2025; Liu et al., 2024; Bai et al., 2025; Gemma Team, 2025). Our results reveal that current VLMs exhibit substantial limitations in correctly distinguishing and aligning sentences with subtle semantic differences that originate from the same structurally ambiguous input.

2 Related Work

2.1 Vision and Language Alignment in VLMs

VLMs have been shown to struggle with aligning linguistic structure to visual scenes, particularly in settings that require sensitivity to fine-grained semantic composition. For example, CLIP (Radford et al., 2021) often fails to correctly associate modifiers such as adjectives with their intended target nouns (Tang et al., 2023). Prior work has proposed benchmarks that probe compositional understanding by altering word order to induce meaning changes (Thrush et al., 2022; Yuksekgonul et al., 2022). Our work extends this line of research by focusing on structural ambiguity, where multiple syntactic interpretations arise from the same surface form and remain simultaneously valid prior to grounding. Unlike compositionality benchmarks that assume a single intended meaning, resolving structural ambiguity requires models to distinguish and align closely related interpretations with corresponding visual scenes.

2.2 Visual Disambiguation

Prior work has explored the use of visual scenes to resolve linguistic ambiguity, both through evaluation benchmarks (Chung et al., 2024; Wang et al., 2025) and task-oriented systems (Inadumi et al., 2025; Kuribayashi and Baldwin, 2025). These benchmarks partially address structural ambiguity. However, their goal is not a systematic investigation of the capabilities of VLMs; rather, it is to realise specific application tasks.

The Language and Visual Ambiguity (LAVA) corpus is one of the few datasets explicitly designed to address structural ambiguity, using handcrafted visual annotations (Berzak et al., 2015). While LAVA has served as a foundational resource for subsequent studies, its limited scale and annotation quality constrain its applicability for evaluating modern VLMs (Mehrabi et al., 2023; Yamaki et al., 2023). The Text-to-Image Ambiguity Benchmark (TAB) extends LAVA by improving the quality and coverage of textual annotations and targeting structural disambiguation in text-to-image generation (Mehrabi et al., 2023). Nevertheless, because TAB is centred on disambiguating prompts for image generation models, the dataset lacks visual scenes, which is not suited for assess-

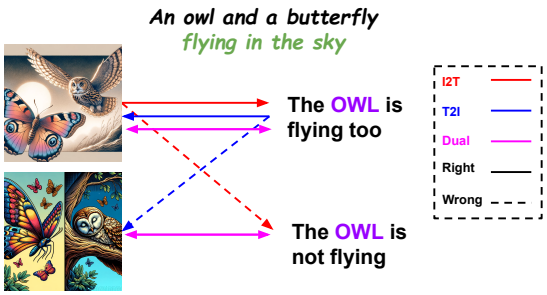


Figure 2: Overall description for classification task

ing whether models can reliably interpret visual scenes. Building on recent advances in large-scale image and language generation (Betker et al., 2023; OpenAI, 2023), we introduce a benchmark that provides paired visual scenes and disambiguated interpretations.

3 ViLStrUB: Vision and Language Structural Understanding Benchmark

We propose the **Vi**sion and **L**anguage **Str**uctural **U**nderstanding **B**enchmark (ViLStrUB) to measure how well VLMs can handle structural ambiguity. We first describe the formulation of the alignment tasks that VLMs should be able to solve under structural ambiguity, and clarify the range of structural ambiguity phenomena covered by our benchmark. We then present the data construction pipeline and report a human evaluation of the quality of the resulting benchmark dataset.

3.1 Task Definition

ViLStrUB formulates vision and language alignment evaluation as a classification task, in which models are required to match structurally distinct interpretations of an identical sentence with their corresponding visual scenes. As shown in Figure 2, each sample contains an ambiguous sentence ("an owl and butterfly flying in the sky"), two sentences that disambiguate the structural ambiguity ("the owl is flying too" and "the owl is not flying"), and two images corresponding to both interpretations.

We subdivide the ViLStrUB task into three settings: Image-to-Text (I2T), Text-to-Image (T2I), and Dual, based on the input–output configuration of each trial. Each trial is associated with a single ambiguous sentence and its corresponding set of structurally disambiguated interpretations. Given either an image or a caption representing one interpretation, the model is required to select the matching caption or image from a set of candidates derived from the same ambiguous source. An overview of the task setup is illustrated in Figure 2.

Since prior work has shown that performance can differ by direction in vision and language alignment tasks (Thrush et al., 2022), we define three task settings: I2T, T2I, and Dual. In the I2T setting, a single image and multiple candidate captions (two or three captions in our case) are provided, and the model selects one caption that best matches the image. The T2I setting is the inverse: given one caption and multiple candidate images (two or three images in our case), the model selects one image that best matches the caption. Here, the caption is provided as the concatenation of two sentences: the original ambiguous sentence and its disambiguated interpretation. The Dual setting counts an instance as correct only when it is solved correctly in both directions.

3.2 Ambiguity Categories

We categorise instances in ViLStrUB to enable a diversified evaluation of structural interpretation in VLMs. Our categorisation is grounded in established typologies of structural ambiguity, from which we select ambiguity categories that are both linguistically distinct and plausibly resolvable using a visual scene. We build ViLStrUB upon the text samples introduced in TAB (Mehrabi et al., 2023), which was originally designed for text-to-image generation and consist of ambiguous sentences. Among the ambiguity categories defined in TAB, five correspond to linguistic phenomena. We further refine and subdivide these into seven categories to better isolate different mechanisms of structural interpretation[1]. Figure 3 illustrates our ambiguity categories defined below.

- Verb Phrase Attachment **(VP)**: Ambiguity arises when a verb phrase can attach to more than one part of the sentence.
- Preposition Phrase Attachment **(PP)**: A prepositional phrase can modify multiple possible heads
- Anaphora **(Anaph)**: A pronoun or referring expression has more than one plausible antecedent.
- Ellipsis **(Ellip)**: An omitted phrase can be interpreted in multiple ways.

[1]Detailed redefinition we used for the data construction are stated in Appendix A.1

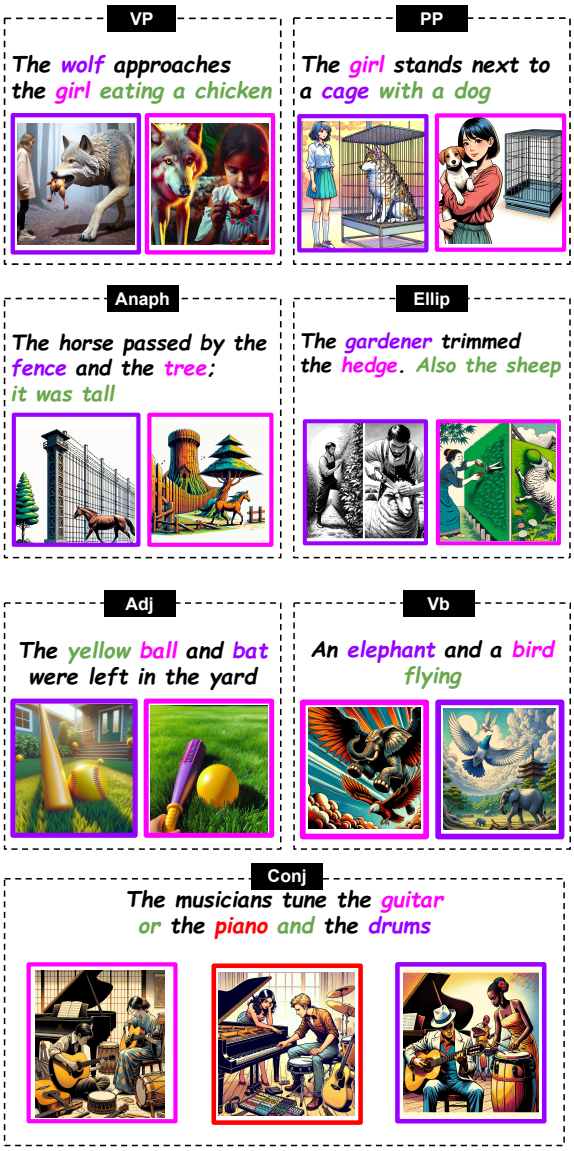


Figure 3: Example sentence and corresponding interpretations with visual scenes from each ambiguity category.

- Adjective Scope **(Adj)**: An adjective can modify either a single noun or an entire coordinated noun phrase.
- Verb Scope **(Vb)**: A verb-derived modifier may apply to one or more coordinated elements.
- Conjunction Scope **(Conj)**: Coordinating conjunctions (e.g. and, or) group sentence elements in more than one way.

3.3 Data Collection

We construct ViLStrUB through a multi-stage data collection pipeline that leverages a VLM (OpenAI, 2023) and an image generation model (Betker et al., 2023) for both caption augmentation and image generation. Prior to data collection, we apply a filtering and modification step to the sentences from TAB (Mehrabi et al., 2023), which we use for seeds for caption augmentation. Specifically, sentences containing violent expressions or references to real world political figures are discarded or minimally modified, as such content is rejected by the image generation model[2].

Ambiguity Category	I2T	T2I	Dual
VP	96.5	98.5	96.0
PP	92.5	97.0	91.5
Anaph	94.5	88.1	85.1
Ellip	89.6	92.6	84.7
Adj	89.5	92.0	85.5
Vb	94.5	92.0	91.5
Conj	93.3	93.0	91.3
All	92.9	93.1	89.6

Table 1: Human evaluation results on ViLStrUB, reported as accuracy (%) for Image-to-Text (I2T), Text-to-Image (T2I), and Dual settings across ambiguity categories.

3.3.1 Caption Augmentation

Building on the selected samples described in Section 3.2, we generate new ambiguous-disambiguated text pairs using GPT-4o[3] (OpenAI, 2023). Each ambiguous sentence is paired with two or three disambiguated counterparts following the format introduced in TAB. The resulting dataset includes 700 ambiguous sentences, with 100 sentences per ambiguity category. Each sentence is paired with two or three disambiguated interpretations, yielding a total of 1,503 disambiguated captions. For example, **Conj** sentences always have three interpretations.

3.3.2 Image Generation

For each disambiguated caption, we generate a corresponding image using DALL-E 3[4] (Betker et al., 2023). To introduce visual diversity, approximately half of the images are generated in a cartoon-style, and the remainder in a photo-realistic-style[5].

[2]Examples of these cases and the corresponding modifications are described in Appendix A.2.

[3]`gpt-4o-mini-2024-07-18`

[4]We used the API from December 11, 2024, to April 5, 2025.

[5]The prompts used for image generation are provided in Appendix B.1.

3.3.3 Human Evaluation

To verify that the collected data are resolvable in principle given visual scenes, we conduct a human evaluation using the task described in Section 3.1. Two annotators not involved in data collection participated in the evaluation, each assigned either T2I or I2T setting for a given ambiguity category. Dual accuracy is computed from the two results. To mitigate memory effects, no annotator evaluates both T2I and I2T settings for the same category[6]. As shown in Table 1, human annotators achieve high performance in both T2I and I2T, while strong Dual performance indicates minimal modality asymmetry, supporting the validity of the dataset.

3.3.4 Image Description Generation

After human evaluation, we generate auxiliary image descriptions focusing on the attributes of main objects using GPT-5.1 (OpenAI, 2025). These descriptions are not used for the main evaluation, but are introduced to conduct later analyses of whether models rely on structurally relevant visual cues or superficial details. We leverage these descriptions in our analysis[7].

4 Experiment

Section 3 introduced a benchmark to evaluate how well existing VLMs can interpret structural ambiguity. In this section, we describe our experimental settings and results.

4.1 Evaluated Models

We test the classification task described in Section 3.1 on representative VLMs, which can be broadly categorised into two groups: contrastive models centred on CLIP (Radford et al., 2021), and LLM-based generative VLMs. Contrastive models are directly relevant to our research goal, as their training objective explicitly optimises cross-modal alignment in a shared embedding space (Chen et al., 2020; Khosla et al., 2020), making them a natural testbed for assessing structural alignment capabilities. Moreover, encoders trained under the CLIP paradigm are widely adopted as backbone components in state-of-the-art generative VLMs (Betker et al., 2023; Gemma Team, 2025; Li et al., 2024; DeepSeek-AI, 2024), meaning that limitations observed at the contrastive level may propagate to downstream models. In contrast, LLM-based generative VLMs represent the dominant paradigm for contemporary multimodal systems. Evaluating both paradigms allows us to assess whether structural alignment failures stem from embedding-level representations or can be mitigated by later-stage multimodal reasoning.

4.1.1 Contrastive VLMs

We evaluate CLIP (Radford et al., 2021) and its variants, covering differences in (i) different vision encoders (Resnet 50, 101 (He et al., 2015) as well as Vision Transformer (Dosovitskiy et al., 2020)), training objectives (SigLIP (Zhai et al., 2023)), training dataset scale (OpenCLIP (Cherti et al., 2023) based on Vision Transformer and ConvNext (Liu et al., 2022), MetaCLIP (Xu et al., 2024), and MetaCLIP2 (Chuang et al., 2025)), and (iv) model scale, exemplified by EVA-CLIP (Sun et al., 2023). Contrastive models are encoder-based models, and their outputs are image and text embeddings. Predictions in both the T2I and I2T settings are obtained by selecting the candidate with the highest cosine similarity to the given query embedding[8].

4.1.2 LLM-based Generative VLMs

LLM-based generative VLMs can produce responses to prompts as outputs. For these models, candidate captions and images are presented with discrete option labels (e.g., A/B/C), and the model is prompted to select the option that best matches the given input. In addition, the models are prompted to generate a brief explanation for their choice, which is used in our analysis[9]. Our evaluation includes both closed-source and open-source models. Among closed-source models, we evaluate GPT5.1 (OpenAI, 2025) to assess the structural comprehension ability of state-of-the-art models. For open-source models, we adopt Qwen-VL-3 (Bai et al., 2025), Llava-Next (Liu et al., 2024), and Gemma3 (Gemma Team, 2025).

4.2 Results

Table 2 reports results grouped by ambiguity category, while Table 3 presents results grouped by image style.

4.2.1 Contrastive VLMs

As shown in Table 2, the performance of contrastive models generally remains close to the ran-

[6]Details of annotator allocation are provided in the Appendix B.2.

[7]Generation prompt in the Appendix B.1.

[8]Model cards we used are provided in the Appendix C.

[9]We provide the prompt in the Appendix D.

	VP			PP			Anaph			Ellip			Adj			Vb			Conj			All		
Model	I2T	T2I	Dual	I2T	T2I	Dual	I2T	T2I	Dual	I2T	T2I	Dual	I2T	T2I	Dual	I2T	T2I	Dual	I2T	T2I	Dual	I2T	T2I	Dual
Contrastive VLMs																								
CLIP-ViT	50.0	51.5	25.0	55.5	50.0	32.0	53.7	49.8	26.9	54.5	49.0	30.7	53.5	52.0	31.0	59.0	53.0	30.5	50.3	37.0	17.0	53.6	48.1	26.9
CLIP-RN50	52.0	48.5	27.5	55.5	51.0	29.5	55.2	51.7	30.9	52.5	53.0	30.7	55.0	50.0	29.5	55.5	51.0	26.0	50.7	36.0	19.7	53.6	47.9	27.2
CLIP-RN101	47.5	50.5	24.5	45.0	49.5	23.0	52.7	48.8	27.9	52.5	50.0	27.7	58.5	50.5	29.5	58.0	51.0	31.5	54.0	35.3	18.7	52.7	47.1	25.6
SigLIP	54.0	49.5	31.0	57.0	50.0	26.5	51.2	50.3	27.9	50.0	50.5	22.3	50.5	50.5	21.5	55.5	50.5	33.0	46.0	38.3	20.3	51.6	47.8	25.7
MetaCLIP	53.0	49.0	26.5	56.0	51.0	30.5	54.2	48.8	25.9	55.9	56.4	30.2	56.5	51.0	33.0	55.0	52.5	31.5	51.7	36.3	19.7	54.4	48.4	27.6
MetaCLIP2	51.0	49.5	27.5	54.5	50.0	26.5	53.2	50.3	29.4	55.9	52.0	30.7	58.5	51.5	31.5	55.0	51.5	33.0	59.0	38.7	22.0	55.6	48.4	28.2
OpenCLIP-ViT	48.5	49.5	23.0	53.0	52.0	26.0	56.7	48.8	28.4	60.4	50.0	31.2	53.0	52.5	25.5	60.0	51.5	33.5	54.0	37.7	19.7	55.0	48.1	26.3
OpenCLIP-convnext	54.0	50.5	30.0	51.5	51.0	26.5	52.2	50.8	29.9	55.5	53.5	33.7	61.5	51.5	30.5	60.0	52.0	31.0	55.7	37.0	21.7	55.8	48.6	28.5
EVA-CLIP	49.0	50.0	28.0	54.5	49.5	27.5	51.7	50.8	31.3	57.9	52.0	29.7	58.5	52.5	31.0	60.5	51.0	36.5	54.0	38.7	20.3	55.1	48.5	28.6
LLM-based Generative VLMs																								
llava-1.6-mistral-7b	51.5	48.5	17.0	53.5	49.5	27.0	49.8	58.2	22.9	51.0	51.0	31.7	50.0	56.5	21.5	53.5	44.5	35.5	40.0	36.0	14.7	49.2	48.3	23.7
Qwen3-VL-8B-Instruct	57.5	65.5	38.5	76.5	77.5	60.0	63.2	63.2	38.3	60.4	61.4	38.6	65.0	87.0	59.0	73.0	76.0	56.5	56.7	61.7	42.3	64.1	69.7	47.2
Gemma3-12b-it	64.0	52.0	29.5	64.5	56.0	42.0	58.2	51.2	19.9	55.9	55.5	31.7	60.0	69.5	53.5	68.0	69.5	57.0	44.0	45.0	15.7	58.2	56.2	34.3
GPT-5.1	86.5	73.0	68.5	91.0	89.0	81.5	82.1	64.2	53.2	65.4	70.8	45.5	82.5	86.0	71.0	84.0	88.0	75.5	63.7	59.3	40.3	78.2	74.7	60.8
Human Evaluation & Random Chance																								
Random Chance	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	50.0	50.0	25.0	33.3	33.3	11.1	—	—	—
Human	96.5	98.5	96.0	92.5	97.0	91.5	94.5	88.1	85.1	89.6	92.6	84.7	89.5	92.0	85.5	94.5	92.0	91.5	93.3	93.0	91.3	92.9	93.1	89.6

Table 2: Model accuracies by ambiguity category. Cell colours indicate deviation from the chance level under a binomial null model: green denotes above-chance performance and red denotes below-chance performance. Colour intensity reflects the magnitude of deviation in units of standard deviations. The bottom rows of the tables report the expected random chance accuracy, which depends on the number of options per trial, as well as human performance from Table 1.

	Cartoon			Photo			All		
Model	I2T	T2I	Dual	I2T	T2I	Dual	I2T	T2I	Dual
Contrastive VLMs									
CLIP-ViT	53.4	48.1	26.9	53.2	46.6	25.6	53.6	48.1	26.9
CLIP-RN50	52.7	47.0	26.9	54.9	49.3	27.5	53.6	47.9	27.2
CLIP-RN101	52.3	46.7	25.0	53.3	47.7	26.7	52.7	47.1	25.6
OpenCLIP-ViT	54.7	47.4	25.2	54.2	46.9	25.3	55.0	48.1	26.3
OpenCLIP-convnext	54.7	47.7	27.8	57.5	50.2	29.8	55.8	48.6	28.5
LLM-based Generative VLMs									
Qwen3-VL-8B-Instruct	67.9	69.1	51.2	57.8	70.7	40.8	64.1	69.7	47.2
GPT-5.1	81.5	75.5	63.6	73.0	73.3	56.1	78.2	74.7	60.8

Table 3: Selective model accuracies by image style

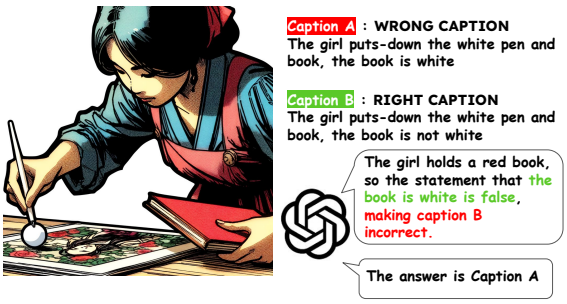


Figure 4: Hallucination case example by GPT-5.1

dom chance level across most settings. An exception is observed in the Conj category, where I2T accuracy is noticeably higher and approaches the random chance level of two-option categories, despite Conj involving three options per trial.

Across all categories, I2T accuracy is consistently higher than T2I accuracy, and Dual accuracy generally hovers around the random chance. This pattern indicates unstable cross-modal alignment, where correct matches in one direction do not reliably coincide with correct matches in the other. Results grouped by image style in Table 3 show similar trends to those observed across ambiguity categories, with no significant performance differences attributable to image style, despite variations in visual encoders.

4.2.2 LLM-based Generative VLMs

In contrast to contrastive models, LLM-based generative VLMs exhibit substantially more varied performance across models (Table 2). LLaVA-NeXT shows the weakest performance, remaining close to random chance across most categories and occasionally falling below it. Gemma-3 performs better overall and exceeds the random baseline in several categories; however, its performance remains limited, particularly in terms of Dual accuracy. Both models exhibit consistently low Dual scores across multiple categories, indicating pronounced asymmetry between I2T and T2I alignment.

Among the open-source models, Qwen3-VL achieves the strongest and most stable performance. GPT-5.1 attains the highest overall accuracy, ex-

Model	I2T	T2I
CLIP-ViT	0.6569	2.3413
CLIP-RN50	0.0051	0.0208
CLIP-RN101	0.0051	0.0201
OpenCLIP-ViT	0.7848	3.8397
OpenCLIP-convnext	0.0075	0.0343

Table 4: Average logit difference between candidate pairs for selective contrastive models, computed per trial. Logits are obtained as the cosine similarity between image and text embeddings, scaled by e^{τ}, where the temperature τ is fixed to 1.0 across all experiments.

	OpenCLIP-convnext		Qwen3-VL	
Type	amb-dis	dis-dis	amb-dis	dis-dis
VP	95.9	99.2	90.4	94.1
PP	96.2	98.0	89.3	94.5
Anaph	98.4	99.3	94.9	97.1
Ellip	94.2	94.7	72.4	69.0
Adj	95.6	98.4	90.8	96.8
Vb	94.8	97.5	90.2	97.4
Conj	95.8	97.1	91.7	97.0

Table 5: Average cosine similarity between caption pairs per sample. *amb-dis* denotes similarity between an ambiguous caption and its disambiguated version, while *dis-dis* denotes similarity among disambiguated candidates derived from the same ambiguous sentence. Results are shown for OpenCLIP-convnext and Qwen3-VL; Qwen3-VL text embeddings are obtained by mean-pooling hidden states from the text backbone (Tang et al., 2015).

ceeding 90% in several settings. These two models consistently outperform the random chance baseline and achieve relatively higher Dual scores, suggesting more stable cross-modal alignment. Nevertheless, their performance remains notably below human performance. Moreover, Qwen3-VL, despite its smaller scale, outperforms GPT-5.1 in certain settings, indicating that model size alone does not guarantee stronger structural alignment.

Results stratified by image style (Table 3) reveal systematic differences for these models: I2T accuracy tends to be higher for cartoon-style images, while T2I accuracy is higher for photo-realistic images. Finally, we observe instances of hallucination in GPT-5.1, where the generated explanation does not align with the model's actual decision (Figure 4). This discrepancy highlights a persistent limitation of current generative VLMs in achieving reliable vision and language structural alignment, posing a challenge for robust visual disambiguation.

5 Analysis

We analyse the performance of VLMs to identify the key challenges in improving their vision and language structural alignment capabilities. Our analysis highlights two primary limitations:

- **Insufficient sensitivity to semantic differences between texts from structural ambiguity**, where models fail to distinguish between alternative syntactic interpretations that are semantically distinct.

- **Dominance on superficial visual features**, which distracts models from visually grounded cues that are directly relevant to resolving structural ambiguity.

5.1 Insufficient Sensitivity to Semantic Differences

The consistent gap between I2T and T2I performance, together with the low Dual accuracy reported in Section 4.2, indicates a modality gap in current VLMs. Compared to visual representations, textual representations exhibit limited diversity, particularly when multiple interpretations share nearly identical surface forms under structural ambiguity.

As shown in Table 4, contrastive models display greater variation across visual candidates than across textual candidates in a general sense. Table 5 reveals that sentence embeddings corresponding to different structural interpretations are highly similar in the text embedding space for OpenCLIP-convnext. Such embeddings collapse makes alternative interpretations difficult to distinguish, which could have caused most performance around the random chance. This collapse could have even reduced the number of distinguishable options, making I2T performance in the Conj category close to the random chance of two option categories despite having three options.

Qwen3-VL shows a similar tendency in text embedding similarity, with the exception of the Ellip category. Nevertheless, its overall performance is substantially higher than that of contrastive models (Table 2), suggesting that some semantic distinctions may be partially recovered during later stages of multimodal fusion. Despite this, performance remains well below human levels, indicating that such recovery is limited.

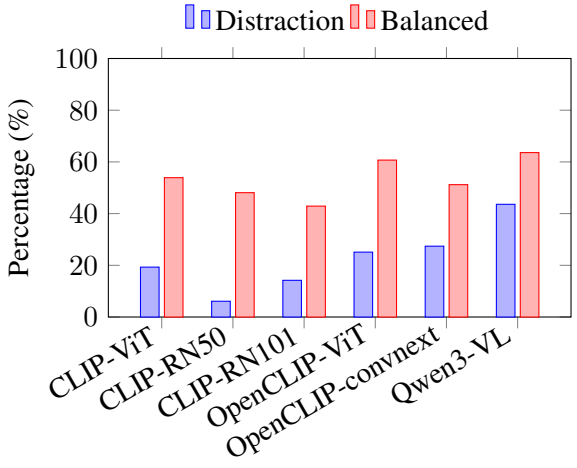


Figure 5: Additional test results of **Distraction** and **Balanced** settings across selected models. **Distraction** is the winning rate of the correct caption against the wrong caption, added with visual description. **Balanced** is the winning rate of the correct caption when both the correct and wrong captions are added with a visual description.

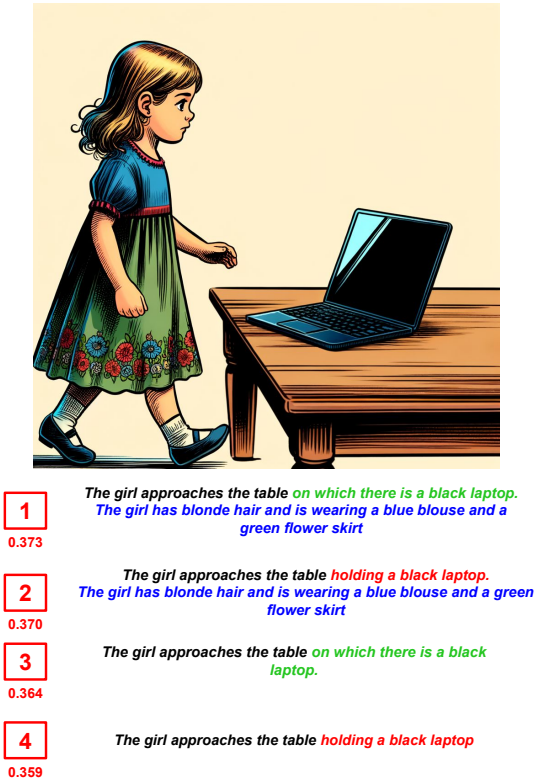


Figure 6: OpenCLIP-ViT error case. Captions are ranked by the image–text matching logit for the given image. Green/red spans indicate correct/incorrect semantics, and the blue sentence is the visual description.

5.2 Dominance of Superficial Visual Features

We conduct an additional I2T evaluation using the visual descriptions introduced in Section 3.3.4. These descriptions capture superficial image features that are largely independent of structural semantics. By augmenting captions with such descriptions, we examine whether VLMs prioritise structural semantics over surface-level visual cues.

In the **Distraction** setting (Figure 5), only the incorrect candidate caption is augmented with a visual description, acting as a distractor that biases the model away from the structurally correct interpretation. We measure the winning rate of the correct caption against this distraction. Results show that the correct caption is selected in fewer than 50% of cases, indicating a strong reliance on superficial visual features. The example in Figure 6 further illustrates that models attend to visually salient yet semantically irrelevant details.

We further evaluate a **Balanced** setting, in which all candidate captions are augmented with visual descriptions. As shown in Figure 5, applying descriptions uniformly improves the winning rate of the correct caption, suggesting that the observed degradation in the Distraction setting cannot be explained solely by increased caption length.

6 Conclusion

We introduced a benchmark for evaluating VLMs on aligning interpretations with subtle semantic differences under structural ambiguity with corresponding visual scenes. Covering a diverse set of ambiguity types, our benchmark enables systematic evaluation across both contrastive and generative model paradigms. Experimental results reveal that current VLMs exhibit limitations in vision and language structural alignment, a fundamental prerequisite for visual disambiguation. Our analysis shows that semantic differences between alternative interpretations are poorly reflected in textual representations, and that models often rely on superficial visual cues rather than structurally relevant semantics. These findings highlight the need for improved cross-modal reasoning and greater sensitivity to structural meaning. Future work should focus on developing models that abstract beyond surface-level features and align syntactic interpretations more reliably with visual scenes.

Limitations

- While our human evaluation of our data suggests its validity in Table 1, more thorough analyses are required regarding the data's statistics. Specifically, diversity in both ambiguous sentences and images would be an important factor justifying our collected dataset.
- While our results suggested that model size isn't yet an important factor for the models' disambiguation ability, further experiments could be done on various sizes from the same model to see more detailed performance differences. Also, more evaluation would be needed on closed models such as Gemini (Google, 2024).
- Our analysis focuses primarily on embedding-level behaviour and does not explicitly examine the role of training objectives in different VLM paradigms. In particular, LLM-based generative VLMs outperform contrastive models despite exhibiting similar text embedding similarities, suggesting that later-stage multimodal fusion plays an important role. A deeper investigation of this aspect is left for future work.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xiong-Hui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, and 46 others. 2025. Qwen3-vl technical report. arXiv:2511.21631.

Kobus Barnard and Matthew Johnson. 2005. Word sense disambiguation with pictures. *Artificial Intelligence*, 167(1):13–30.

Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. 2015. Do you see what I mean? visual resolution of linguistic ambiguities. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1477–1487.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science.*, 2(3):8.

Anna Bodonhelyi, Efe Bozkir, Shuo Yang, Enkelejda Kasneci, and Gjergji Kasneci. 2024. User intent recognition and satisfaction with large language models: A user study with ChatGPT. arXiv:2402.02136.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607.

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

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke S. Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen tau Yih, Shang-Wen Li, and Hu Xu. 2025. Meta CLIP 2: A worldwide scaling recipe. arXiv:2507.22062.

Jiwan Chung, Seungwon Lim, Jaehyun Jeon, Seungbeen Lee, and Youngjae Yu. 2024. Can visual language models resolve textual ambiguity with visual cues? let visual puns tell you! In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2452–2469.

Jiwan Chung, Seungwon Lim, Sangkyu Lee, and Youngjae Yu. 2025. MASS: Overcoming language bias in image-text matching. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence*, 3, pages 2591–2599.

DeepSeek-AI. 2024. Deepseek-v3 technical report. arXiv:2412.19437.

David DeVault and Matthew Stone. 2009. Learning to interpret utterances using dialogue history. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 184–192.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 8th International Conference on Learning Representations*.

Manjuan Duan, Ethan Hill, and Michael White. 2016. Generating disambiguating paraphrases for structurally ambiguous sentences. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016*, pages 160–170.

Gemma Team. 2025. Gemma 3 technical report. arXiv:2503.19786.

Gemini Team Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Fabian Hutmacher. 2019. Why is there so much more research on vision than on any other sensory modality? *Frontiers in Psychology*, 10.

Shun Inadumi, Nobuhiro Ueda, and Koichiro Yoshino. 2025. Disambiguating reference in visually grounded dialogues through joint modeling of textual and multimodal semantic structures. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 11183–11198.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 18661–18673.

Tatsuki Kuribayashi and Timothy Baldwin. 2025. Does vision accelerate hierarchical generalization in neural language learners? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1865–1879.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy visual task transfer. arXiv:2408.03326.

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. https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11966–11976.

Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. 2023. Resolving ambiguities in text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 14367–14388.

OpenAI. 2023. GPT-4 technical report. arXiv:2303.08774.

OpenAI. 2025. GPT-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1):137–169.

Deb Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1):170–205.

Deb K. Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(1):1–12.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10740–10749.

Elias Stengel-Eskin, Jimena Guallar-Blasco, Yi Zhou, and Benjamin Van Durme. 2023. Why did the chicken cross the road? rephrasing and analyzing ambiguous questions in VQA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 10220–10237.

Quan Sun, Yuxin Fang, Ledell Yu Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved training techniques for CLIP at scale. arXiv:2303.15389.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1014–1023.

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

Stefanie Tellex, Ross A. Knepper, Adrian Shuai Li, Daniela Rus, and Nicholas Roy. 2014. Asking for help using inverse semantics. In *Proceedings of the Robotics Science and Systems*.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the 25th AAAI Conference*

on Artificial Intelligence, volume 25, pages 1507–1514.

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. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5228–5238.

Xiaolong Wang, Zhaolu Kang, Wangyuxuan Zhai, Xinyue Lou, Yunghwei Lai, Ziyue Wang, Yawen Wang, Kaiyu Huang, Yile Wang, Peng Li, and Yang Liu. 2025. MUCAR: Benchmarking multilingual cross-modal ambiguity resolution for multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15026–15048.

Ruhiyah Widiaputri, Ayu Purwarianti, Dessi Lestari, Kurniawati Azizah, Dipta Tanaya, and Sakriani Sakti. 2023. Speech recognition and meaning interpretation: Towards disambiguation of structurally ambiguous spoken utterances in Indonesian. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16813–16824.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2024. Demystifying CLIP data. In *Proceedings of the 12th International Conference on Learning Representations*.

Yutaro Yamada, Yingtian Tang, and Ilker Yildirim. 2023. When are lemons purple? the concept association bias of CLIP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14333–14348.

Ryosuke Yamaki, Tadahiro Taniguchi, and Daichi Mochihashi. 2023. Holographic CCG parsing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 262–276.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Y. Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *The 11th International Conference on Learning Representations*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

A Modification from TAB

A.1 Revision of Ambiguity Categories

TAB (Mehrabi et al., 2023) dataset has defined the following seven categories of ambiguity:

- Anaphora: Ambiguity arises when pronouns or similar expressions refer to a previously mentioned entity, but there are multiple possible referents. (e.g. "The girl looks at the bird and the butterfly; it is red.")
- Ellipsis: Ambiguity caused by omitted elements in a sentence, resulting in multiple possible interpretations. (e.g. "The lion eats the chicken. Also the cat.")
- Fairness: Ambiguity occurs when the caption lacks specific attributes of an object, resulting in multiple possible visual interpretations. (e.g. "The man dusting the floor.")
- Syntax-PP: Ambiguity arises when it's unclear which part of the sentence a prepositional phrase is modifying. (e.g. "The woman approached the chair with a bag.")
- Syntax-VP: Ambiguity arises when it's unclear which part of the sentence a verb phrase is modifying (e.g. "The man looked at a boy talking to a telephone.")
- Conjunction: Ambiguity caused by the scope of verbs or adjectives connected to multiple nouns via conjunctions like "and" or "or". (e.g. "The girl holds the green chair and bag.")
- Miscellaneous: A collective set of ambiguities which are not affiliated to either of the above six, but not quantitatively enough to form a separate category. (e.g. "The chicken is ready to eat.")

Among the original ambiguity types, Fairness was found to be exclusively related to image generation and was therefore unsuitable for our research, which focuses on semantic diversity rather than visual representation. Additionally, the Miscellaneous category contained too few instances. As a result, we excluded both of these categories from our study. Furthermore, we redefined the Conjunction category as a scope ambiguity problem and subdivided it into three finer-grained types: adjective scope, verb scope, and conjunction scope.

A.2 Inappropriate Samples from TAB

While TAB was originally designed for image generation tasks, some of its samples included inappropriate content that was rejected by the generation model. One common issue involved violent verbs,

Collecion Process	Prompt
Caption Augmentation	Hi, I'm making a dataset by extending the following examples. Output sentences in the following format: – An ambiguous sentence having 2 or 3 possible meanings: Avoid repeating common phrases and use a wide range of vocabulary and creative expression, a variety of synonyms and idioms. – Disambiguated sentences corresponded to an ambiguous sentence: Do not say something else, but just 2 or 3 sentences. These sentences are connected slash. – If I'm not satisfied, I will give you feedback. If I say good, then generate another round. – Create a text filled with detail that allows one to easily visualise the scene. The topic is `{AMB_TYPE}`. From now on, I will show you some of the examples. **Example:** `{EXAMPLE_FROM_TAB}`
Image Generation	Follow the given caption prompt and visual style to generate a faithful image. **Prompt:** `{CAPTION}` **Style:** `{coloured cartoon OR coloured photography}`
Image Description Generation	You are a vision-language agent that outputs visual attributes ONLY for objects explicitly mentioned in the caption. Your behaviour rules: - Describe ONLY objects that appear both in the caption and the image. - For each object, output at most TWO visual attributes. - Attributes must be concise (e.g., "red," "wooden," "large," "striped shirt"). - Ignore objects not mentioned in the caption. - Ignore actions, relationships, and scene-level descriptions. - Keep it concise and factual. Example style: "the boy is wearing a striped shirt and the dog has brown fur" **Caption:** `{CAPTION}` **Image:** `{IMAGE}` Now look at the image and output the object descriptions.

Table 6: Prompt templates used for data collection.

such as kill, threaten, or hit (e.g., "The girl killed the boy with a gun."). Another issue was the inclusion of real-world political figures from the contemporary era, which also triggered rejection (e.g., "Biden sits next to a girl worshipping Trump.").

To address these issues, we made the following modifications: violent verbs were replaced with neutral alternatives (e.g., "greet"), and named political figures were replaced with descriptive phrases (e.g., "the old man and the blonde man") to preserve the intended ambiguity while avoiding rejection by the model.

B Data Collection

B.1 Prompts used for Generation Models

For data collection, we used the following prompts in Table 6. For caption augmentation, previous samples from TAB were given to the generation model to grant it a sense of the sentences it was supposed to create. `{AMB_TYPE}` was formatted with the name and a description of the ambiguity type as follows:

- vp: VP Attachment Ambiguity, occurring when it is unclear which part of a sentence a verb phrase is intended to modify
- pp: PP Attachment Ambiguity, occurring when it is unclear which part of a sentence a prepositional phrase is intended to modify
- anaph: Anaphoric Ambiguity, which occurs when it is unclear which antecedent a particular anaphor refers to within a given context
- ellip: Ellipsis Ambiguity, involving the omission of words or phrases that are understood from the context
- adj: Adjective Scope Ambiguity, occurring when it is unclear how far the influence of an adjective extends within a sentence
- vb: Verb Scope Ambiguity, occurring when it is unclear how far the influence of a verb extends within a sentence

Annotator	I2T	T2I
A	7, 8, 9, 0, 1, 2	3, 4, 5, 6
B	3, 4, 5, 6	0, 1, 2, 7, 8, 9

Table 7: Two human annotators denoted as **A** and **B** were given splits so that one annotator doesn't evaluate a sample in both directions.

Setting	p_0	**N**	**95% CI (Wilson)**
Conj (Dual)	0.11	300	[0.079, 0.151]
Conj (T2I / I2T)	0.333	300	[0.282, 0.389]
Other Types (Dual)	0.25	200	[0.195, 0.314]
Other Types (T2I / I2T)	0.50	200	[0.431, 0.569]

Table 8: 95% Wilson confidence intervals around the random chance levels for each task setting.

- conj: Conjunction Scope Ambiguity, occurring when it is unclear how far the influence of a conjunction coordinate, such as AND/OR extends within a sentence

Image generation prompts were carefully made to have the same semantic structure as that of the texts used for the experiments. For image styles, "coloured cartoon" and "coloured photography" were used.

B.2 Annotator Allocation

ViLStrUB consists of 700 ambiguous sentences, each equipped with two or three vision and language interpretations. We constructed human evaluation sets in two settings: I2T and T2I, each consisting of 700 samples. 700 samples in each setting were split into 10, which we labelled from 0 to 9. Two annotators were given the splits in a way that one annotator doesn't evaluate the same sample in both directions, in order to mitigate memory effects. Annotators were given the splits as in Table 7.

C Model Card

- **CLIP-ViT**
  `openai/clip-vit-large-patch14-336` (HuggingFace)
- **CLIP-RN50**
  RN50 model trained by openai from `https://github.com/mlfoundations/open_clip`
- **CLIP-RN101**
  RN101 model trained by openai from `https://github.com/mlfoundations/open_clip`
- **SigLIP**
  `google/siglip-so400m-patch14-384` (HuggingFace)
- **MetaCLIP**
  `facebook/metaclip-h14-fullcc2.5b` (HuggingFace)
- **MetaCLIP2**
  `facebook/metaclip-2-worldwide-114` (HuggingFace)
- **OpenCLIP-ViT**
  `laion/CLIP-ViT-H-14-laion2B-s32B-b79K` (HuggingFace)
- **OpenCLIP-convnext**
  convnext_xxlarge model trained by laion2b/s34b/b82k_augreg from `https://github.com/mlfoundations/open_clip`
- **EVA-CLIP**
  `BAAI/EVA-CLIP-18B` (HuggingFace)
- **LLaVA-NeXT**
  `llava-hf/llava-v1.6-mistral-7b-hf` (HuggingFace)
- **Qwen3-VL**
  `Qwen/Qwen3-VL-8B-Instruct` (HuggingFace)
- **Gemma3**
  `google/gemma-3-12b-it` (HuggingFace)

D Prompts used for LLM-based Generative Models for Experiments

Table 9 shows the prompts used for LLM-based generative VLMs at the main experiment in Table 2 and Table 3. Additional analysis in Figure 5 was done using the same prompt for I2T.

E Confidence Interval

Table 8 is the 95% Wilson confidence intervals around the random chance levels for the main experiment in Table 2.

	Prompt
I2T	**SYSTEM:** You are a careful vision-language classifier. Task: - You will receive ONE image and {num_captions} candidate captions. - The captions correspond to labels {labels_str}. Your job is to choose which single caption best matches the image. Rules: - Respond in a forced choice format: choose ONLY one label from: {labels_str}. Then provide a short explanation. Output format (must follow EXACTLY): <LETTER> Explanation: <your reasoning in one short sentence> **USER:** You will see the image first. Then you will see the candidate captions.
T2I	**SYSTEM:** You are a careful vision-language classifier. Task: - You will receive ONE caption and {num_images} candidate images in a fixed order. - The images correspond to labels {labels_str} in the same order. Your job is to choose which single image best matches the caption. Rules: - Respond in a forced choice format: choose ONLY one label from: {labels_str}. Then provide a short explanation. Output format (must follow EXACTLY): <LETTER> Explanation: <your reasoning in one short sentence> **USER:** Now you will see the images shown in the same order as the labels.

Table 9: Two human annotators denoted as **A** and **B** were given splits so that one annotator doesn't evaluate a sample in both directions.