A Benchmark for Evaluating Structural Ambiguity Resolution in Vision & Language Models

Anonymous ACL submission

Abstract

Structural ambiguity in natural language, where a single sentence permits multiple meanings arising from syntax hierarchy, is a crucial challenge for language understanding. Visual context offers a valuable source of additional information for resolving such ambiguity, making Vision & Language Models (VLMs) a promising solution. As a first step towards evaluating the ability of VLMs to capture such structural ambiguity, we constructed a large-scale benchmark covering a variety of ambiguity types and including both classification and generation tasks. Quantitative results on recent models reveal clear limitations, and our analysis identifies persistent challenges in aligning visual and structural semantics, offering insights for future research.

1 Introduction

Structural ambiguity, where a sentence supports multiple interpretations due to its syntactic structure, remains a key challenge in natural language understanding. For example, task-oriented dialogue systems require accurate interpretation of user instructions for executing them correctly (Bodonhelyi et al., 2024). Unlike lexical ambiguity, structural ambiguity arises beyond the word level and demands deeper integration of linguistic reasoning and contextual understanding. Resolving it not only prevents misinterpretation, but also grants the systems capacities for syntactic reasoning and deeper linguistic understanding.

Disambiguation typically requires additional contextual information, such as dialogue history, prosody, or visual input (DeVault and Stone, 2009; Widiaputri et al., 2023; Kuribayashi and Baldwin, 2025). Among these sources, visual input is particularly valuable, as it is one of the most informative and pervasive modalities available to real-world systems (Hutmacher, 2019). Figure 1 illustrates a use case of such visual information in a task-oriented system. Given the ambiguous instruction "Bring me a yellow cap and sneakers", the system identifies two possible interpretations based on the visual scene, where only the cap is clearly yellow, but two kinds of sneakers exist. Recognising the ambiguity, the system infers that the colour attribute is underspecified and asks for clarification.

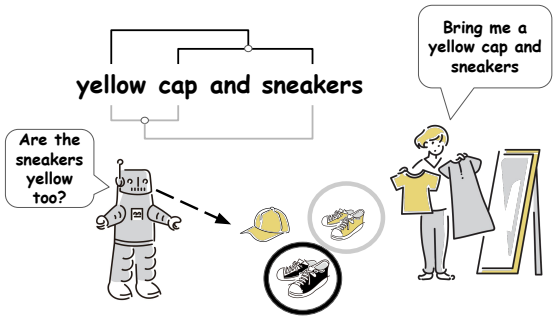

Figure 1: Use case of a task-oriented dialogue system equipped with visual disambiguation ability. The system grounds multiple candidate interpretations from an ambiguous instruction and requests clarification.

When we address such tasks using Vision & Language Models (VLMs), the models must recognise that the given sentence or image may have multiple interpretations, and appropriately identify where the ambiguity lies. While previous studies (Thrush et al., 2022; Yuksekgonul et al., 2022) have explored the limitations of VLMs in understanding linguistic structure, ambiguity arising from syntactic interpretation has not been sufficiently discussed. Existing datasets that address structural ambiguity (Berzak et al., 2015; Mehrabi et al., 2023) face challenges in terms of both scale and quality, making it difficult to comprehensively evaluate the capabilities of recent VLMs.

As a first step towards enabling VLMs to resolve such ambiguity, this study proposes a new benchmark for evaluating how VLMs solve structural ambiguity using visual information. We also assess the extent to which existing VLMs can handle such

challenges.

By following previous work, we categorise structural ambiguity into seven types and construct a large-scale dataset consisting of ambiguous sentences, their possible interpretations, and corresponding visual scenes. To generate the visual scenes, we used a large-scale image generation model (Betker et al., 2023), and the outputs were manually evaluated for quality.

Using this benchmark, we conducted a comprehensive evaluation of existing VLMs and also tested whether humans can successfully resolve the same ambiguity. Our results reveal that while most of the tasks in the benchmark are solvable by humans, existing VLMs exhibit a significant performance gap, indicating an explicit limitation in their current ability to resolve structural ambiguity using visual context.

2 Related Work

2.1 VLMs

The core idea behind VLMs is to pre-train on large-scale datasets using contrastive learning, aligning images and their corresponding captions so that they share similar representations in a joint embedding space. Models following this paradigm, such as CLIP and SigLIP (Radford et al., 2021; Zhai et al., 2023), have demonstrated strong zero-shot image classification performance. The text and image encoders from these models are often reused as backbone components in downstream applications, including text-to-image generation (Rombach et al., 2021; Ramesh et al., 2023) and multimodal response generation (Laurençon et al., 2024). Our research evaluates both these contrastive and derived generation models for resolving structural ambiguity using visual context.

2.2 VLMs and Compositional Understanding

Despite their impressive performance, VLMs have been shown to struggle with capturing structural compositionality, the way words combine to form meaning in a sentence. For instance, CLIP (Radford et al., 2021) has difficulty correctly associating adjectives with their intended target nouns (Tang et al., 2023). Other studies have proposed benchmarks that test compositional understanding by altering word order to shift sentence meaning (Thrush et al., 2022; Yuksekgonul et al., 2022). Our research extends this line of inquiry by evaluating VLM's ability to resolve structural ambiguity, a challenge that goes beyond compositional understanding alone and requires distinguishing between multiple valid syntactic interpretations of the same input.

2.3 Visual Disambiguation

Resolving structural ambiguity with visual input presents unique challenges, particularly due to the specificity of linguistic context, which makes it difficult to reuse existing datasets. 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). It has served as a foundational resource for subsequent studies in the field. However, its limited size and annotation quality have posed challenges for broader applicability (Mehrabi et al., 2023; Yamaki et al., 2023). The Text-to-Image Ambiguity Benchmark (TAB) expanded upon LAVA by improving the quality and quantity of textual annotations, aiming to support structural disambiguation in text-to-image generation (Mehrabi et al., 2023). Nevertheless, due to the nature of its generation task, the benchmark lacks annotated visual references, limiting its use in evaluating how models interpret visual input. Building on recent advances in generation models (Betker et al., 2023; OpenAI, 2023), our work aims to collect more comprehensive and well-aligned data, enabling a clearer evaluation of VLMs' ability to use visual information for resolving structural ambiguity.

3 Data Construction

Our motivation lies in constructing a benchmark containing structural ambiguity that can be appropriately resolved by referencing both linguistic and visual information to improve the performance of VLMs. According to this motivation, we built a benchmark which incorporates both classification and generation tasks, corresponding to seven ambiguity types inspired by those defined in the TAB dataset. In this section, we describe the procedure used to construct the benchmark.

3.1 Ambiguity Type Definition

Our categorisation builds on the ambiguity types defined in TAB. We exclude one category unrelated o linguistic structure ("fairness") and subdivide the original conjunction category into three (Appendix A). As a result, we define seven categories of structural ambiguity, which were selected

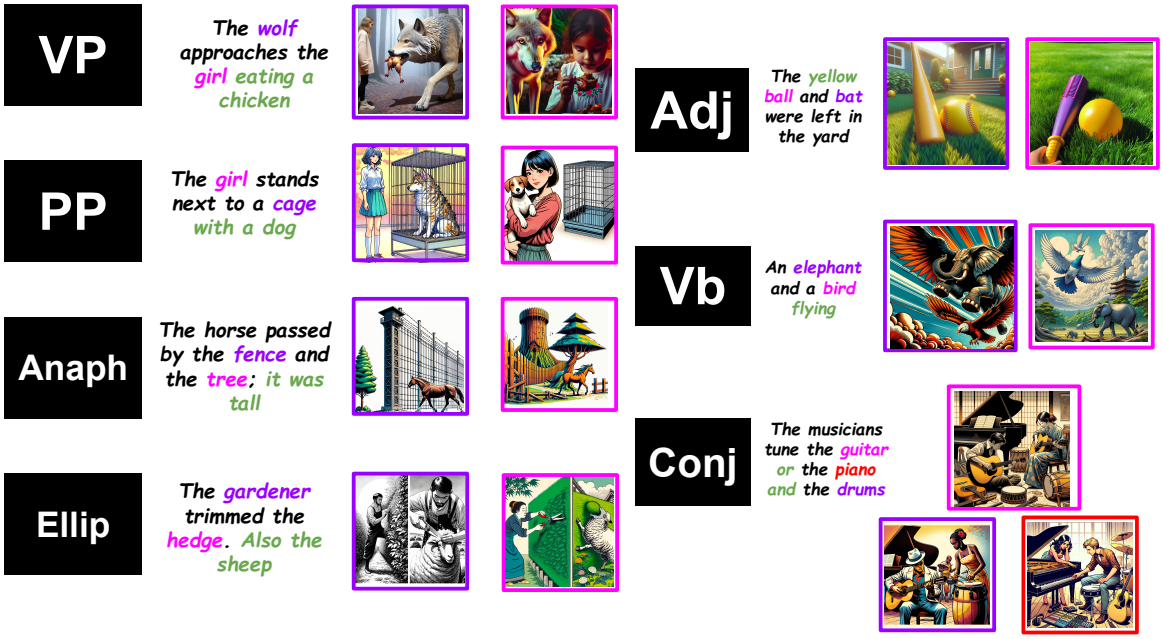

Figure 2: Example sentence and corresponding visual interpretations from each type defined in Section 3.1

to balance linguistic representativeness with visual clarity, ensuring that each ambiguity can be effectively grounded in image-text pairs. Figure 2 presents one example for each category along with its alternative interpretations.

- Verb Phrase Attachment (VP): Ambiguity arises when a verb phrase could attach to more than one part of the sentence. (e.g. The wolf approaches the girl eating a chicken)
- Preposition Phrase Attachment (PP): A prepositional phrase can modify multiple possible heads (e.g. The girl stands next to a cage with a dog)
- Anaphora (Anaph): A pronoun or referring expression has more than one plausible antecedent. (e.g. The horse passed by the fence and the tree; it was tall)
- Ellipsis (Ellip): An omitted phrase can be interpreted in multiple ways. (e.g. The gardener trimmed the hedge. Also the sheep)
- Adjective Scope (Adj): An adjective can modify either a single noun or an entire coordinated noun phrase. (e.g. The yellow ball and bat were left in the yard)
- Verb Scope (Vb): A verb-derived modifier may apply to one or more coordinated elements. (e.g. An elephant and a bird flying)
- Conjunction Scope (Conj): Coordinating conjunctions (e.g. and, or) group sentence elements in more than one way. (e.g. The musicians tune the guitar or the piano and the drums)

3.2 Data Collection

Building on the structure of TAB, we generate new ambiguous-disambiguated text pairs using GPT-4[1] (OpenAI, 2023). Each ambiguous sentence is paired with two or three disambiguated counterparts, following the format introduced in TAB (e.g. concatenation: "The boy approaches the chair with a bag. The bag is on the chair."). The dataset includes 700 ambiguous sentences (100/type), each paired with 2 to 3 disambiguated interpretations (e.g. for Conjunction Scope, every sentence has three interpretations), resulting in 1503 disambiguated sentence pairs with corresponding images. For each disambiguated caption, we generate an image using DALL-E 3[2] (Betker et al., 2023). To introduce visual diversity, images out of half of the sentences are cartoon-style, the rest photo-realistic (for prompts used for generation models, refer to Appendix B.). In cases where certain prompts (e.g. involving violence or political figures) were re-

[1]We used the `gpt-4o-mini` API variant from December 3 to 9, 2024.

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

jected by the generation model, we rephrased them to preserve the intended ambiguity while ensuring compatibility. (Appendix C).

4 Experimental Settings

To assess how well current VLMs can resolve structural ambiguity, we conduct experiments using our dataset across two tasks: classification and generation. These tasks are designed to evaluate model performance and highlight specific limitations in aligning visual and linguistic information.

4.1 Classification

VLMs trained for zero-shot classification via large-scale contrastive pretraining aim to align representations of images and their corresponding textual descriptions. This setting allows us to evaluate how well subtle semantic differences are reflected and matched across the visual and textual modalities.

4.1.1 Task Specifics

The classification task involves correctly matching disambiguated captions with the corresponding images. We report classification accuracy (Acc) as the evaluation metric. Below, we describe the setup assuming two disambiguated interpretations per ambiguous caption (note: for conjunction scope ambiguity, there are three options).

- Text-to-Image (T2I): Given an ambiguous caption A, we assume two disambiguated versions C_1, C_2, and their corresponding images I_1, I_2. In a trial, the model receives one caption (e.g., C_1) and both images (I_1, I_2). The model succeeds if it assigns a higher similarity score to the correct image (e.g., $sim(C_1, I_1) > sim(C_1, I_2)$). Accuracy is computed as the proportion of successful trials over the dataset.

- Image-to-Text (I2T): In this direction, the model is given one image (e.g., I_1) and both captions (C_1, C_2). A trial is considered successful if the model assigns a higher similarity score to the correct caption. Accuracy is computed in the same manner as T2I.

- Dual: This task evaluates whether a disambiguated image-caption pair is correctly matched in both directions. The model is given both images and both captions, and the match is counted as successful if, for a given pair (C_1, I_1), both $sim(C_1, I_1) > sim(C_1, I_2)$ and $sim(I_1, C_1) > sim(I_1, C_2)$. The accuracy is calculated as the proportion of correctly matched pairs (0, 1, or 2 per instance).

4.1.2 Evaluated Models

Given their strong zero-shot capabilities and broad applicability, our evaluation focuses primarily on contrastive VLMs based on the CLIP paradigm. Specifically, we evaluate the following models: CLIP (Radford et al., 2021), SIGLIP (Zhai et al., 2023), and its variants. As a result, our targets are CLIP[3] (Radford et al., 2021), OpenCLIP[4] (Cherti et al., 2023), MetaCLIP[5] (Xu et al., 2024), EVA-CLIP[6] (Sun et al., 2023), SigLIP[7] (Zhai et al., 2023), and SigLIP2[8] (Tschannen et al., 2025), with the versions as large as possible.

4.1.3 Human Evaluation

For comparison, we also report human performance on the classification task. Two annotators evaluated the entire dataset, with each annotator handling half of the samples shuffled across ambiguity types. To avoid potential memory effects, the two were given disjoint subsets for the T2I and I2T conditions. Overall human performance is reported as the aggregate number of correct decisions across both annotators. For the Dual condition, a sample was considered correct only if both annotators selected the correct match independently in their respective directions.

4.2 Generation

The generation task assesses whether a model can revise or preserve a caption based on accompanying visual input. This includes rewriting ambiguous capions to resolve structural ambiguity, preserving accurate disambiguated captions, or correcting captions that mismatch the image. This task setup reflects practical use cases where a model must interpret language in the context of a visual scene and adjust output accordingly.

4.2.1 Task Details

Each input to the model consists of a caption, an image, and an instruction prompt. Depending on the input, the model is expected to either preserve

[3] openai/clip-vit-large-patch14-336
[4] hf-hub:laion/CLIP-ViT-g-14-laion2B-s12B-b42K
[5] facebook/metaclip-h14-fullcc2.5b
[6] BAAI/EVA-CLIP-18B
[7] google/siglip-so400m-patch14-384
[8] google/siglip2-so400m-patch16-512

the caption, revise it to correct a mismatch, or disambiguate it using the visual context. We define three input scenarios:

- Ambiguous caption + disambiguating image: The caption contains structural ambiguity, and the model must rewrite it into a disambiguated one that aligns with the image.
- Disambiguated caption + matching image: The caption is already correct. The model should preserve the semantic structure, optionally rephrasing it without altering its meaning.
- Disambiguated caption + mismatching image: The caption does not match the visual input. The model must revise it to reflect the content of the image.

To guide the model, we design two types of instruction prompts: one general (PROMPT-GENERAL), and one elaborated (PROMPT-ELABORATED) with explicit mention of the ambiguity type. This allows us to assess whether models benefit from task-specific guidance during disambiguation. The full list of prompts and hyper parameters is provided in the Appendix D.

4.2.2 Metrics

The generation task is evaluated by comparing the model-generated caption with a gold (reference) caption that aligns with the intended structural semantics of the input image. We adopt two complementary evaluation metrics.

BERTScore BERTScore (Zhang et al., 2020) computes similarity between text sequences using contextualized embeddings from pre-trained language models, capturing semantic similarity beyond surface-level matching. Unlike traditional metrics such as BLEU (Papineni et al., 2002), which rely on exact n-gram overlap, BERTScore evaluates how well the generated caption captures the meaning of the gold caption. Given the semantic focus of our task, we find BERTScore particularly suitable.

Smatch Smatch (Cai and Knight, 2013) measures the similarity between two Abstract Meaning Representation (AMR) graphs (Banarescu et al., 2013), which encode the meaning of sentences in a predicate-logic-like form. This allows comparison of deeper structural semantics, abstracting away superficial textual differences. We convert captions into AMR graphs using amrlib[9]. While the AMR parser is not flawless, Smatch provides an additional perspective on how well models preserve or recover the underlying meaning. We use it to gauge the alignment between generated and gold captions beyond surface-level text similarity.

4.2.3 Evaluated Models

We evaluate one closed-source model (GPT-4o (OpenAI, 2023)) alongside 7 open-source models (Gemma3[10] (Team, 2025), LLaVA1.6[11] (Liu et al., 2024), Qwen2.5-VL[12] (Bai et al., 2025), Pixtral[13] (Agrawal et al., 2024), Idefics3[14] (Laurençon et al., 2024), and Chameleon[15] (Team, 2024)). Due to hardware limitations, we were unable to test the largest versions of each open model. However, previous evaluations suggest that the smaller variants used here demonstrate broadly similar performance patterns, with only slight degradations in accuracy. Our goal is to analyse general trends in how these models handle structural ambiguity. The inclusion of GPT-4o allows us to benchmark performance at the higher end of model capability, providing a reference point for future research.

5 Results

5.1 Classification

Table 1 presents the classification results. As outlined in the Appendix E, the expected accuracy by random chance is 50% for ambiguity types with two possible interpretations, and approximately 33% for the Conjunction Scope (Conj) type, which involves three options. For the Dual task, where success requires correct matches in both directions (T2I and I2T), the random baseline is 25% for two-option types and approximately 11% (1/9) for Conj.

Across most two-option ambiguity types, both T2I and I2T performance hovers near the random baseline, indicating that current VLMs struggle to reliably resolve structural ambiguity. In contrast, human performance mostly exceeds 0.9, demonstrating that the task is well-posed and that the benchmark captures structurally resolvable cases.

We observe a performance gap between task directions: I2T generally outperforms T2I, and

[9] https://github.com/bjascob/amrlib
[10] google/gemma-3-12b-it
[11] llava-hf/llava-v1.6-vicuna-13b-hf
[12] Qwen/Qwen2.5-VL-7B-Instruct
[13] mistral-community/pixtral-12b
[14] HuggingFaceM4/Idefics3-8B-Llama3
[15] facebook/chameleon-7b

Model	VP	PP	Anaph	Ellip	Vb	Adj	Conj	All
Text-to-Image (T2I)								
CLIP	0.525	0.510	0.493	0.495	0.515	0.525	0.377	0.484
SigLIP	0.505	0.530	0.502	0.559	0.545	0.530	0.347	0.492
SigLIP2	0.530	0.550	0.522	0.515	0.510	0.495	0.390	0.494
MetaCLIP	0.515	0.505	0.493	0.550	0.525	0.535	0.357	0.488
OpenCLIP	0.505	0.510	0.498	0.540	0.475	0.505	0.380	0.480
EVA-CLIP	0.500	0.495	0.507	0.520	0.510	0.525	0.387	0.485
Human	0.985	0.970	0.881	0.926	0.905	0.920	0.930	0.931
Image-to-Text (I2T)								
CLIP	0.505	0.565	0.517	0.540	0.570	0.535	0.520	0.535
SigLIP	0.570	0.590	0.552	0.515	0.645	0.545	0.463	0.548
SigLIP2	0.515	0.595	0.562	0.545	0.585	0.510	0.517	0.545
MetaCLIP	0.500	0.540	0.567	0.545	0.565	0.570	0.510	0.540
OpenCLIP	0.540	0.545	0.532	0.525	0.590	0.605	0.573	0.560
EVA-CLIP	0.490	0.550	0.522	0.579	0.605	0.590	0.543	0.554
Human	0.965	0.925	0.945	0.896	0.895	0.945	0.933	0.929
Dual								
CLIP	0.260	0.335	0.274	0.327	0.300	0.325	0.190	0.281
SigLIP	0.325	0.310	0.274	0.322	0.400	0.310	0.187	0.300
SigLIP2	0.285	0.355	0.294	0.287	0.240	0.245	0.160	0.259
MetaCLIP	0.290	0.280	0.264	0.327	0.320	0.335	0.183	0.279
OpenCLIP	0.310	0.330	0.303	0.312	0.330	0.325	0.237	0.302
EVA-CLIP	0.290	0.285	0.338	0.322	0.385	0.315	0.203	0.299
Human	0.960	0.915	0.851	0.847	0.855	0.915	0.913	0.896

Table 1: Classification results. The best score across per type is doublelined and per model is boldfaced.

Dual accuracy often exceeds the respective baseline. This directional asymmetry is particularly pronounced in the Conj type, where T2I performance remains near chance while I2T shows a marked improvement, resulting in Dual accuracy substantially higher than the random chance. One possible explanation is that the two modalities have different bias which don't work in accordance with structural ambiguity

Finally, it is worth noting that EVA-CLIP, despite its 18B parameter size, does not show notable superiority over smaller models. This may imply that current contrastive pre-training alone, regardless of model scale, is not yet sufficient to handle structural ambiguity effectively.

5.2 Generation

Table 2 presents the results of the generation task. Although all models were instructed to perform the same task using the equivalent information, the stability of generated responses differed significantly. Gemma3, Idefics3, and Chameleon often produced unstable outputs, such as off-topic answers or captions containing multiple conflicting interpretations. In particular, Gemma3 occasionally failed to recognize the provided image, resulting in the weakest performance. Meanwhile, other models generated more stable responses, scoring around 0.8 in BERTScore, indicating sound performance, though with room for improvement. Notably, while it might be expected that GPT-4o, with the largest parameter size, would excel at a task requiring multi-step inference such as disambiguation, LLaVA1.6 (13B) outperformed it, and the best performance came from Qwen2.5 (7B). This suggests that disambiguation ability currently shows little correlation with model size, and structural ambiguity has not yet been a core focus of large-scale VLM training. Prompts with more elaborated descriptions of the ambiguity type generally led to improved performance, but not to a degree considered reliably effective. Additionally, while Smatch scores followed a similar trend to BERTScore, the overall performance on Smatch was lower, indicating further limitations in structural-level understanding.

6 Discussion

To better understand model behaviour, we further analyse embedding similarities and alignment patterns across modalities. We identify two primary limitations that contribute to the overall low accu-

Model	VP	PP	Anaph	Ellip	Vb	Adj	Conj	All
				BERTScore				
GPT-4o	0.824 (+0.004)	0.806 (+0.010)	0.870 (+0.017)	0.827 (+0.011)	0.738 (+0.004)	0.797 (+0.010)	0.741 (+0.008)	0.793 (+0.009)
Gemma3	0.678 (+0.175)	0.625 (+0.183)	0.615 (+0.188)	0.605 (+0.015)	0.668 (+0.090)	0.654 (+0.131)	0.543 (+0.208)	0.617 (+0.149)
LLaVA1.6	0.852 (+0.003)	0.828 (+0.023)	0.846 (+0.078)	0.808 (+0.029)	0.784 (+0.022)	0.838 (-0.007)	0.755 (+0.019)	0.808 (+0.023)
Qwen2.5	**0.859** (+0.002)	**0.866** (-0.008)	**0.911** (-0.006)	**0.863** (+0.006)	**0.875** (-0.029)	**0.846** (-0.023)	**0.822** (+0.030)	**0.858** (+0)
Pixtral	0.834 (+0.009)	0.829 (+0.012)	0.830 (+0.054)	0.846 (+0.021)	0.765 (-0.002)	0.804 (+0)	0.755 (+0.079)	0.802 (+0.031)
Idefics3	0.677 (+0.023)	0.666 (+0.018)	0.665 (+0.043)	0.666 (+0.011)	0.646 (+0.007)	0.663 (+0.009)	0.591 (+0.035)	0.646 (+0.023)
Chameleon	0.667 (+0.074)	0.679 (+0.069)	0.698 (0.043)	0.621 (+0.024)	0.580 (+0.060)	0.683 (+0.038)	0.602 (+0.107)	0.642 (+0.065)
				Smatch				
GPT-4o	0.670 (+0.020)	0.630 (+0.020)	0.700 (+0.030)	0.640 (+0.020)	0.520 (+0)	0.680 (+0.020)	0.520 (+0.020)	0.623 (+0.019)
Gemma3	0.360 (+0.340)	0.290 (+0.350)	0.330 (+0.240)	0.240 (+0.010)	0.370 (+0.220)	0.400 (+0.310)	0.270 (+0.350)	0.323 (+0.260)
LLaVA1.6	**0.730** (-0.010)	0.670 (+0.040)	0.700 (+0.100)	0.610 (+0.040)	0.570 (-0.020)	0.710 (-0.050)	0.560 (+0.040)	0.650 (+0.020)
Qwen2.5	0.720 (-0.010)	**0.710** (+0)	**0.770** (+0)	0.680 (+0.030)	**0.730** (-0.020)	**0.770** (-0.020)	0.620 (+0.060)	**0.714** (+0.006)
Pixtral	0.690 (+0)	0.640 (+0.030)	0.660 (+0.070)	**0.690** (+0.030)	0.610 (+0.020)	0.760 (+0.030)	**0.630** (+0.080)	0.669 (+0.037)
Idefics3	0.470 (+0.020)	0.450 (+0.010)	0.490 (+0.040)	0.480 (+0.010)	0.370 (+0.010)	0.520 (+0.010)	0.370 (+0.030)	0.450 (+0.019)
Chameleon	0.450 (+0.060)	0.420 (+0.050)	0.480 (+0.040)	0.330 (+0.030)	0.300 (+0.050)	0.500 (+0.020)	0.390 (+0.10)	0.410 (+0.037)

Table 2: Generation results. Inside the brackets are score differences by PROMPT-ELABORATED. The best performance per type is boldfaced.

racy of current VLMs:

- A lack of sensitivity to structural differences in textual meaning
- Overreliance on surface-level visual features that distract from the disambiguation-relevant semantics.

6.1 Structural Meaning not Reflected in Embeddings

Our classification results indicate that current models struggle to resolve structural ambiguity, often performing near random chance. While generation performance is higher particularly in BERTScore, Smatch which emphasises structural accuracy reveals persistent gaps. One possible hypothesis is that current text encoders are not sufficiently sensitive to syntactic distinctions between lexically similar sentences.

Table 1 shows that I2T accuracy consistently outperformed T2I. Prior work such as Winoground (Thrush et al., 2022), attributes this to stronger text encoders or modality-specific biases. Our findings support this, showing that the text modality may carry systematic biases that limit semantic separation.

	OpenCLIP		Qwen2.5	
Type	amb-dis	dis-dis	amb-dis	dis-dis
VP	0.971	0.994	0.970	0.993
PP	0.971	0.983	0.969	0.989
Anaph	0.987	0.993	0.990	0.999
Ellip	0.953	0.949	0.940	0.941
Vb	0.952	0.976	0.951	0.996
Adj	0.963	0.986	0.973	0.998
Conj	0.966	0.988	0.988	0.997
All	0.966	0.983	0.970	0.990

Table 3: Cosine similarity between the captions. amb-dis signifies the comparison between the ambiguous caption and its disambiguated version, and dis-dis signifies the similarity between the disambiguated candidates from the same ambiguous sentence.

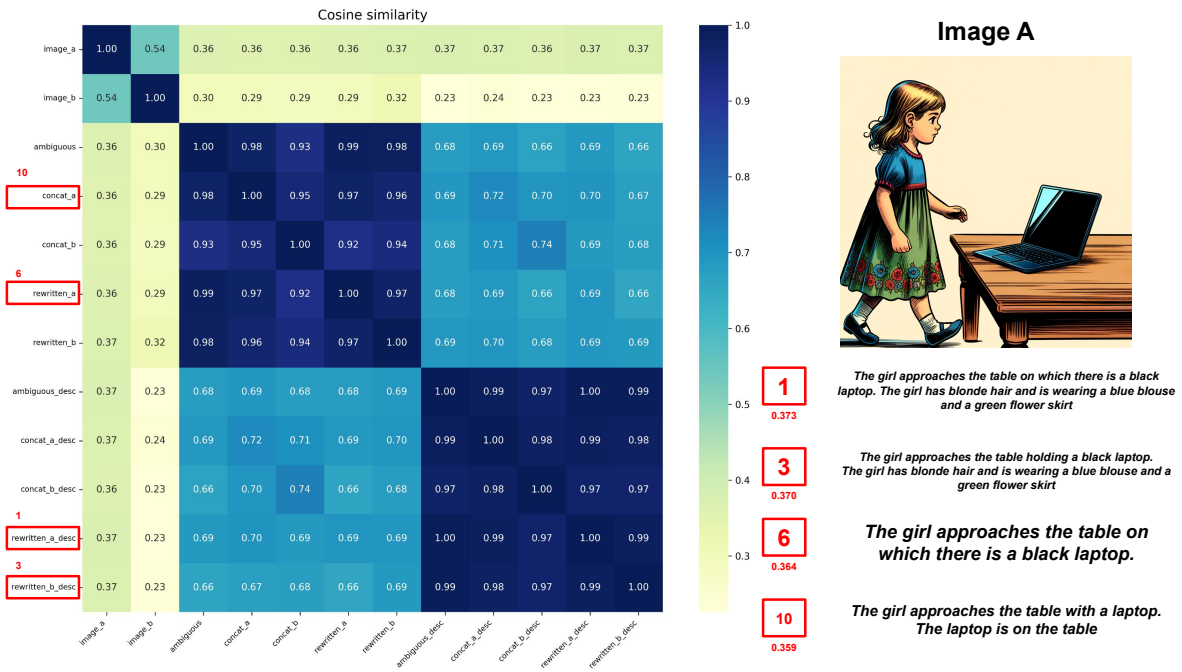

Figure 3: Error case analysis from OpenCLIP model. The left side shows a cosine similarity heatmap between two images and ten caption variants, including the original ambiguous caption ("The girl approaches the table with a black laptop."), two disambiguated captions via concatenation (concat_a, concat_b), and two via rewriting (rewritten_a, rewritten_b), each with and without added visual description. The visual description is: "The girl has blonde hair and is wearing a blue blouse and a green flower skirt." The right panel shows cosine similarity rankings of the ten captions relative to the image embedding

To further investigate, we compute cosine similarities between the captions (Table 3), which reveals that disambiguated captions are often too similar to both their ambiguous originals and to each other in embedding space. This overlap likely contributes to poor classification performance explains why Smatch scores are lower in generation, given its sensitivity to structural mismatches.

6.2 Visual Detail Dominance

Visual modality allows a wider range of surface expressions than text, often distracting models from structural cues. In Figure 3, OpenCLIP fails to match a disambiguated PP caption (concat_a) with its image. We analysed rewritten variants with and without added superficial visual descriptions. Cosine similarity reveals clustering based on visual detail rather than syntactic meaning. Even an incorrect caption with visual cues outperformed the correct one without, indicating that the models prioritise superficial information over structural information, hindering effective disambiguation.

These findings imply that VLMs may overfit to descriptive visual features and fail to generalise across expression styles. Effective disambiguation will require models to abstract visual meaning beyond literal object descriptions and better integrate this with structural cues from text.

7 Conclusion

We introduced a benchmark for evaluating VLMs on structural ambiguity resolution using visual information. Covering seven ambiguity types, our dataset supports both classification and generation tasks to assess model behaviour. Results show that classification performance remains near random chance, and although generation outputs score well for BERTScore, structural evaluation with Smatch reveals major gaps. Analysis indicates that semantic differences between captions are poorly reflected in embedding space, and that models often focus on superficial visual details rather than disambiguating cues. These findings highlight the need for improved cross-modal reasoning and structural sensitivity. Future work should aim to develop models that abstract beyond surface-level features and align syntactic interpretation more reliably with visual context.

Limitations

- While our human evaluation on our data suggests its plausibility in Table 1, more thorough analysis is required regarding the data's statistics. Specifically, diversity in both ambiguous sentences and images would be an important factor justifying our collected dataset in assessing the VLMs' disambiguation ability.

- 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 difference. Also, more evaluation would be needed on closed model such as Gemini (Reid et al., 2024).

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Singh Chaplot, Jessica Chudnovsky, Saurabh Garg, Théophile Gervet, Soham Ghosh, Am'elie H'eliou, Paul Jacob, Albert Q. Jiang, Timothée Lacroix, Guillaume Lample, Diego de Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, and 18 others. 2024. Pixtral 12B. arXiv:2410.07073.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. arXiv:2502.13923.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

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*, abs/2402.02136.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 748–752.

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.

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.

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

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.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. arXiv:2408.12637.

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.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

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.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2023. Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Ben jamin Lee, and 655 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

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.

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.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. arXiv:2405.09818.

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

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.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim M. Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier H'enaff, Jeremiah Harmsen, Andreas Steiner, and Xiao-Qi Zhai. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv:2502.14786.

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 *The 12th International Conference on Learning Representations*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with bert. In *Proceedings of the 8th International Conference on Learning Representations*.

A Revision of Ambiguity Type from TAB

TAB (Mehrabi et al., 2023) dataset has defined the following seven types 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.)
- Miscellaneous: Ambiguity arising from objects having multiple possible roles or properties. (e.g. The chicken is ready to eat.)
- 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.)

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—only three samples were present in the entire TAB dataset—to support meaningful experiment. As a result, we excluded both of these types 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.

B Prompts used for Generation Models for Data Collection

For data collection, we used the following prompts

- Text Generation: 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 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 visualize the scene. The topic is `{AMB_TYPE}`. From now on, I will show you some of the examples.

- Image Generation: Follow the prompt and styles to create a faithful image.
  Prompt: `{args.prompt}`
  Styles: `{args.style}`

For text generation, previous samples from TAB were given to the generation model to grant it a sense of 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, occuring 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

- adjscope: Adjective Scope Ambiguity, occurring when it is unclear how far the influence of an adjective extends within a sentece

- verbscope: Verb Scope Ambiguity, occurring when it is unclear how far the influence of a verb extends within a sentence

- conjscope: 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 photograpy" were used.

C Example Cases of Inappropriate Texts in 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, 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.

D Prompts and Hyper Parameters for Generation Experiment

Instruction prompts given to the generation models for our experiment in Section 4.2 are designed to convey the identical meaning across the models assessed. However since some models had different requirements for the format, we adjusted them accordingly:

- GPT-4o: I am giving you an image and its caption. The caption might contain in itself structural ambiguity.
  Your job is to output the final caption with modification if necessary.
  Your job process is gonna be as follows:

  1. If the caption is unambiguous, then look at the image to see if the caption matches the image semantics. If so, the final caption stays the same from the original. Don't change it.
  2. If the caption isn't ambiguous but doesn't match the image semantics, then rewrite the caption to match the image content. Do it like the following example:

     e.g. The man approached the chair with a bag. the bag is in the man's arms.
     But if the bag in the image is on the chair, then just change it like:
     The man approached the chair with a bag. the bag is on the chair.

     You should not focus on the visual details too much.
  3. If the caption is ambiguous, then disambiguate it by looking at the given image.

  Your output should be a single sentence, the final caption.

- Chameleon: Look at this caption `{input_caption}` and the image `<image>`. And rewrite the caption in the following process:

  1. If there is no structural ambiguity, output the caption as it is.
  2. However, if the caption is unambiguous but doesn't match the image semantics, rewrite it.
  3. If there is structural ambiguity, output the modified caption given the image information.

  Your output should be a single sentence, the final caption.

- Others: I am giving you an image and its caption. The caption might contain in itself structural ambiguity. Your job is to output the final caption with modification if necessary by looking at the image.

  Job process:

  1. If there is no ambiguity, output the caption as it is.
  2. However, if the caption is unambiguous but doesn't match the image semantics, rewrite it.
  3. If there is structural ambiguity, output the modified caption given the image information.

  Your output should be a single sentence, the final caption.

  [Image]

  Caption: `{input_caption}`

Above prompts are input prompts without

E Random Chance

Task	Model	VP	PP	Anaph	Ellip	Vb	Adj	Conj	All
T2I	CLIP	0.5	0.5	0.498	0.495	0.5	0.5	0.333	0.468
	SigLIP	0.5	0.5	0.498	0.495	0.5	0.5	0.333	0.468
	SigLIP2	0.5	0.5	0.498	0.495	0.5	0.5	0.333	0.466
	MetaCLIP	0.5	0.5	0.498	0.495	0.5	0.5	0.333	0.468
	OpenCLIP	0.5	0.5	0.498	0.495	0.5	0.5	0.333	0.466
	EVA-CLIP	0.5	0.5	0.498	0.495	0.5	0.5	0.333	0.466
I2T	CLIP	0.5	0.5	0.498	0.495	0.5	0.5	0.333	0.568
	SigLIP	0.5	0.5	0.498	0.495	0.5	0.5	0.333	0.468
	SigLIP2	0.5	0.5	0.498	0.495	0.5	0.5	0.333	0.466
	MetaCLIP	0.5	0.5	0.498	0.495	0.5	0.5	0.333	0.468
	OpenCLIP	0.5	0.5	0.498	0.495	0.5	0.5	0.333	0.466
	EVA-CLIP	0.5	0.5	0.498	0.495	0.5	0.5	0.333	0.466
Dual	CLIP	0.23	0.295	0.274	0.262	0.26	0.32	0.103	0.240
	SigLIP	0.275	0.255	0.239	0.238	0.275	0.3	0.093	0.230
	SigLIP2	0.255	0.245	0.259	0.267	0.22	0.315	0.077	0.224
	MetaCLIP	0.225	0.235	0.279	0.277	0.285	0.315	0.077	0.231
	OpenCLIP	0.215	0.215	0.249	0.277	0.29	0.31	0.333	0.220
	EVA-CLIP	0.24	0.28	0.239	0.252	0.31	0.28	0.07	0.228

Table 4: Classification results based on random chance trials. The input text is the original ambiguous sentence.