Bridging the Visual Gap: Fine-Tuning Multimodal Models with Knowledge-Adapted Captions

Moran Yanuka[τ] **Assaf Ben Kish**[τ] **Yonatan Bitton**[G] **Idan Szpektor**[G] **Raja Giryes**[τ]

[τ]Tel Aviv University [G]Google Research

Abstract

Recent research increasingly focuses on training vision-language models (VLMs) with long, detailed image captions. However, small-scale VLMs often struggle to balance the richness of these captions with the risk of hallucinating content during fine-tuning. In this paper, we explore how well VLMs adapt to such captions. To quantify caption quality, we propose Decomposed NLI (DNLI), an evaluation framework that breaks down generated captions into individual propositions, assessing each in isolation. This fine-grained analysis reveals a critical balance between capturing descriptive details and preventing hallucinations. Our findings show that simply reducing caption complexity or employing standard data curation techniques does not effectively resolve this issue. To tackle this challenge, we introduce Knowledge Adapted (KnowAda) fine-tuning, a data-centric approach that automatically adapts training data with the model's existing knowledge and visual understanding. KnowAda minimizes hallucinations while preserving high descriptiveness. We validate this approach across several small-scale VLMs (up to 7B parameters) and dense caption datasets, demonstrating that KnowAda effectively balances hallucination reduction and descriptiveness. Our results show that KnowAda outperforms various baselines in both automatic metrics and human evaluations. The code is available here.

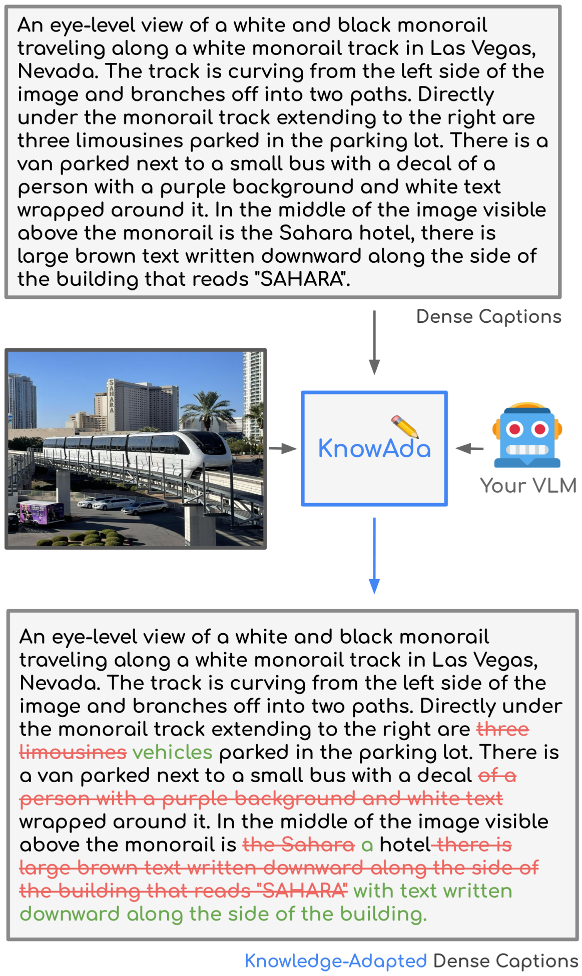

Figure 1: KnowAda identifies knowledge gaps of a VLM and adapts the dense caption accordingly. The KnowAda dense captions are better suited for downstream fine-tuning of the VLM.

1 Introduction

Fine-tuning pretrained multimodal models for generating dense image captions is common in both research and practical applications, such as assisting visually impaired individuals. Recent work has focused on creating high-quality, descriptive captions through human annotations (Onoe et al., 2024; Garg et al., 2024; Deitke et al., 2024) and synthetic generations from models like GPT-4 (Chen et al., 2023a, 2024a) and Gemini (Singla et al., 2024), including extensions that integrate expert models for enhanced detail (Li et al., 2024). These datasets enable the fine-tuning of models to create detailed descriptions in specific styles, distinguishing them from the zero-shot capabilities of pretrained models. However, smaller multimodal models (e.g., up to 7 billion parameters), which are essential for real-time applications, frequently face challenges

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 10497–10518
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

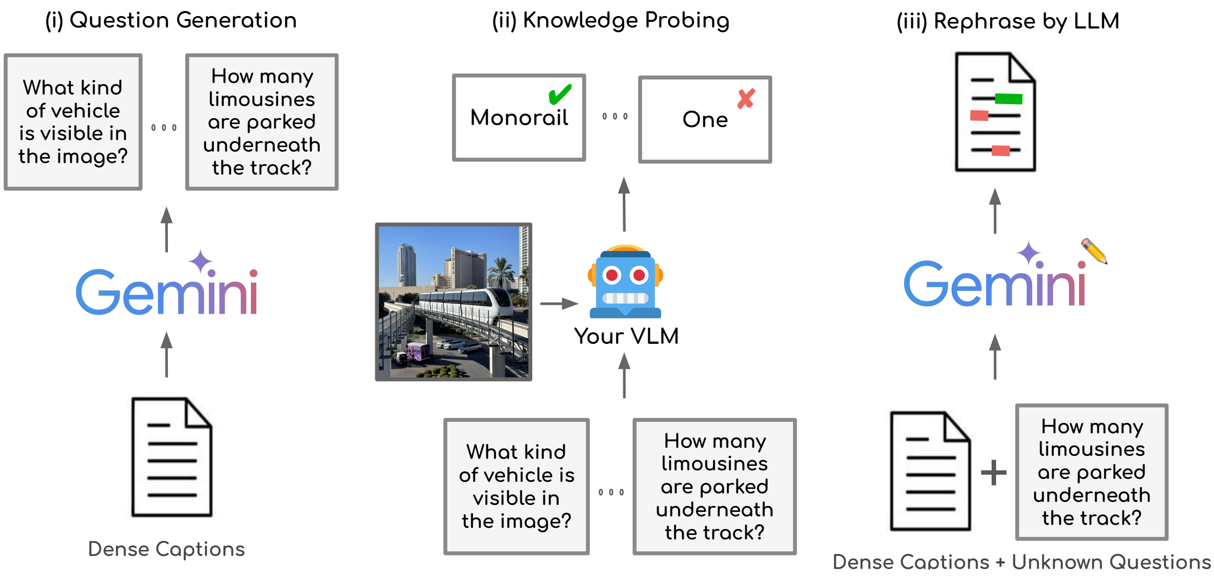

Figure 2: **Our proposed KnowAda pipeline.** We first probe the knowledge of the VLM, identifying the known and unknown parts of the image description, by generating questions about the visual content of the image mentioned in the caption. Then, KnowAda identifies the knowledge gaps by judging the VLM answers to these questions. Finally, KnowAda adapt the description to match these gaps (e.g., removing the number of limousines mentioned in the caption, which relates to a question the model failed to answer).

in capturing fine-grained visual details during fine-tuning, resulting in hallucinations.

Consider a fine-tuning dataset for dense captioning, like the one in Figure 1 from DOCCI (Onoe et al., 2024), alongside a pretrained vision-language model trained on tasks like captioning, VQA, and OCR. If, for instance, the model has only encountered low-resolution images during pretraining, it may struggle to identify fine details, such as the drawing on the purple van or the hotel name in the background, which require higher resolution. This issue extends beyond resolution to other visual challenges in modern VLMs (Tong et al., 2024; Wu et al., 2024; Zhang et al., 2024a). We hypothesize that fine-tuning the model on overly complex captions may increase hallucinations, as the model is compelled to predict details it cannot accurately perceive or understand.

Recent works on large language models (LLMs) have shown that fine-tuning primarily adapts pre-existing factual knowledge for specific tasks, with the majority of this knowledge being encoded during the pretraining phase (Geva et al., 2020; Meng et al., 2022; Roberts et al., 2020). Rather than acquiring new information, fine-tuning typically *activates and refines pretrained knowledge*, which remains largely stable throughout the process (Zhou et al., 2024b). Furthermore, Gekhman et al. (2024) demonstrated that fine-tuning on content not grounded in a model's pre-existing factual-knowledge can lead to an increase in hallucinations. Similarly, Yu et al. (2024) showed that attempting to distill GPT4V into a smaller, less capable VLM significantly increases hallucinations. Motivated by these findings, we hypothesize that one possible cause of increased hallucinations is the excessive visual complexity of captions relative to the model's pretrained capabilities. We propose a method to mitigate this effect.

To better adapt pretrained models to dense caption datasets, we introduce KnowAda, a model-specific adaptation technique that simplifies complex details in dense captions. The KnowAda pipeline, shown in Figure 2, automatically identifies visual knowledge gaps between a pretrained VLM and an image-caption pair by generating questions related to the image's visual content. It then modifies the captions to align with the model's visual knowledge and capabilities, producing adapted captions that enhance fine-grained control while balancing a low hallucination rate with high descriptiveness.

Evaluating dense image captions requires attention to two critical factors: descriptiveness, which captures the image's details, and hallucination rate, which measures factual accuracy. Traditional metrics, such as those based on similarity to reference captions (Papineni et al., 2002; Banerjee and Lavie, 2005; Zhang et al., 2019; Reimers, 2019) or CLIP-based alignment (Sarto et al., 2023; Radford et al., 2021), often fall short when applied to long captions. These approaches penalize valid variations in phrasing and fail to distinguish between factual accuracy and token overlap or semantic similarity, rendering them inadequate for identifying hallucinations in detailed captions. Existing hallucination metrics are similarly limited, focusing primarily on short captions or object-level errors (Rohrbach et al., 2018; Li et al., 2023c; Ben-Kish et al., 2024), or relying on other VLMs (Jing et al., 2023; Liu et al., 2023), which themselves can hallucinate during the verification process. To address these limitations, we propose Decomposed NLI (DNLI), a novel evaluation framework that breaks captions into propositions and assesses their entailment with the detailed ground truth description. DNLI offers a more reliable measure of both descriptiveness and accuracy, demonstrating strong alignment with human judgments.

Our results show that training with KnowAda captions offers a favorable balance between descriptiveness and fidelity when fine-tuning LLaVA-7B, outperforming other data-centric baselines. To demonstrate the consistency of KnowAda across different models and datasets, we fine-tuned several multimodal models, ranging from 2 billion to 7 billion parameters, on two different dense captioning datasets. Across all models, KnowAda consistently reduced hallucinations compared to training with the original captions, as confirmed by both automatic and human evaluations.

Stated explicitly, our contributions are: (I) We show that small-to-medium-scale VLMs underperform when fine-tuned directly on dense caption datasets, exhibiting increased hallucinations and reduced descriptive accuracy; (II) To address this we propose KnowAda, a model-dependent augmentation method that fills knowledge gaps in captions, reducing hallucinations while preserving high descriptive accuracy; (III) We demonstrate the effectiveness of KnowAda through extensive experiments, supported by both quantitative metrics and qualitative human evaluations; (IV) We introduce DNLI, a novel evaluation framework for dense captions that offers a more fine-grained analysis and shows strong correlation with human annotations.

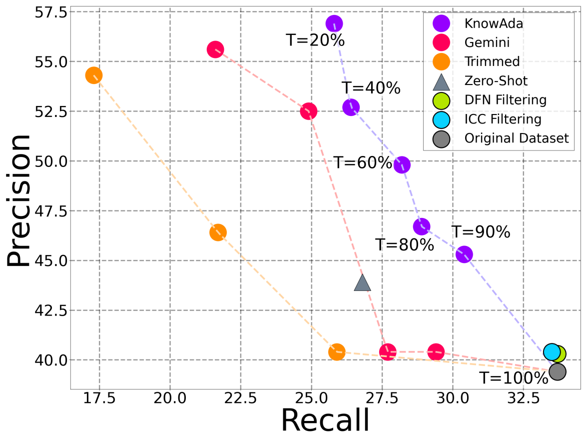

Figure 3: **Dense captioning descriptiveness precision-recall results** for LLaVA-7B fine-tuned with DOCCI captions, adapted using different methods. "Trimmed" refers to naive removal of sentences, while "Gemini" involves prompting Gemini to simplify the caption by removing difficult details of varying degrees. KnowAda consistently achieves better precision-recall balance. Original captions corresponds to KnowAda with a threshold of $T = 100\%$, where no information is classified as unknown.

2 KnowAda

We begin by introducing our caption adaptation method for dense captioning datasets. KnowAda comprises three stages, as shown in Figure 2. First, we use an LLM to generate visual questions from each dense caption. Next, these questions are employed to probe the VLM's pretrained visual knowledge and identify parts of the image caption the model struggles with. Finally, an LLM adapts the image descriptions by editing out the unknown parts. Below, we elaborate on each step of the pipeline.

2.1 VLM Knowledge Probing

In order to detect the visual attributes of the image that are unknown to the VLM but are mentioned in the image description, we probe the VLM knowledge. It is done by generating visual questions which can be answered by the image description, then letting the VLM answer these questions, and finally measuring if the responses are correct. We provide further details regarding each stage below.

Finding the unknown questions. Given a dataset D containing image descriptions, we aim to find all the parts of the description that are unknown to the VLM. Following prior work on uncertainty estimation (Cheng et al., 2024; Gekhman et al., 2024), the steps are as follows:

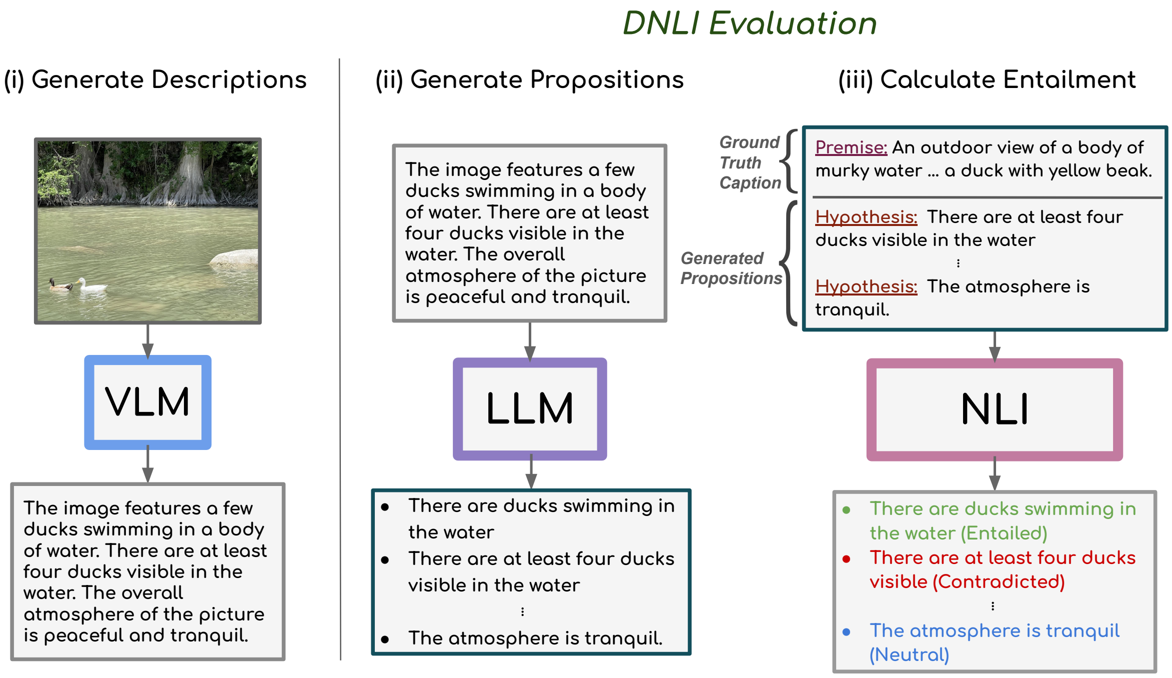

Figure 4: **DNLI Evaluation**. Given a generated description by a VLM, we decompose it to atomic propositions. Then, we classify each proposition to either entailed, contradicted or neutral, conditioned on the ground-truth description. Finally, we calculate the descriptiveness and contradiction based on the number of entailed and contradicted propositions.

1. For each image caption $d \in D$, generate n questions $Q = \{q_1, q_2, \ldots, q_n\}$. Note that n is caption dependent (e.g., more questions are generated for longer captions).

2. For each question $q_i \in Q$, sample m answers. Let $A_i = \{a_{i1}, a_{i2}, \ldots, a_{im}\}$ represent the set of m answers for question q_i.

3. Evaluate each answer $a_{ij} \in A_i$ to determine how difficult it is for the specific VLM. Let $C_i \subseteq A_i$ be the set of correct answers and $I_i \subseteq A_i$ be the set of incorrect answers for question q_i.

4. Calculate the difficulty of each question q_i based on the ratio of incorrect to incorrect and correct answers. The difficulty score df_i for question q_i is given by:

$$df_i = \frac{|I_i|}{|C_i| + |I_i|}$$

Here, $|C_i|$ and $|I_i|$ are the number of correct and incorrect answers respectively. A higher value of df_i indicates a more difficult question. For a threshold T, a question with $df_i > T$ is defined as an *unknown question*.

We assess the accuracy of the model's responses by prompting a LLM to evaluate each generated answer in relation to the given question and the ground truth description.

For example, consider the two questions shown in Figure 2 (ii). If the model correctly answered "`Monorail`" for the question "`What kind of vehicle is visible in the image?`" in 6 out of 10 sampled instances, with 4 incorrect answers, the difficulty for this question is calculated as $df = \frac{4}{10}$. For the second question, "`How many limousines are parked underneath the track?`", the model correctly answered "`3`" only 4 out of 10 times, resulting in a difficulty of $df = \frac{6}{10}$. With a threshold of $T = 50\%$, the first question is classified as known, while the second as unknown. However, if the threshold is raised to $T = 70\%$, both questions would be considered unknown.

For example, consider the two visual questions shown in Figure 2 (ii) with the following prediction accuracies:

- For the question "`What kind of vehicle is visible in the image?`", the model correctly answers "`Monorail`" in $6/10$ instances, resulting in $df = 4/10 = 0.4$.
- For the question "`How many limousines are parked underneath the track?`", the model correctly answers "`three`" in $4/10$ instances, yielding $df = 6/10 = 0.6$.

With a difficulty threshold T, the classification of these questions changes:

Model	FT Captions	Contradiction ↓				Descriptiveness ↑				# Words
		Precision		Recall		Precision		Recall		
		Auto	Human	Auto	Human	Auto	Human	Auto	Human	
PaliGemma	Synthetic	38.9	19	30	15.5	47.9	81	**39.2**	**67.8**	72
PaliGemma	Synthetic KA	**32.4**	**18.4**	**20.8**	**13**	**55.2**	**81.6**	36.4	61.2	54
TinyLLaVA	Synthetic	38.1	52.1	49.1	45	35.1	47.9	**26.3**	40.8	71
TinyLLaVA	Synthetic KA	**22.9**	**34**	**40.2**	**22.2**	**48.1**	**66**	25.4	**40.9**	51
LLaVA-7B	Synthetic	39.1	19.3	39.1	16.2	47.3	80.7	**39.7**	**65.3**	79
LLaVA-7B	Synthetic KA	**31**	**15.8**	**31**	**9.3**	**58.4**	**84.2**	34.5	47.4	54
PaliGemma	Human	41.6	20.4	24.6	14.9	46.7	79.6	24.6	**43.2**	62
PaliGemma	Human KA	**38.3**	**18.3**	**22.2**	**13.6**	**49.4**	**81.7**	**28.7**	38.9	65
TinyLLaVA	Human	53	51.8	39.5	38.8	31.9	51.8	**22.4**	**31.4**	100
TinyLLaVA	Human KA	**42.6**	**34.5**	**19.6**	**12.5**	**46.9**	**65.5**	19.1	22.5	53
LLaVA-7B	Human	47.2	33.4	39.7	33.2	39.4	66.6	**33.7**	**48.1**	109
LLaVA-7B	Human KA	**33.7**	**17.1**	**16.7**	**11.2**	**56.9**	**82.9**	25.8	31.8	55

Table 1: **Dense captioning results** over the test sets of DOCCI when fine-tuning on original human-annotated captions, synthetic captions, and KnowAda-adapted captions (denoted as KA) with a threshold of 20%. “Automatic (Auto)” refers to model-based NLI evaluation, while “Human” refers to evaluations based on human labeling.

1. For $T = 50\%$:
   - The first question is Known (since $df = 0.4 < 0.5$)
   - The first question is Unknown (since $df = 0.6 > 0.5$)
2. For $T = 70\%$:
   - The first question is Unknown (since $df = 0.4 < 0.7$)
   - The second question is Unknown (since $df = 0.6 < 0.7$)

Image description rewriting. After identifying the questions unknown to the model for each image description, we prompt a LLM, namely, Gemini with several in-context examples to remove or edit the parts of the caption that answer these unknown questions, while keeping the other details intact. For instance, in Figure 1, the questions about the number and type of cars parked in the parking lot is unknown to the small VLM. Therefore the corresponding questions are passed to Gemini, which edits “`three limousines`” to “`vehicles`”.

Formally, for each image description d_i, we use all the corresponding questions classified as unknown to the VLM according to the threshold T, denoted as $Q_i = \{q_{i,1}, q_{i,2}, \dots, q_{i,\bar{n}}\}$, to prompt the LLM to rewrite the description by removing information associated with Q_i. See Appendix A.1 for the prompts used at each stage.

2.2 KnowAda Data Analysis

We start by examining the characteristics of the dense captioning datasets, the impact of KnowAda on the adapted captions, and the types of questions utilized for visual probing within the KnowAda pipeline.

Datasets. We use two variations of dense caption datasets in our experiments: DOCCI (Onoe et al., 2024), a human-annotated dataset rich in visual details, and DOCCI images paired with synthetic captions generated by Gemini-Pro-1.5, which are designed to be highly visually descriptive. These datasets differ in caption style and level of visual detail, allowing us to demonstrate the robustness of KnowAda across varying data distributions.

Dataset characteristics. Figure 6 illustrates the overlap of unknown questions across the different models. While there is a core set of unknown questions common to all three models, each model also has its own unique set of unknown questions.

Table 2 presents statistics for each dataset, including the average number of unknown questions per model and dataset, as well as the average word count in both the original and KnowAda captions.

The data indicates that the average number of unknown questions and the average word count in the rewritten captions are relatively consistent across different models, where the human-authored DOCCI captions containing slightly more challenging questions compared to the synthetic captions.

Visual questions category distribution. To verify that KnowAda generates diverse types of visual questions, we use LLaMa-3-70B (Dubey et al., 2024), quantized to 4 bits, to classify 12,422 questions generated for 1,000 images from the DOCCI test set into categories defined by SeedBench (Li et al., 2023a). For the classification task, we provide one-few shot examples from each SeedBench category. The resulting distribution of questions is presented in Figure 5. As observed, the diversity in the question types suggests that KnowAda performs knowledge probing across a wide range of visual tasks.

3 Decomposed NLI Evaluation

Next, we introduce DNLI, a proposition-decomposition evaluation framework for dense-captioning. It evaluates the quality of a generated dense image caption using two criteria: *descriptiveness*, which measures accuracy and detail, and *contradiction*, which measures how much the caption contradicts the ground truth. Inspired by prior work on paragraph summarization evaluation (Ernst et al., 2021; Zhang and Bansal, 2021; Ernst et al., 2021) and retrieval (Chen et al., 2023b), DNLI assesses descriptiveness and consistency through proposition extraction, as outlined below and shown in Figure 4. Additional qualitative examples are in the appendix.

Propositional decomposition. Given a generated image description, we use a Gemini to decompose it into a set of atomic propositions, capturing individual, verifiable claims. This enables fine-grained evaluation. To avoid duplicates, we instruct the model to include only unique propositions.

Natural Language Inference (NLI) analysis. We next evaluate the entailment of propositions generated from the image using both our automatic method, where Gemini is used to compute textual entailment, and through human annotators who directly assess visual entailment:

Automatic Textual Entailment: Each atomic proposition is assessed with a Natural Language Inference (NLI) model. This model compares each proposition (the hypothesis) to a ground truth description of the image (the premise), determining whether the proposition is entailed, contradicted, or neutral with respect to the description. This evaluation forms the basis for our descriptiveness and contradiction metrics. See details and prompts in Appendix A.1.

Human-Based Visual Entailment Evaluation: We employ Amazon Mechanical Turk to engage three independent annotators for each proposition, tasking them with directly assessing its entailment based on the corresponding image. To ensure quality, we administer a qualification test and select the top-performing annotators. Each proposition is evaluated by a distinct set of three annotators, and the final entailment label is determined by a majority vote. For consistency, we randomly sample 20 images per model, generating an average of 175 propositions per model. In total, 2,812 unique propositions were evaluated. Further details are provided in Appendix A.3.

Descriptiveness metric. We quantify the *descriptiveness* of the generated caption using two measures: recall and precision.

Descriptiveness Recall: The proportion of ground truth propositions that are entailed by the generated description:

$$\text{Descriptiveness Recall} = \frac{|\text{Entailed}|}{|\text{Ground Truth}|},$$

where the entailed propositions are those identified by the NLI model as described above.

Descriptiveness Precision: The proportion of entailed propositions relative to the total number of propositions in the generated description:

$$\text{Descriptiveness Precision} = \frac{|\text{Entailed}|}{|\text{Generated}|},$$

where the generated propositions are those found in the model's caption. The precision represents the likelihood that a given proposition extracted from the generated caption, would be entailed.

Contradiction metric. The *contradiction* precision and recall are calculated in a similar manner:

$$\text{Contradiction Precision} = \frac{|\text{Contradicted}|}{|\text{Ground Truth}|}$$

$$\text{Contradiction Recall} = \frac{|\text{Contradicted}|}{|\text{Generated}|}$$

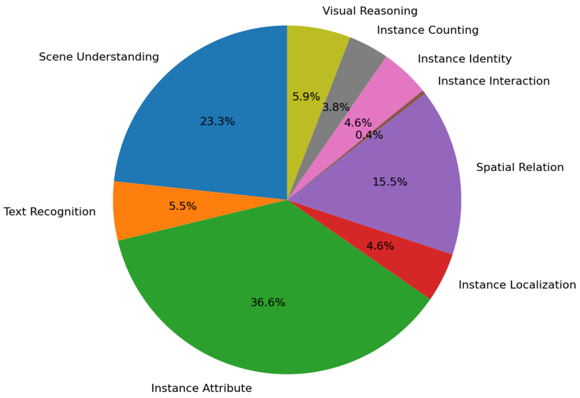

Figure 5: **Distribution of visual question categories** generated from captions in the first stage of KnowAda, classified using SeedBench (Li et al., 2023a) definitions.

Captions	Model	$\bar{C}_o$	$\bar{C}_r$	$\bar{Q}_{unk}$
Human	PaliGemma	122	84.2	5.5
Human	TinyLLaVA	122	80.9	6.3
Human	LLaVA-7B	122	80.7	6.3
Synthetic	PaliGemma	92	65.8	4.4
Synthetic	TinyLLaVA	92	63.3	5.2
Synthetic	LLaVA-7B	92	63.7	5.2

Table 2: Statistics of the original and KnowAda adapted captions with $T = 20\%$ over the human annotated and synthetic version of DOCCI captions. $\bar{C}_o$ and $\bar{C}_r$ denotes the mean number of words in the original caption and caption after applying KnowAda, and $\bar{Q}_{unk}$ denotes the mean number of questions unknown to the model.

Together, these metrics provide a comprehensive evaluation of the generated captions. Descriptiveness precision and recall assess the accuracy and coverage of the content, respectively. Contradiction precision indicates the likelihood of a false proposition in a caption, while contradiction recall measures the total number of contradictions.

Note that the neutral labels are discarded in the automatic evaluation, as they could represent either subjective propositions (e.g., "vibrant atmosphere") or visual claims not described in the original captions. These labels could be either entailed or contradicted, but it is unclear which.

4 Experimental Settings

Our experiments concentrate on training dense captioning models. We evaluate the performance of models fine-tuned on captions curated through various methods against those fine-tuned on captions adapted using KnowAda.

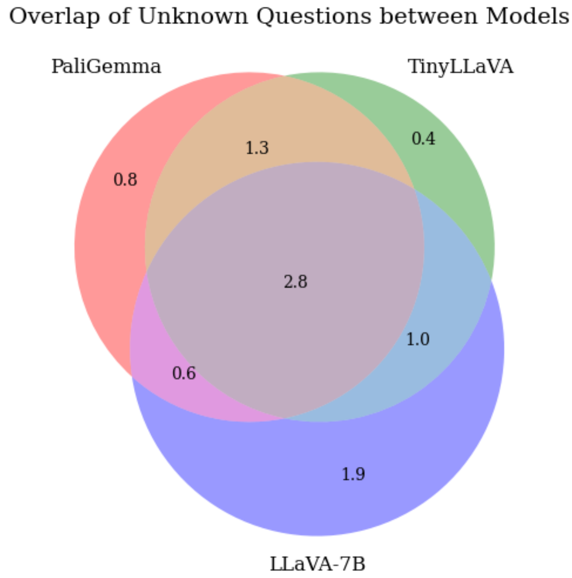

Figure 6: Overlap in $\bar{Q}_{unk}$ (average number of unknown questions per caption) across different models on the DOCCI training set for $T = 20\%$. Each model has a unique set of unknown questions, with about 50% of these shared among all three models.

We start by comparing KnowAda to data curation methods for multimodal datasets according to the DataComp challenge (Gadre et al., 2024), specifically ICC (Yanuka et al., 2024) and DFN (Fang et al., 2023). We also compare our approach to the following baselines that might address the hallucination-descriptiveness trade-off:

- *Caption Trimming*: A progressive method that removes varying numbers of sentences to simplify captions.
- *Gemini Simplification*: An approach prompting Gemini to remove difficult details from captions to varying degrees.

For all methods in this experiment, we fine-tuned LLaVA-1.5-7B (Liu et al., 2024) using Low Rank Adaptation (Hu et al., 2021).

To ensure that KnowAda is robust across multiple models and datasets, we fix the threshold at 20% for classifying questions as unknown and fine-tune three models: PaliGemma (Beyer et al., 2024), TinyLLaVA (Zhou et al., 2024a), and LLaVA-1.5-7B (Liu et al., 2024). We fine-tune on two DOCCI variations: one using the original DOCCI captions, and another using synthetically generated captions created by Gemini, which were prompted to be visually descriptive. We refer to Appendix A.4 for an experiment on the larger-scale PixelProse (Singla

Method	Contradiction ↓		Descriptiveness ↑		# Words
	Precision	Recall	Precision	Recall	
KnowAda Random	34.7	17.2	57.9	23.9	55
KnowAda	**33.7**	**16.7**	**58.9**	**25.8**	55

Table 3: **Design choices ablations**.We ablate the effect of removing unknown information versus removing random information in the KnowAda pipeline. Removing unknown information improves performance across all metrics.

et al., 2024) dataset. We evaluate the models using an automatic NLI model and human annotators (Section 3). In all experiments, we split the DOCCI test set into 1,000 samples for evaluation, while 4,000 samples are used for the reported test set.

5 Results

This section presents our experiments, highlighting the impact of KnowAda over competing baselines.

5.1 KnowAda Achieves Better Descriptiveness-Hallucination Trade-off

Figure 3 illustrates that fine-tuning on KnowAda captions consistently provides the best balance between high descriptiveness and low hallucination rates compared to competing baselines. While ICC and DFN offer slight improvements in precision, they do not facilitate further gains due to a lack of control over the trade-off. Although trimming and rephrasing captions allow for some control over precision and recall, they yield inferior results compared to KnowAda across all thresholds.

5.2 KnowAda Is Consistent Across Multiple Models and Datasets

Table 1 shows that fine-tuning on KnowAda captions consistently reduces the hallucination rate while maintaining high descriptiveness across different trained models. This superior performance is evident for both human-annotated and synthetically generated dense captions, as confirmed by our automatic and human evaluation pipelines.

Specifically, KnowAda significantly reduces the contradiction rate in terms of both precision and recall. However, while it improves precision in descriptiveness, it decreases recall. The results in Table 1 are obtained using a stringent threshold of $T = 20\%$, prioritizing hallucination reduction. KnowAda allows increased recall by enabling the selection of a less stringent threshold, facilitating a trade-off between precision and recall, as illustrated in Figure 3. Future work may explore how to maintain KnowAda's improvement in hallucinations while ensuring high recall in descriptiveness.

5.3 Textual Entailment is Highly Correlated to Human Annotated Visual Entailment

We computed the Phi correlation coefficient between the majority agreement labels (ground-truth human annotations) and the labels generated by the NLI model for each proposition. The results yielded a Phi coefficient of $\phi = 0.73$ for the DOCCI original captions and $\phi = 0.67$ for the synthetic captions. These strong positive correlations indicate a significant relationship between the human annotations and our proposed automatic evaluation. It suggests that our text-only NLI model effectively aligns with human judgments of visual entailment, demonstrating its reliability in distinguishing between contradictory and entailed propositions based solely on dense captions.

Moreover, the correlation difference between original and synthetic captions shows that DOCCI human-authored captions are generally more detailed and reliable compared to the synthetic captions, enhancing the effectiveness of the automatic entailment computation using them.

Note that the results from our human annotation process showed an average majority agreement of 82.6% for the "Contradicted" labels and 89.7% for the "Entailed" labels, indicating a high level of consensus among annotators.

5.4 Necessity of Removing Unknown Information

We perform an ablation study to assess the importance of removing VLM-unknown information by following the KnowAda procedure for a fixed $T = 20\%$, but instead of removing information linked to unknown questions, we randomly remove information, a method referred to as *KnowAda Random* in Table 3. As expected, this results in worse performance across all metrics, demonstrating that unknown information is indeed a problematic factor in captions that impacts the hallucinations.

6 Conclusions

This work focuses on fine-tuning small-to-medium-scale vision-language models by aligning dense captions with the models' existing knowledge and visual understanding. It aims to address the challenge of balancing rich descriptiveness with factual correctness in multimodal models, especially when models have limited capacity to process complex visual details in dense captions. By probing models with visual questions and adapting captions to exclude unknown information, we reduce hallucinations while maintaining high descriptive accuracy. Our results, supported by both human and automatic evaluations, suggest that this strategy can lower hallucinations compared to training on original captions or competing data curation baselines.

Furthermore, we introduce a proposition-based evaluation framework that provides fine-grained analysis of generated captions, offering deeper insights into the balance between descriptiveness and factuality. We believe that our findings contribute to improving fine-tuning methods in VLMs from a data-centric approach, particularly in resource-constrained environments where balancing descriptiveness and accuracy is important for real-world applications. We hope that this work will encourage further research into addressing the challenges of dense captions in VLMs.

7 Related Work

Our research is connected to the field of dense image captioning, particularly in addressing unknown information and enhancing fine-tuning datasets. Below, we provide an overview of previous works in each of these areas.

7.1 Dense Image Captioning

Traditional image captioning datasets like COCO (Chen et al., 2015) often feature brief captions with limited visual detail. Recently, interest has grown in creating longer and more complex captions. For example, Onoe et al. 2024 and Garg et al., 2024 developed a dataset of 15K human-annotated, richly descriptive captions. Shabtay et al. (2025) used captions of figures of arXiv papers. Additionally, Chen et al., 2023a and Singla et al., 2024 employed GPT-4 and Gemini to generate detailed captions. We show that these datasets are more effective for fine-tuning when adapted to the specific model being trained.

7.2 Training On Unknown Information

Several works examine the relationship between unknown information and downstream model performance. Gekhman et al. (2024) show that fine-tuning LLMs on low-confidence examples encourages hallucinations, and suggest that fine-tuning should not introduce new knowledge, but only teach the model to make use of existing knowlegde. Zhang et al. (2024b) suggest optimizing for higher-confidence responses via Direct Preference Optimization (DPO), where the confidence of each response is estimated via self-evaluation. Piché et al. (2024) train a utility function that encourages predicting only high-certainty responses. The function is trained over synthetic data, which is collected iteratively, and is composed of fused high-confidence predicted facts. Additionally, Xu et al. (2024) introduce a refusal mechanism which encourages the model to reject questions that do not align with its existing knowledge. We demonstrate that training a model on questions it is fundamentally lacking the ability to answer, rather than solely on factual knowledge it lacks—particularly in the multimodal domain—results in similar performance degradation. Moreover, our method only rejects the unknown parts of a data sample, while maintaining the information that is useful for training, thus avoiding throwing out useful training samples which only include some unknown parts.

7.3 Data Curation in Fine-Tuning

Recent research has focused on refining the fine-tuning process for pretrained models, especially through instruction tuning. Zhou et al. (2024b) demonstrated that fine-tuning on a small, high-quality instruction dataset can yield superior results. Li et al. (2023b) explored training a language model on a subset of instruction data, using loss discrepancies to inform filtering strategies. However, this approach doesn't extend to image captioning, where no instructions are available. Lin et al. (2024) proposed reducing noisy token influence by assigning lower weights, while Chen et al. (2024b) trained a network to filter out low-loss instruction samples. In contrast, our approach adapts specific parts of captions misaligned with the model's capabilities, rather than filtering or reweighing the entire caption. Our strategy leverages visual knowledge and emphasizes sequences and conceptual coherence, moving beyond token-level adjustments.

8 Limitations

While KnowAda demonstrates notable effectiveness, several limitations must be considered.

Firstly, KnowAda is model-dependent, probing the knowledge of each model in isolation. This means that the data must be tailored for each model independently, leading to increased computational overhead.

Secondly, while we show that KnowAda achieves better performance in balancing the descriptiveness-hallucination trade-off, it does not fully resolve this issue. Reducing contradictions to the ground truth often decreases descriptiveness recall (but not precision).

Lastly, our analysis focused specifically on dense caption generation, which is a limited task. Expanding KnowAda to other tasks, such as visual question answering (VQA), could represent an interesting avenue for future research.

Ethics Statement

This work focuses on measuring and mitigating hallucinations in visual-language models (VLMs). As such, it is expected to increase the reliability of VLMs and the ability to measure their performance, which is important when using them in real-world systems. This is expected to have a positive impact on the use of VLMs in society. However, we recognize that the foundation models used in the KnowAda construction and evaluation pipeline could propagate biases. We anticipate further research into such biases before relying on our work beyond the research environment. The human annotation study performed in this work received the required IRB approval from our institution's ethics committee.

Acknowledgments

This work was partially supported by Google and the TAU Center for Artificial Intelligence and Data Science. The authors thank Morris Alper and Nimrod Shabtay for their valuable feedback and assistance.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. 2024. Mitigating open-vocabulary caption hallucinations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22680–22698, Miami, Florida, USA. Association for Computational Linguistics.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. 2024. Paligemma: A versatile 3b vlm for transfer. *Preprint*, arXiv:2407.07726.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. 2024b. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. *arXiv preprint arXiv:2402.12501*.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Dong Yu, and Hongming Zhang. 2023b. Dense x retrieval: What retrieval granularity should we use? *arXiv preprint arXiv:2312.06648*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? *arXiv preprint arXiv:2401.13275*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger, and Ido Dagan. 2021. Proposition-level clustering for multi-document summarization. *arXiv preprint arXiv:2112.08770*.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023. Data filtering networks. *arXiv preprint arXiv:2309.17425*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. 2024. Imageinwords: Unlocking hyper-detailed image descriptions. *arXiv preprint arXiv:2405.02793*.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Junlong Jia, Ying Hu, Xi Weng, Yiming Shi, Miao Li, Xingjian Zhang, Baichuan Zhou, Ziyu Liu, Jie Luo, Lei Huang, and Ji Wu. 2024. Tinyllava factory: A modularized codebase for small-scale large multimodal models. *arXiv preprint arXiv:2405.11788*.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023b. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. 2024. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, et al. 2024. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. 2024. Docci: Descriptions of connected and contrasting images. *arXiv preprint arXiv:2404.19753*.

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.

Alexandre Piché, Aristides Milios, Dzmitry Bahdanau, and Chris Pal. 2024. Llms can learn self-restraint through iterative self-reflection. *Preprint*, arXiv:2405.13022.

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

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6914–6924.

Nimrod Shabtay, Felipe Maia Polo, Sivan Doveh, Wei Lin, M. Jehanzeb Mirza, Leshem Chosen, Mikhail Yurochkin, Yuekai Sun, Assaf Arbelle, Leonid Karlinsky, and Raja Giryes. 2025. Livexiv – a multi-modal live benchmark based on arxiv papers content. In *ICLR*.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. 2024. From pixels to prose: A large dataset of dense image captions. *arXiv preprint arXiv:2406.10328*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*.

Qiucheng Wu, Handong Zhao, Michael Saxon, Trung Bui, William Yang Wang, Yang Zhang, and Shiyu Chang. 2024. Vsp: Assessing the dual challenges of perception and reasoning in spatial planning tasks for vlms. *arXiv preprint arXiv:2407.01863*.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback. *Preprint*, arXiv:2403.18349.

Moran Yanuka, Morris Alper, Hadar Averbuch-Elor, and Raja Giryes. 2024. ICC : Quantifying image caption concreteness for multimodal dataset curation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11048–11064, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Jiarui Zhang, Jinyi Hu, Mahyar Khayatkhoei, Filip Ilievski, and Maosong Sun. 2024a. Exploring perceptual limitation of multimodal large language models. *arXiv preprint arXiv:2402.07384*.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. *arXiv preprint arXiv:2109.11503*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024b. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *Preprint*, arXiv:2402.09267.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024a. Tinyllava: A framework of small-scale large multimodal models. *Preprint*, arXiv:2402.14289.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024b. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A Appendix

In this appendix, we describe the implementation details in Appendix A.1, qualitative examples in Appendix A.2, additional human annotations details in Appendix A.3 and additional experiments in Appendix A.4.

A.1 Implementation Details

KnowAda Implementation Details For question generation, we use the `gemini-1.5-flash-001` model, following the prompt in Figure 14. To generate answers, we set the temperature to 0.4 for all models, sampling 10 answers per question. These answers are evaluated using the `gemini-1.5-flash-001` model, with prompts provided in Figure 15. For rewriting image descriptions, we employ `gemini-1.5-pro`, utilizing manually curated few-shot examples as illustrated in Figure 16. In all experiments, we use greedy sampling when generating outputs with Gemini.

Proposition Entailment Evaluation Implementation Details We use `gemini-1.5-flash-001` both for the proposition extraction, as well as the textual entailment task. We constrain the output to be in JSON format, following the proposition extraction prompt shown in Figure 17 and textual entailment prompt in Figure 18.

Training Details We provide the training hyperparameters in Table 4, all other hyperparameters are set to default. LLaVA and PaliGemma were trained using LLaMA-Factory framework (Zheng et al., 2024) and TinyLLaVA was trained using TinyLLaVA-Factory framework (Jia et al., 2024). We note that this work fully complies with the licenses of all used scientific artifacts (e.g. DOCCI, LLaVA, PaliGemma, etc.). All use of scientific artifacts is consistent with their intended use. All models were train on a single NVIDIA A6000 GPU.

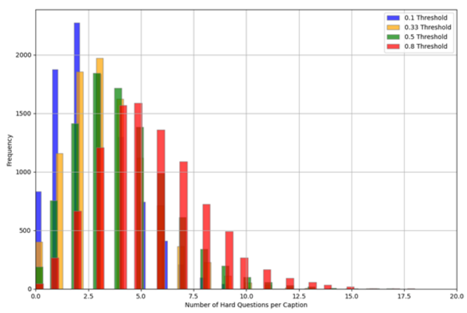

Figure 7: Distribution of number of unknown questions per caption for each ratio threshold of correct and incorrect answers from the vlm.

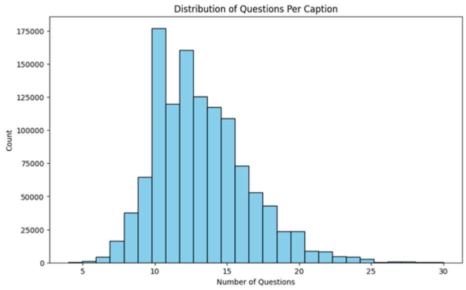

Figure 8: **Distribution of number of questions per caption** that are generated by the LLM based on the ground truth image description during the first stage of KnowAda.

A.2 Qualitative Examples

In Figure 11, we show example of our evaluation pipeline, including the original description, the generated description and the labels produce by the NLI model.

Figure 12 and Figure13 shows outputs generated by LLaVA-1.5-7B, trained on KnowAda-adapted captions using various thresholds. As the threshold T decreases, the model's tendency to hallucinate decreases. However, this comes at the cost of reduced detail coverage. The threshold T allows for fine-grained control over this trade-off between hallucination and detail retention.

A.3 Additional Annotation Details

To recruit high-quality annotators, we required them to have a HIT rate greater than 97% and at least 5,000 approved HITs, without imposing any geographical constraints. Additionally, we designed a qualification test consisting of 10 questions of varying difficulty, which we manually evaluated. A total of 14 annotators who passed the test with fewer than 2 mistakes were selected to perform the annotations for all experiments.

We present an example from the Amazon Mechanical Turk user interface, as shown to the annotators in Figure 10. Alongside the generated task guidelines, we provided supplementary slides with a detailed explanation of the visual entailment task, including 10 manually annotated examples that feature different labels and varying levels of difficulty for the annotators to review.

Annotators were paid 0.10$ per annotation, with an average hourly wage of 10$. The annotation process was approved by the institution's ethics committee.

Model	Epochs	BS	LoRA Rank	LR	Checkpoint
LLaVA-1.5-7B	3	8	64	0.0001	`llava-hf/llava-1.5-7b-hf`
TinyLLaVA	3	8	64	0.0001	`tinyllava/TinyLLaVA-Gemma-SigLIP-2.4B`
PaliGemma	15	8	128	0.0001	`google/paligemma-3b-mix-224`

Table 4: Training hyperparameters for the models. BS and LR refers to batch size and learning rate.

		Contradiction ↓		Descriptiveness ↑		
Model	**FT Captions**	**Precision**	**Recall**	**Precision**	**Recall**	**# Words**
PaliGemma	PixelProse	44.1	40.2	43.4	**38.4**	70.2
PaliGemma	PixelProse KA	**35**	**25.4**	**54.2**	38.2	47.7

Table 5: **Dense captioning results** using the automatic version of DNLI on the PixelProse (Singla et al., 2024) test set, when fine-tuning on original captions and KnowAda-adapted captions (denoted as KA) with a 20% threshold.

A.4 Additional Experiments

Larger-Scale Dense Caption Dataset We fine-tune PaliGemma on a dataset that is 10 times larger, sampling 100,000 image-caption pairs from the PixelProse (Singla et al., 2024) dataset. Of these, 95,000 pairs are used for training and 5,000 for testing. We follow the same configurations as in the main paper, except that we train for a single epoch, and report the results in Table 5. As shown, KnowAda significantly reduces the contradiction rate while maintaining high descriptiveness, demonstrating its effectiveness in generalizing to larger-scale datasets.

Relative Location of Contradicted Propositions. In Figure 9, we illustrate the distributions of the relative locations of propositions categorized as contradicted with respect to the generated caption. Consistent with prior research (McKenna et al., 2023) We observe that errors tend to increase with the distance from the beginning of the text: as more text is written, the rate of contradictions increases.

DNLI Evaluation with a Smaller Model To justify the use of Gemini in the question difficulty evaluation stage, we compared its performance against Gemma-2B by using both models as judges on 1,000 question-answer pairs from the DOCCI test set. The results showed a 75% agreement between the two models. To further analyze discrepancies, we manually inspected 20 random samples where the judgments differed, finding that in 18 out of 20 cases, Gemini provided the accurate assessment. These findings highlight a trade-off between model size and judgment accuracy. While Gemma-2B is a more lightweight alternative, we prioritized the superior accuracy and consistency offered by Gemini for our evaluations.

Additional KnowAda Statistics In Figure 8, we show the distribution of the number of questions generated per caption for the DOCCI training set. The data indicates that the distribution is centered around approximately 13 visual questions per caption.

In Figure 7, we demonstrate how the number of unknown questions changes when adjusting the threshold T. As T increases, more questions are classified as unknown, shifting the distribution.

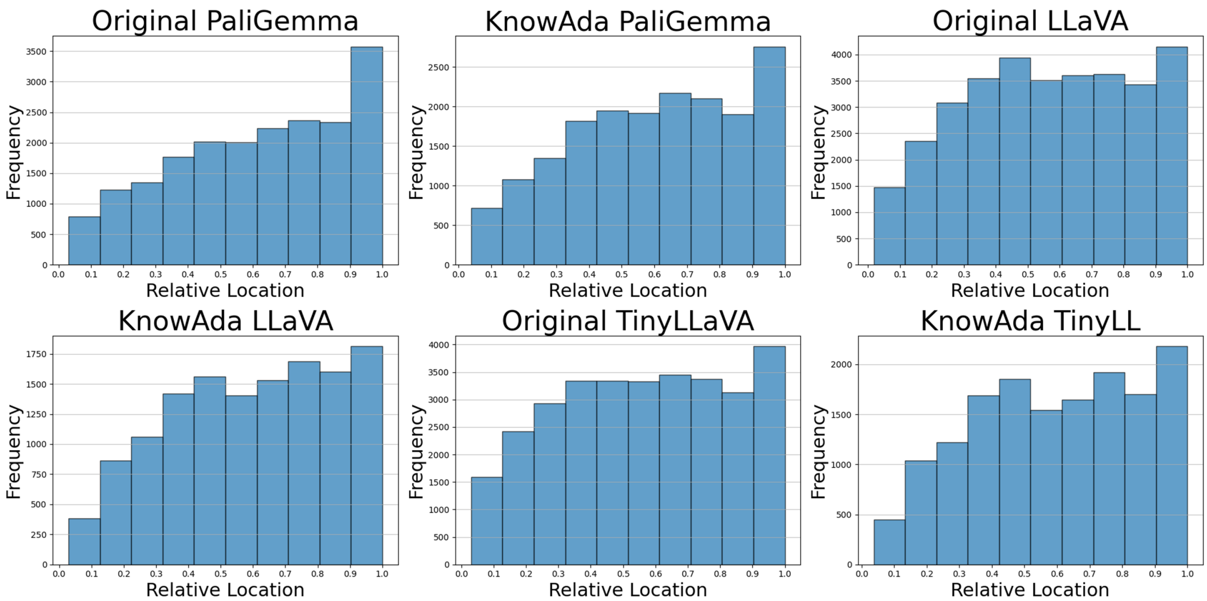

Figure 9: **Distribution of contradicted propositions' relative location** in generated captions across different models, tested on the DOCCI test set and measured using our proposition evaluation framework. Contradictions increase progressively as the captions become longer.

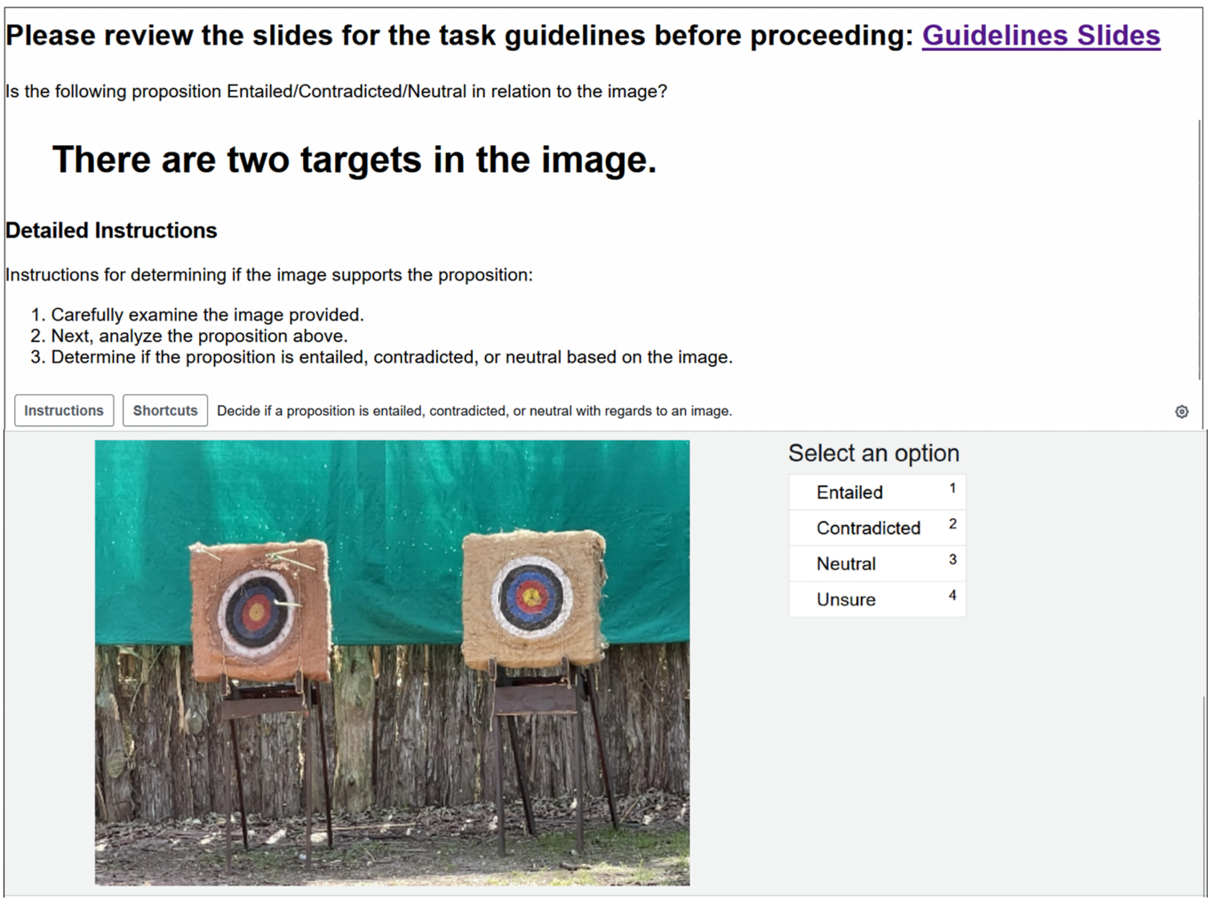

Figure 10: **Example of annotation interface in Amazon Mechanical Turk**

Image	Proposition	Judgement
	The image shows a roulette wheel.	Entailed
	The roulette wheel has a brown wooden frame.	Entailed
	The numbers on the roulette wheel are black.	Contradicted
	...	
	Numbers are on the outer edge of the wheel.	Entailed
	There is a black pole in the top right corner.	Neutral
	Black box is placed in a black and gray carpet.	Contradicted

Predicted Caption (LLaVA-1.5-7B)

A close up view of a roulette wheel with a brown wooden frame and a clear glass top. The numbers are in black and are placed on the outer edge of the wheel. The wheel is placed on a black metal platform. The platform is placed inside of a black box. The black box is placed in a black and gray carpet. The carpet is placed over a black platform. A black pole is seen in the top right corner of the view. The pole has a black base and a black top. A white light is seen shining on the carpet and the black platform in the background.

Original Caption

A roulette wheel is centered in view; the wheel has a glass dome covering the top, which is reflecting the lights above, and the outer border of the wheel is darkly polished wood. Inside of the roulette wheel is a circular stretch of the roulette numbers, which are white, and a checkered pattern of red and black beneath them; three of the squares on the right side are a muted green and have the numbers "0","000", and "00". Inside of the numbers is another circular border that is silver, and inside of it is the same dark polished wood as the outside. Surrounding the wheel is a black table, as well as other gambling games in a casino. A wall of plexiglass surrounds the right and bottom of the wheel, leaving the left and top sides open. The scene is well lit, with parts having a slightly purple hue.

Figure 11: **Qualitative Example of Our Proposition-Based Evaluation Metric**. Given a generated description by a VLM (middle), we decompose it into atomic propositions (top, center part). Then, we classify each proposition as either ’Entailed’, ’Contradicted’ or ’Neutral’ (top, right part), conditioned on the ground-truth image description (bottom). Finally, we calculate the consistency and contradiction based on the number of entailed and contradicted propositions.

Original:

A medium-close-up view of a dark brown wooden dresser that is sitting on a light brown tile floor. The dresser has a glass vase sitting on top of it, and the vase is filled with water. The vase has a clear base and a clear neck, and it is filled halfway with water, and on the neck of the vine there is a small crystal bowl. The light bulb is sitting inside of a white glass lamp shade. The lamp is sitting at the top of the dresser. On the right side of the lamp, there is an oval mirror that is reflecting the lamp and the dressers. The mirror is being held up by a silver metal pole. The walls surrounding the dressing room are made up of wooden planks that run vertically.

T=90%:

A medium-close-up view of a wooden dresser that is placed on a white tiled floor. The dresser is made up of dark brown wooden panels that run horizontally, and along the panels there are small drawers that have gold handles. The top portion of the dresser has a small lamp that is made of glass, and the glass is being lit up by a light bulb. On the left side of the lamp, there is a small vase that is being held up by two thin sticks. On top of the vase, there are two small glasses that are being held together by a thin metal rod. On both sides of the mirror, there's a wooden wall that is covered in wooden planks that run vertically.

T=80%:

A view of a wooden desk with a lamp on it. The desk is dark brown and has a drawer on the left and one on the right. The drawer in the middle has a small black hole in it. A small black box is on the floor in front of the desk. The lamp is on top of the drawer. It has a white shade on it and a glass base. The base has a reflection of the lamp on the inside of it. On the left of the frame is a white outlet with a cord plugged into it.

T=60%:

A view of a wooden desk with a lamp on it. The desk has drawers on it and a small shelf under it. A lamp is on the desk. The lamp has a white shade on it, and it has a glass base. The base has a clear glass base with a small round base on it that has a small white light bulb in it. There is a small glass vase on the left of the lamp. The vase has a round base and a clear base. A mirror is on a wall behind the desks. The mirror has a frame around it.

T=40%:

A view of a wooden desk with a lamp on it. The desk has drawers on it and a small shelf under it. A mirror is on the wall above the desk. The mirror has a frame around it. Light is shining on the desks and the wall. The floor is tiled

T=20%:

A view of a brown wooden desk with a lamp on it. The desk has drawers on it and a small shelf under it. A mirror is on the wall above the desk. The mirror has a frame around it. It is hanging on the wood wall. The wall is made of wood planks. The floor is tiled.

Figure 12: **Qualitative Example of KnowAda's Threshold T effect**. We apply KnowAda's rephrasing to the original caption (bottom) using different T values.

Original:
A medium-close-up view of a statue that is placed on a gray cement platform. The statue is of a woman who is sitting down and is facing right. The woman is wearing a dress that is made up of a golden material, and along the dress there are golden markings. The dress is being held up by a golden belt that is wrapped around the woman's waist. The left arm of the woman is placed along the right side of the statue, and the right arm is placed behind the woman. The right hand of the arm is holding onto the left arm. The head of the sculpture is facing slightly to the right, and it is a bust of a man who has short hair and a beard. The man is weeping, and his right hand is placed against his chest. The bust is being lit up by the sun. Behind the statue there is a gray building that is covered in shade.

T=90%:
A medium-close-up view of a statue that is placed on a gray cement platform. The statue is of a woman who is sitting on a chair, and she is facing forward. The woman is wearing a dress that is made up of gold and silver. The dress is made of a long skirt, and along the skirt there are two gold lines that run vertically. The skirt is made out of a gold and white dress, and it is made to look like a bell. The top portion of the dress is white, and on top of the white portion there is a gold one. The head of the woman is made like a bust, and the neck of the bust is made from gold. The arms of the statue are thin, and they are made of gold. On the left arm of the lady, there is something that is being held. The face of the person is made with a lot of detail, and surrounding the face there is gold. To the right of the chair, there are statues of two people who are sitting on their knees. The person on the left is a man, and he is facing to the right. The man is weating a white dress that has gold trimming. The other person is a woman, and her face is made in a similar way to the woman's. The legs of the people are thin and made of white marble. The window of the building is made into a rectangle, and running along the window there are thin lines that are made vertically and horizontally

T=80%:
A medium-close-up view of a statue that is placed on a gray cement platform. The statue is of a woman who is sitting on a pedestal. The woman is facing forward, and she is wearing a dress that is made up of white fabric. The dress is being held up by a gold belt that is wrapped around the woman's waist. The head of the woman is being supported by a golden crown that is being worn by the woman. The crown is being placed on the woman by a woman' figure that is sitting behind the woman and is facing the right side. The figure is weary, and it is being lit up by the sun. The right arm of the figure is being wrapped around a woman, and the left arm is being used to hold up the crown. The face of the statue is being covered by a veil. The cement pedestals are being held together by a metal pole that runs vertically.

T=60%:
A view of a statue of a woman and a man. The woman is sitting on a platform, and she is facing forward. She has a crown on her head. She is wearing a dress. The man is sitting in front of her, and he is facing her. He has no clothes on. He is holding his right arm up. The statue is on a gray stone platform. The sun is shining on it. A building is behind the statue

T=40%
A view of a statue of a woman sitting on a pedestal. The woman is facing forward and has her right arm around the boy. The boy is facing away from the woman and has his right arm wrapped around her. The statue is placed on a gray cement pedestals. The pedestall is placed in front of a building.

T=20%
A view of a statue of a woman and a man. The woman is sitting on a pedestal. The man is sitting in front of her. The statue is on a gray stone pedestals. The pedestall is on top of a gray cement base. The base is on the ground. The ground is covered in shadows.

Figure 13: **Additional Qualitative Example of KnowAda's Threshold T effect**.

SYSTEM PROMPT:

Task Overview:
You are an AI visual assistant observing a single image. Accompanying this image is a paragraph that describes it.

Guidelines:
Formulate Relevant Questions: Your primary task is to generate multiple questions as if you are actively viewing the image.

Adopt an AI Tone: Frame your questions in a tone consistent with that of a visual AI assistant, focusing on analyzing the visual content.

Focus on Visual Details: Your questions should pertain to specific aspects of the image, such as:
Types of objects present, Counting the objects, Actions performed by the objects, Locations and relative positions of objects, Background knowledge related to the objects

Ensure Definitive Answers: Only ask questions that can be confidently answered based on the visible content of the image. Each question should meet the following criteria:
- It can be confidently answered from what is seen in the image, as described in the caption.
- It is clear that the answer can be determined as absent if the information is not present.

Avoid Uncertain Details: Do not include questions that cannot be answered confidently. Focus on complex, relevant inquiries that delve into the visual elements and contextual significance within the image.

Response Structure:

Please structure your response as follows:

Questions:
<Your first question>
<Your second question>
<Your third question>
...

FEW-SHOT PROMPT:

An indoor upwards angled view of a wooden beam light fixture with six incandescent light bulbs hanging down from the frayed edges of a rope wrapped around the wooden beam. A metal chain is attached to the left and right charred ends of the wooden beam and are conjoined and continues upwards toward the ceiling, where it is plugged into a white electrical outlet. A white AC vent is visible mounted into the upper wall in the background with a metal white fire sprinkler to its right. A white barn door with a black metal railing is visible below and behind the light bulbs, with a black menu board and white letters that read,"$3500/ $500/ $2500" to its left. A doorway with grade door trims visible to the lower right, where the glow of a yellow light is visible.

1. Question: What is the perspective of the image?
2. Question: What material is the light fixture made from?
3. Question: How many light bulbs are hanging from the fixture?
4. Question: What is the condition of the rope holding the light fixture?
5. Question: What material is the chain attached to the light fixture made from?
6. Question: What is the color of the chain attached to the light fixture?
7. Question: What is the condition of the ends of the wooden beam?
8. Question: What is mounted on the upper wall in the background of the image?
9. Question: What color is the fire sprinkler next to the AC vent?
10. Question: What type of door is visible below the light fixture?
11. Question: What color is the railing next to the barn door?
12. Question: What is the material of the railing next to the barn door?
13. Question: What is located to the left of the barn door?
14. Question: What is color of the board?
15. Question: What color are the letters on the menu board?
16. Question: What is written on the menu baord?
17. Question: What is visible through the doorway in the lower right corner of the image?
18. Question: What color is the light glowing from the doorway?

A top-down, close-up view of a brown, fiery skipper sitting on top of a leaf that is slightly wide. On the far left side of the fiery skipper, there is a discolored flower that used to be blue, while the one underneath it is a dark blue. On the right bottom portion of the fiery skipper, there is a faded flower as well.

1. Question: What is the overall perspective of the scene?
2. Question: What is the distance and zoom level of the viewpoint?
3. Question: What type of insect is seen on top of the leaf?
4. Question: What is the primary color of the insect in the scene?
5. Question: How wide is the leaf that the insect is sitting on compared to its body?
6. Question: Is there a flower to the left to the insect?
7. Question: What is the condition of the flower to the left bottom of the insect?
8. Question: Is there a flower to the left bottom of the insect?
9. Question: What is the condition of the flower to the right bottom of the insect?
10. Question: What is the color of the flower that is underneath the insect?
11. Question: Is there a flower to the right bottom of the insect?
12. Question: What is the condition of the flower to the left of the insect?

A distorted and blurry outdoor image of a red colored octagon shaped sign in front of various leaves and trees with the clouded gray sky visible at the top of the view. The sign says "STOP" in white, with a white colored border around the outside of the sign.

1. Question: What is the overall clarity of the image?
2. Question: Is the image taken indoors or outdoors?
3. Question: What is the basic shape of the sign in the image?
4. Question: What color is the sign in the image?
5. Question: Is there any text on the sign?
6. Question: what color is the text on the sign?
7. Question: Is there a border around the sign?
8. Question: If there is a border around the sign, what color is the border?
9. Question: What is visible in the background behind the sign?
10. Question: What is the color of the sky visible in the image?

Figure 14: **Question Generation Prompt**.

PROMPT:
Your task is to evaluate how accurately a vision-language model answers a given question, using a scale of 1 to 3.

3: The model's answer is fully correct.
2: The answer is somewhat correct but has minor discrepancies.
1: The answer is incorrect or contradicts the image description and ground truth.

You will be provided with the image description, the model's answer, and the ground truth answer.

When scoring, consider how much the model's answer deviates from the ground truth and the image description. The more the answer contradicts these, the lower the score should be.

For example:

If the ground truth is "square" and the model answers "rectangle," it should receive a 2, as the answer is close but not exact.
If the model answers "circle," which contradicts the ground truth, it should receive a 1.

Similarly, if the ground truth is "brown" and the model answers "tan," the answer should receive a 2, but if the answer is "green," it should receive a 1.

Image description:
===
<description>
===

Question:
===
<question>
===

Vision-language model answer:
===
<model-answer>
===

Please provide a numerical score (1-3) for the model's answer. Do not include any additional explanations or comments.

Figure 15: **Evaluate VLM Answer Prompt**.

SYSTEM PROMPT:

Task Overview:
You are provided with an image description and a set of questions related to that image. Your objective is to modify the image description so that it no longer includes information that can be used to answer any of the questions.

Guidelines:
1. Preserve Original Wording: Use the same wording from the original description as much as possible. Only make changes that pertain directly to the questions provided.
2. Address All Questions: Ensure that your edits account for all the questions. No question should be overlooked in your modifications.
3. Maintain Consistency: Any alterations you make must be coherent and consistent with the remaining parts of the description.
4. Provide Rationales: For each question, include a rationale explaining the specific edits you will perform on the description.

Response Format:
Your responses should adhere to the following structure:

Rationale:
<Your rationale for which details should be removed based on each question.>

New Description:
<The edited description reflecting your rationale.>

FEW-SHOT PROMPT:

Description:
An indoor upwards angled view of a wooden beam light fixture with six incandescent light bulbs hanging down from the frayed edges of a rope wrapped around the wooden beam. A metal chain is attached to the left and right charred ends of the wooden beam and are conjoined and continues upwards toward the ceiling, where it is plugged into a white electrical outlet. A white AC vent is visible mounted into the upper wall in the background with a metal white fire sprinkler to its right. A white barn door with a black metal railing is visible below and behind the light bulbs, with a black menu board and white letters that read,"$3500/ $500/ $2500" to its left. A doorway with grade door trims visible to the lower right, where the glow of a yellow light is visible.

Questions:
What is the perspective of the image?
How many light bulbs are hanging from the fixture?
What is the condition of the rope holding the light fixture?
What is written on the menu baord?

Rational:
The angled view should not be mentioned because it answers the first question.
The amount of light bulbs (six) should not be mentioned because it answers the second question.
The frayed edges of the rope should not be mentioned because it answers the third question.
The text written on the menu ("$3500/ $500/ $2500") should not be mentioned because it answers the fourth question.

New Description:
An indoor view of a wooden beam light fixture with incandescent light bulbs hanging down from a rope wrapped around the wooden beam. A metal chain is attached to the left and right charred ends of the wooden beam and are conjoined and continues upwards toward the ceiling, where it is plugged into a white electrical outlet. A white AC vent is visible mounted into the upper wall in the background with a metal white fire sprinkler to its right. A white barn door with a black metal railing is visible below and behind the light bulbs, with a black menu board and white letters to its left. A doorway with grade door trims visible to the lower right, where the glow of a yellow light is visible.

Description:
A top-down, close-up view of a brown, fiery skipper sitting on top of a leaf that is slightly wide. On the far left side of the fiery skipper, there is a discolored flower that used to be blue, while the one underneath it is a dark blue. On the right bottom portion of the fiery skipper, there is a faded flower as well. A dog is playing with a ball in the distance.

Questions:
What type of insect is seen on top of the leaf?
What is the primary color of the insect in the scene?
How wide is the leaf that the insect is sitting on compared to its body?
What is the condition of the flower to the left bottom of the insect?
Is there a flower to the right bottom of the insect?
What is the dog playing with?

Rational:
The type of insect (fiery skipper) should be removed because it answers the first question.
The primary color of the insect (brown) should not be mentioned because it answers the second question
How wide the leaf is open (slightly wide) should not be mentioned because it answers the third question.
The condition of the flower to the left bottom of the insect (discolored) should not be mentioned because it answers the forth question.
The flower to the right bottom of the insect should not be mentioned because it answers the fifth question.
The ball the dog is playing with should not be mentioned because it answers the sixth question.

New Description:
A top-down, close-up view of an insect sitting on top of a leaf. On the far left side of the insect, there is a flower, and another underneath it is in a dark blue color. On the right bottom portion of the insect, there is a faded flower. A dog is playing with something in the distance.

Description:
A distorted and blurry outdoor image of a red colored octagon shaped sign in front of various leaves and trees with the clouded gray sky visible at the top of the view. The sign says "STOP" in white, with a white colored border around the outside of the sign.

Questions:
What is the basic shape of the sign in the image?
Is there any text on the sign?
what color is the text on the sign?

Rational:
The octagon shape of the sign should not be mentioned because it answer the first question.
The text and the color of the text on the sign ("STOP" in white) should not be mentioned because it answer the second and third questions.

New Description:
A distorted and blurry outdoor image of a red colored sign in front of various leaves and trees with the clouded gray sky visible at the top of the view. There is something written on the sign and there is a white colored border around the outside of the sign.

Description:
A black Labradoodle is walking on a concrete road with a curly tail. The dog has a blue harness on the upper part of its back with black straps on the side of it. A small metal clasp is on the top of the vest. Metal links are going across wooden planks to form a fence. Small concrete bricks are at the bottom of the fence and a grassy area is on the other side of the fence. Leaves and dirt are in front of the concrete bricks. Sunlight is shining on the grassy area behind the fence.

Questions:
What color are the straps on the dog's harness?
What is on the top of the harness?
What kind of fence is in the image?
What is at the bottom of the fence?

Rational:
The black color of the straps should not be mentioned because it answer the first question.
The small metal clasp is on the top of the vest should not be mentioned because it answer the second question.
The metal links that are going across wooden planks to form a fence should not be mentioned because it answer the third question.
The bricks at the bottom of the fence should be edited because it answers the forth question.

New Description:
A black Labradoodle is walking on a concrete road with a curly tail. The dog has a blue harness on the upper part of its back with straps on the side of it. There are some small concrete bricks. There is a grassy area is on the other side of the fence. Leaves and dirt are in front of the concrete bricks. Sunlight is shining on the grassy area behind the fence.

Figure 16: **Rewriting Prompt**.

PROMPT:
Decompose the given image caption into clear and simple propositions, ensuring they are interpretable out of context.

Follow these guidelines:
1. Split compound sentences into simple sentences.
2. For any named entity that is accompanied by additional descriptive information, separate this information into its own distinct proposition.
3. Decontextualize the propositions by:
- Adding necessary modifiers to nouns or entire sentences to clarify context.
- Replacing pronouns (e.g., "it", "he", "she", "they", "this", "that") with the full name of the entity they refer to, ensuring no references rely on prior information.
4. If there are any contradicting propositions, present both as separate propositions.
5. Present the results in JSON format with the following structure:
- `"propositions"`: an array of objects, each containing:
- `"id"`: a unique number for each proposition, starting from 1.
- `"proposition"`: the decomposed, decontextualized proposition as a string.
Only use the json format, without trailing \n, ```, or the word JSON etc. Make sure you use signle " quotes and not double "" in the json output when representing strings.
It is extremely important to have the right JSOM format, otherwise the evaluation will fail, as is shown in the example below.

Example:

Input:
The image shows a concrete sidewalk. A diagonally oriented rectangular section of the sidewalk is textured with 17 parallel lines. To the right of the textured section, a red stripe runs parallel to the edge of the picture. The words "FIRE LANE" are inscribed within the stripe, with the top of the letters oriented toward the top right of the image. To the left of the textured section, the word "ROW" is written in orange. Above the "ROW", an orange line, parallel to the top edge of the textured section, bisects a small orange circle.

Output:
```
{
"propositions": [
{ "id": 1, "proposition": "There is a sidewalk in the image." },
{ "id": 2, "proposition": "The sidewalk is made of concrete." },
{ "id": 3, "proposition": "There is a rectangular section of the sidewalk." },
{ "id": 4, "proposition": "The rectangular section is oriented diagonally." },
{ "id": 5, "proposition": "The rectangular section is textured with lines." },
{ "id": 6, "proposition": "There are 17 parallel lines on the rectangular section." },
{ "id": 7, "proposition": "There is a red stripe in the image." },
{ "id": 8, "proposition": "The red stripe is to the right of the textured section." },
{ "id": 9, "proposition": "The red stripe runs parallel to the edge of the picture." },
{ "id": 10, "proposition": "The words 'FIRE LANE' are written." },
{ "id": 11, "proposition": "The words 'FIRE LANE' are inscribed within the stripe." },
{ "id": 12, "proposition": "The top of the letters of the words 'FIRE LANE' is oriented toward the top right of the image." },
{ "id": 13, "proposition": "There is text to the left of the textured section." },
{ "id": 14, "proposition": "The word 'ROW' is written." },
{ "id": 15, "proposition": "The word 'ROW' is written in orange." },
{ "id": 16, "proposition": "There is an orange line above the word 'ROW'." },
{ "id": 17, "proposition": "The orange line is parallel to the top edge of the textured section." },
{ "id": 18, "proposition": "There is a small orange circle in the image." },
{ "id": 19, "proposition": "The orange line bisects the small orange circle." }
]
}
```

Figure 17: **Proposition Extraction Prompt**.

PROMPT:
You are given a ground truth image description and a list of propositions. Your task is to analyze each proposition and check whether it is entailed by the ground truth image description. A proposition is considered entailed if all the information in it can be inferred from the ground truth description. If the proposition introduces new information that is not present in the ground truth or contradicts it, it is not entailed.

Additional Criteria:
1. If the proposition contains neutral information (subjective information, such as describing the environment as "lively" or "pleasant"), it should not be counted as either entailed or contradicted. Instead, it should be judged as "Neutral."
2. If the proposition introduces additional visual information that is not in the ground truth, it should be judged as "Contradicted."

For each proposition, respond with the proposition number and its corresponding judgment as either "Entailed," "Contradicted," or "Neutral." Your output should maintain the same number of propositions as the input list.

Provide the results in the following JSON format:
- `"propositions"`: an array of objects, each containing:
- `"id"`: the number of the proposition.
- `"judgment"`: the result for each proposition as "Entailed," "Contradicted," or "Neutral."
- `"summary"`: an object containing:
- `"contradicting_count"`: the number of contradicting propositions.
- `"entailed_count"`: the number of entailed propositions.
- `"neutral_count"`: the number of neutral propositions.

If the list of propositions is empty, provide an empty json as the output, with the same format.

Example output:

```
{
"propositions": [
{ "id": 1, "judgment": "Entailed" },
{ "id": 2, "judgment": "Contradicted" },
{ "id": 3, "judgment": "Neutral" },
{ "id": 4, "judgment": "Entailed" }
],
"summary": {
"contradicting_count": 1,
"entailed_count": 2,
"neutral_count": 1
}
}
```

Figure 18: **Proposition Judgement Prompt**.