Do More Details Always Introduce More Hallucinations in LVLM-based Image Captioning?

Anonymous ACL submission

Abstract

Generating descriptions using large vision-language models (LVLMs) often involves hallucinations, in which the output text misrepresents the visual information in the input image. While previous studies attribute the occurrence of hallucinations to the inclusion of more details, our study finds technical flaws in existing metrics, leading to unreliable evaluations of models and conclusions about hallucinations. This has sparked a debate on the question: *Do more details always introduce more hallucinations in LVLM-based image captioning?*

In this paper, we address this debate by proposing a novel decoding strategy, Differentiated Beam Decoding (DBD), along with a reliable new set of evaluation metrics: CLIP-Precision, CLIP-Recall, and CLIP-F1. DBD decodes the wealth of information hidden in visual input into distinct language representations called unit facts in parallel. This decoding is achieved via a well-designed differential score that guides the parallel search and candidate screening. The selected unit facts are then aggregated to generate the final caption. Our proposed metrics evaluate the comprehensiveness and accuracy of image captions by comparing the embedding groups of ground-truth image regions and generated text partitions. Extensive experiments on the Visual Genome dataset validate the effectiveness of our approach, demonstrating that it produces detailed descriptions while maintaining low hallucination levels.

1 Introduction

Large Vision-Language Models (LVLMs) (Liu et al., 2023b; Long et al., 2022; Zhu et al., 2023; OpenAI, 2024) have been broadly employed in a range of multimodal applications, such as image captioning (Li et al., 2022; Liu et al., 2023d; Chen et al., 2023b; Wang et al., 2023; Xuan et al., 2023), where the objective is to generate textual descriptions that accurately encapsulate visual data. On the one hand, LVLMs are expected to generate descriptions that are detailed and comprehensive rather than overly succinct, ensuring pivotal visual information is not omitted. On the other hand, detailed and long descriptions from these LVLMs often suffer from a significant challenge known as *hallucination* (Rohrbach et al., 2018), where output text is semantically coherent but misaligned with the actual objects present in the input image. In this work, we mainly consider *object hallucination*. Such issues, whether of omission or hallucination, limit practical applications in safety-critical fields such as medical imaging (Hu et al., 2023; Alsharid et al., 2022, 2021) or autonomous driving (Chen et al., 2023c; Jin et al., 2023).

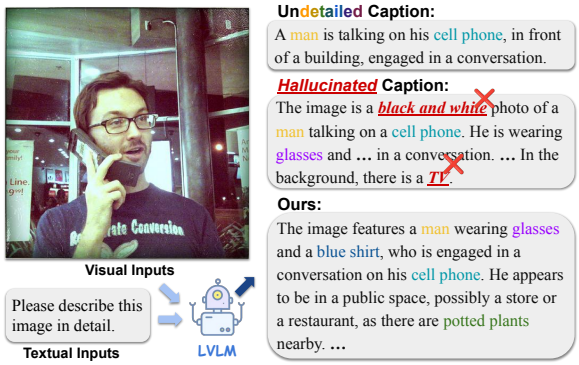

Figure 1: Comparison of captions. The top caption ignores much visual information. The middle one includes details but also introduces hallucinations. Our method (bottom) provides detailed and accurate captions.

As shown in Figure 1, a short caption accurately describes the main objects, *man* and *cell phone*, in the center of the image but overlooks details such as the man's *glasses* and the *potted plant* in the corner. However, when we encourage LVLMs to generate a longer caption to include more image details, we can see the occurrence of hallucinations, such as a nonexistent object, *TV*, and inaccurate attributes, *black and white*. A question naturally arises: ***Do more details always introduce more hallucinations in LVLM-based Image Captioning?***

There remains a debate about answering this question. On the one hand, substantial evidence (Zhou et al., 2023; Deng et al., 2024) indicates that hallucinations are more likely to appear in later sentences of long text sequences where models tend to describe details. Moreover, the native design of the model results in the subsequent description of the details receiving more interference than the previous text, which is misleading from the previous text itself. On the other hand, our in-depth manual analysis reveals that existing metrics overstate the presence of hallucinations in detailed descriptions. This inaccuracy stems from two main factors: first, human-generated ground-truth captions often lack detail; second, rule-based evaluations fail to grasp the true intent of image captioning, sometimes even contradicting it. The more detailed the description, the more serious the impact of these two factors, rendering the evidence less credible.

In this work, we aim to ***settle down this debate by demonstrating that, with proper guidance, LVLMs can produce detailed descriptions while maintaining low hallucination levels.*** To achieve this goal, we propose a novel decoding strategy, Differentiated Beam Decoding (DBD), along with a new set of evaluation metrics: CLIP-Precision, CLIP-Recall, and CLIP-F1. Our proposed DBD is grounded in parallel decoding to eliminate misleading information from pre-generated texts and differentiated searching to compel the LVLM to describe the image from various perspectives. It operates in three phases: differentiated parallel search, post-search selection, and unit facts summarization. Our introduced metrics utilize different divisions of the image and the caption to obtain a group of embeddings, respectively. The percentage that one group of embeddings can express the other is calculated, serving as CLIP-precision and CLIP-recall, which are then combined to produce CLIP-F1.

Our contributions are fourfold:

- We introduce a novel, training-free decoding method that generates in parallel multiple facts of distinct aspects of an image and synthesizes them into a comprehensive and accurate description.
- We conduct an in-depth analysis of widely used hallucination metrics, revealing that over 50% of identified object hallucinations are unjust.
- We propose a novel set of metrics that specifically and separately evaluate the comprehensiveness and hallucination level of image captioning.
- We perform extensive experiments to validate the effectiveness of our approach. We tested various LVLMs on the Visual Genome dataset using the proposed metric and human evaluation.

2 Related Work

2.1 LVLM-based Image Captioning

Image captioning (Bai and An, 2018) aims to bridge the gap between visual perception and language understanding. Compared with traditional models (Vinyals et al., 2015; Wang et al., 2020), by training on extremely large-scale datasets, Large Vision-Language Models (LVLMs) can give us more detailed and contextually rich descriptions to improve the performance of image understanding. Representative works include LLaVA-1.5 (Liu et al., 2023d), MiniGPT-4 (Zhu et al., 2023), and mPLUG-Owl2 (Ye et al., 2023b).

However, the use of LVLMs in the image captioning task sometimes comes with hallucination (Gunjal et al., 2024), which is an issue in that LVLMs generate descriptions containing objects or details not present in the image. Existing methods to mitigate hallucinations include reinforced grounding (Favero et al., 2024), adversarial training (Park et al., 2024), and visual verification steps (Liu et al., 2023a; Lu et al., 2023).

2.2 Evaluation Metrics of Image Captioning

Three metrics are widely used to evaluate the performance of the image captioning task, including Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al., 2018), Consensus-based Image Description Evaluation (CIDEr) (Vedantam et al., 2015), and the CLIP Score (CLIP-S) (Hessel et al., 2022).

CHAIR evaluates object hallucination in image captioning, consisting of (1) CHAIR_S, measuring the proportion of sentences with hallucinated objects, and (2) CHAIR_I, assessing the proportion of hallucinated objects among all mentioned objects. CIDEr evaluates the match between generated captions and reference captions by measuring the consensus across multiple human-annotated reference captions. The reference-free CLIP-S is also used to evaluate image captioning. However, each of these metrics has limitations. CLIP-S lacks detail, which is crucial for high-quality image captioning. CHAIR and CIDEr metrics are often limited by humans' undetailed ground-truth caption annotations. These challenges highlight the need for more robust and detailed evaluation metrics in captioning.

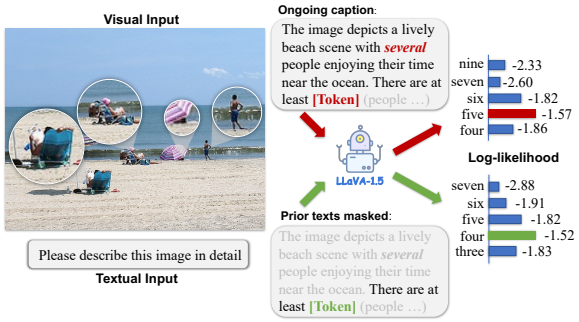

Figure 2: Misleading from pre-generated texts. With prior output engaged, LLaVA-1.5 predicts most likely *five* people in the scene. However, with the first sentence masked during the coding, the number decreases to *four*.

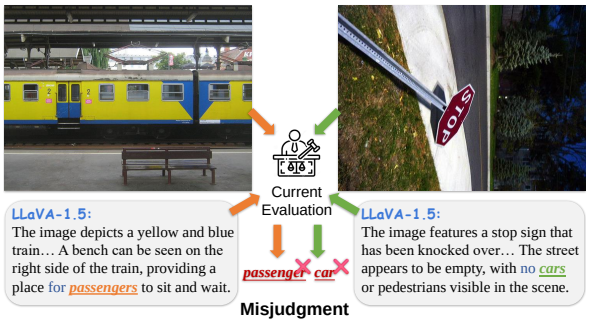

Figure 3: Misjudgement from rule-based identification. The current framework misunderstands elaborate captions, where many object words appear not for existence but for decorative features or even for non-existence.

3 Detail-Hallucination Debate

3.1 Co-occurrence of Detail and Hallucination

The arguments regarding the co-occurrence of detail and hallucination fall into two main points.

First, substantial experimental evidence from prior research indicates that hallucinations tend to appear more frequently in the later parts of captions. For instance, Figure 2a. in Deng et al. (2024) shows a consistent pattern across multiple LVLMs, where the probability of hallucination increases as the sentence progresses. Additionally, Figure 1c. in Zhou et al. (2023) highlights that the end of descriptions is a predominant high-density area for hallucinatory objects. Although this observation is intuitive, our further experiment of LVLM-based captioning demonstrates that models tend to describe details in the later part of the caption. The results and settings for this experiment and more discussions are provided in the Appendix A.1.

Second, the inherent attention mechanism of looking back causes the subsequent detailed description to receive more interference. In a typical sequential generation, it is reasonably expected that earlier text influences later text, and a rational consensus is that this influence should only affect the direction of the narrative, not the critical factual content. However, in vision-language tasks such as image captioning, where complex information is hidden in the visual embedding rather than explicitly expressed in language, the previous descriptions can interfere with the generation of subsequent details. For example, as illustrated in Figure 2, in the original caption, LLaVA-1.5 initially identifies *five* people in the scene. However, when we mask the pre-generated texts during inference, the number decreases to *four*, indicating the misleading from the prior words such as *"several."*

3.2 Inflated Hallucination Presence

Nevertheless, our in-depth analysis of the classic hallucination evaluation method, CHAIR, reveals that the presence rate of hallucination is inflated. Specifically, we generated captions for 500 randomly selected images from the MSCOCO validation dataset using LLaVA-1.5 (Liu et al., 2023d) and evaluated them using the CHAIR (Rohrbach et al., 2018) framework. Upon manually inspecting the samples identified as hallucinated, we found that over 55% of identified hallucinations are unjustified. The misjudgment is due to two main reasons, and both are exacerbated in detailed captions.

First, the rule-based object extraction is rigid and prone to errors, leading to a misunderstanding of the caption. The current rule assumes that if an object word appears in the caption, the description expresses the presence of the object in the image. However, detailed captions often include decorative text that contains object words without indicating their presence. For instance, as shown in Figure 3 (left), the object word, *passenger*, is used to describe a characteristic of the seat, for passengers to sit and wait, not to indicate the presence of a passenger. Similarly, detailed captions might explicitly state the absence of certain *objects* related to the scene, such as *car* in Figure 3 (right), yet are still labeled as hallucinations. Other inaccurate extractions include identifying *orange* from *orange juice*, *bowl* from *toilet bowl*, and so on.

Second, the language-reference-required comparisons are restricted by the quality of annotations. Any ground-truth caption is essentially a lossy compression of the original visual information, which falls short when more detailed predictions arise. This phenomenon is inevitable, akin to our perpetual quest for additional decimal places in π. This issue is particularly prevalent in the current evalu-

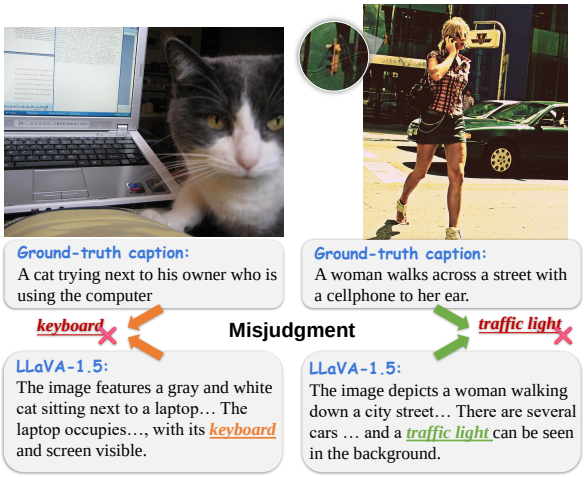

Figure 4: Misjudgement from undetailed ground truths. The current framework identifies hallucination based on language ground truths, restricting the capability to evaluate captions in more detail than ground truths.

ation framework that widely uses MSCOCO. For example, as presented in Figure 4, ground-truth captions from humans often omit subassemblies, like *keyboard* of a laptop or small details such as *traffic light*. Notably, ground truths shown in Figure 4 are obtained by summarizing all human captions provided by MSCOCO. Although the CHAIR framework uses both caption and object annotations, ground truths sometimes still lack detail. We provide more examples of failure in Appendix A.1.

Intuitively, both reasons lead to increased false positives in the descriptions of details, refuting the experimental evidence from the co-occurrence side. Moreover, direct evidence suggests that there are more false positives in later parts of the captions. Specifically, we calculate the relative positions where false positives appear in the caption and find that only 8.6% of false positives appear in the top 25% of the caption, 15.2% in 25-50%, 19.6% in 50-75%, and 56.5% in the bottom 25%.

3.3 Statement of Purpose

While the validity of experimental evidence that highlights the co-occurrence of detail and hallucination using current evaluation metrics has been challenged, the inherent problem posed by the attention mechanism's design persists. To mitigate the issue of the extra interference from pre-generated text and allow for high-fidelity and detailed image captioning, we introduce a novel decoding method, Differentiated Beam Decoding (DBD).

Despite a new method to reduce interference issues, the current metrics still have significant limitations, as emphasized by the opposing viewpoint. To provide a more comprehensive and reliable evaluation of both the thoroughness of detail and the accuracy of content, we propose a novel set of metrics, CLIP-Recall, -Precision, and -F1.

4 Methodology

4.1 Preliminary

LVLMs. We consider a Large Vision-Language Model (LVLM), denoted as $\mathcal{M}$, which receives an image v and a text prompt x, including the pre-generated tokens, to infer the next token probability. This process is formalized as follows:

$$\mathcal{M}(\boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}_{<t}) := p_{\mathcal{M}}(\cdot \mid \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}_{<t}).$$

Given each token's probability from the LVLM, the sentence-level log-likelihood is defined as:

$$\log p_{\mathcal{M}}(\boldsymbol{y}_{\leq t} \mid \boldsymbol{v}, \boldsymbol{x}) = \sum_i \log p_{\mathcal{M}}(\boldsymbol{y}_i \mid \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}_{<i}).$$

Decoding strategy. Decoding strategies dictate how to harness the next token probability to select the next token y_t from the set S of all possible tokens in the vocabulary, which includes special tokens such as <BOS>, <EOS>, and <PAD>. Greedy decoding, the simplest strategy, selects the next token with the highest probability, formalized as

$$y_t = \arg\max_{s \in S} \log p_{\mathcal{M}}(s \mid \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}_{<t}).$$

4.2 Differentiated Beam Decoding

As depicted in Figure 5, our method, Differentiated Beam Decoding (DBD), comprises three phases. First, we search for multiple candidates of the image's unit facts in parallel, ensuring there is no precedence and no extra interference. Subsequently, we select a subset from these unit-fact candidates to eliminate redundancy and improve efficiency. Finally, we task the LVLM to finalize the caption based on selected unit facts and the image.

Differential score. We first introduce the sequence-level and set-level *differential score*, which are utilized in the search and selection phases to promote diversity. For two sequences, $\boldsymbol{y}^{(1)} = (y_1^{(1)}, ...y_{L_1}^{(1)})$ and $\boldsymbol{y}^{(2)} = (y_1^{(2)}, ...y_{L_2}^{(2)})$, the sequence-level differential score is computed as the negative cosine similarity between their tokens' average hidden state vectors from the LVLM, or equivalently,

$$d_{\mathcal{M}}(\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}) = -\frac{1}{L_1 L_2} \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} \cos(h_i^{(1)}, h_j^{(2)}),$$

where h_i denotes the hidden states of the token y_i. For a set of candidates $\mathcal{Y} = \{\boldsymbol{y}^{(1)}, ...\boldsymbol{y}^{(k)}\}$, the

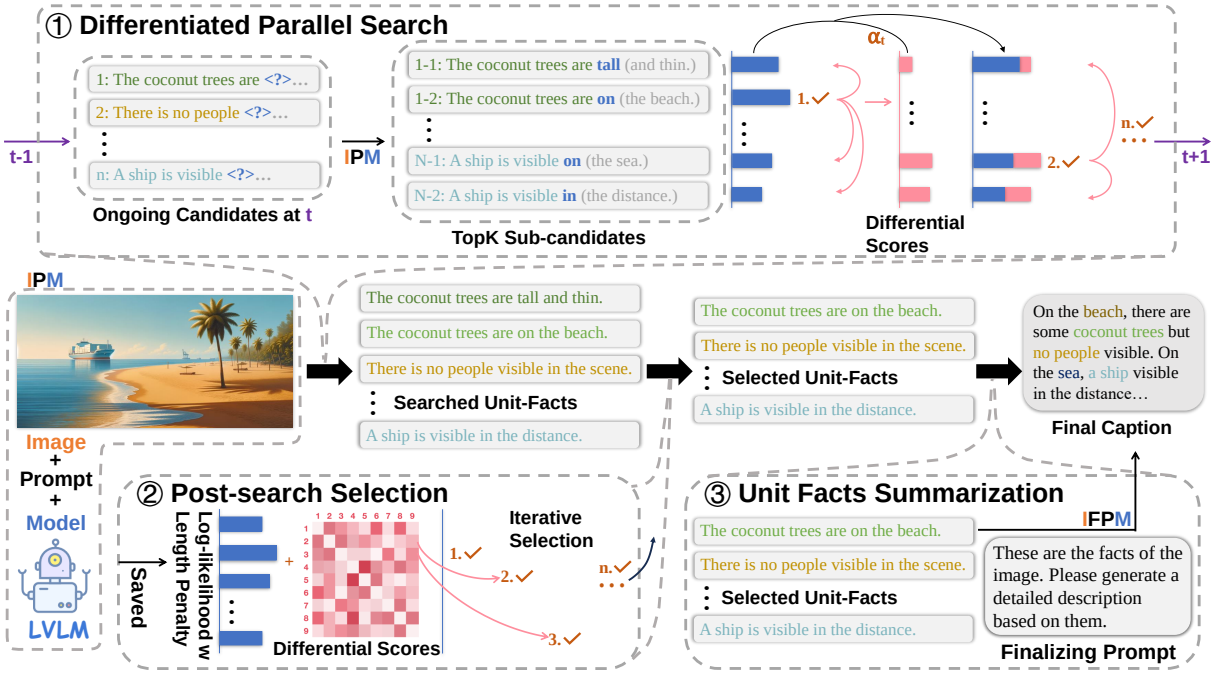

Figure 5: Overview of DBD method. In the Differentiated Parallel Search phase, we balance log-likelihoods from the model and differential scores to search distinct unit facts. In the Post-search Selection phase, we select multiple representative unit facts using length-penalized log-likelihood and differential scores. In the Unit Facts Summarization phase, we task the LVLM to generate the final caption based on selected unit facts and the image.

set-level differential score, $D_{\mathcal{M}}(\mathcal{Y})$ aggregates the sequence-level differential scores across all unique candidate pairs $(\boldsymbol{y}^{(i)}, \boldsymbol{y}^{(j)})$ for $i \neq j$, as follows:

$$D_{\mathcal{M}}(\mathcal{Y}) = \sum_{i=2}^{k} \sum_{j=1}^{i-1} d_{\mathcal{M}}(\boldsymbol{y}^{(i)}, \boldsymbol{y}^{(j)}).$$

Differentiated parallel search. In this phase, we leverage differential scores to guide the generation of independent and distinct unit facts in parallel.

The objective is to identify the optimal set of unit facts, $\mathcal{Y} = \{\boldsymbol{y}^{(1)}, ...\boldsymbol{y}^{(n)}\}$, that contains n unit facts and maximizes a combination of the log-likelihoods and a set-level differential score,

$$\max_{\mathcal{Y}} [\sum_{i=1}^{n} \log p_{\mathcal{M}}(\boldsymbol{y}^{(i)} \mid \boldsymbol{v}, \boldsymbol{x}) + \alpha D_{\mathcal{M}}(\mathcal{Y})]$$

where n is the predefined size of the candidate set and α is a weight that balances the impact of the differential score. Directly addressing this optimization is complex, so we adopt an iterative approach to resolve the sub-problems for $i \in [n]$,

$$\max_{\mathcal{Y}} [\log p_{\mathcal{M}}(\boldsymbol{y}^{(i)} \mid \boldsymbol{v}, \boldsymbol{x}) + \sum_{j=1}^{i-1} d_{\mathcal{M}}(\boldsymbol{y}^{(i)}, \boldsymbol{y}^{(j)})].$$

Specifically, at each step t, we maintain a set of parallel candidates $\mathcal{Y}_t = \{\boldsymbol{y}_{\leq t}^{(1)}, ..., \boldsymbol{y}_{\leq t}^{(n)}\}$, with each candidate representing a unit fact in progress. At the next step $t+1$, for each candidate $\boldsymbol{y}_{\leq t}^{(i)}$, we obtain the top K most probable next tokens predicted by the model, where K is a hyperparameter. These total nK next tokens, along with their source candidates, form a sub-candidate set $\tilde{\mathcal{Y}}_{t+1} = \{\tilde{\boldsymbol{y}}_{\leq t+1}^{(1)}, ..., \tilde{\boldsymbol{y}}_{\leq t+1}^{(nK)}\}$. From this sub-candidate set, we pick next-step candidates, $\mathcal{Y}_{t+1}$. Initially, the sub-candidate with the highest sentence-level log-likelihood is chosen first and removed from $\tilde{\mathcal{Y}}_{t+1}$:

$$\boldsymbol{y}_{\leq t+1}^{(1)} = \underset{\tilde{\boldsymbol{y}}_{\leq t+1} \in \tilde{\mathcal{Y}}_{t+1}}{\arg\max} \ \log p_{\mathcal{M}}(\tilde{\boldsymbol{y}}_{\leq t+1} \mid \boldsymbol{v}, \boldsymbol{x}).$$

Subsequently, we iteratively pick and remove the sub-candidate with the highest hybrid score, defined as the log-likelihood adjusted by a weighted sum of differential scores with existing next-step candidates, from the remaining sub-candidate set. It can be formalized as follows for $2 \leq j \leq n$:

$$\boldsymbol{y}_{\leq t+1}^{(j)} = \underset{\tilde{\boldsymbol{y}}_{\leq t+1} \in \tilde{\mathcal{Y}}'_{t+1}}{\arg\max} \ [\log p_{\mathcal{M}}(\tilde{\boldsymbol{y}}_{\leq t+1} \mid \boldsymbol{v}, \boldsymbol{x}) + \alpha_{t+1} \sum_{k=1}^{j-1} d_{\mathcal{M}}(\tilde{\boldsymbol{y}}_{\leq t+1}, \boldsymbol{y}_{\leq t+1}^{(k)})],$$

where $\tilde{\mathcal{Y}}'_{t+1}$ denotes the remaining sub-candidate set, and α_{t+1} is a parameter that adjusts the influence of differential scores, decreasing as t progresses to encourage differentiation in early stage. Suppose a picked token is the end-of-sequence token (<EOS>). In that case, this completed fact is

moved to a set of final unit-fact candidates, $\mathcal{Y}_{\text{fact}}$, for the next phase, and the next sub-candidate with the highest hybrid score is picked as a replacement. Once n unfinished sub-candidates are chosen, we obtain the next-step candidate set $\mathcal{Y}_{t+1}$.

The procedure repeats until the predefined maximum time step is reached. Throughout the process, multiple completed unit facts are added to the final unit-fact candidate set $\mathcal{Y}_{\text{fact}}$. Typically, $\mathcal{Y}_{\text{fact}}$ contains redundancies that need to be further refined.

Post-search selection. In this phase, we select a subset $\bar{\mathcal{Y}} = \{\bar{\boldsymbol{y}}^{(1)}, ...\bar{\boldsymbol{y}}^{(\bar{n})}\}$, from the obtained facts set $\mathcal{Y}_{\text{fact}}$, to remove redundancy and improve efficiency. The goal is to identify the most representative unit facts that balance sentence-level log-likelihood with a length penalty and differential score. The selection process operates in a manner similar to the first phase. Specifically, it begins by identifying the candidate with the highest pure likelihood score. Then, it is followed by an iterative selection from the remaining facts, prioritizing those with high log-likelihoods and minimal redundancy with the already selected facts. The dynamic criterion is formalized as follows: for $j \in [\bar{n}]$

$$\bar{\boldsymbol{y}}^{(j)} = \arg\max_{\boldsymbol{y} \in \mathcal{Y}_{\text{fact}}} \Big(\frac{1}{|\boldsymbol{y}|} \log p_{\mathcal{M}}(\boldsymbol{y} \mid \boldsymbol{v}, \boldsymbol{x}) + \bar{\alpha} \sum_{k=1}^{j-1} d_{\mathcal{M}}(\boldsymbol{y}, \bar{\boldsymbol{y}}^{(k)}) \Big),$$

where $|\boldsymbol{y}|$ represents the length of the candidate and $1/|\boldsymbol{y}|$ serves as the length penalty factor.

Unit facts summarization. In this phase, the set $\bar{\mathcal{Y}}$ of selected unit facts are presented to the LVLM alongside the image. The model synthesizes these unit facts into a coherent final description. This step not only leverages the explicit information in the unit facts but allows for optional re-evaluation based on the image itself, ensuring that the final caption remains closely aligned with the vision.

4.3 CLIP-Recall, -Precision, and -F1

Current evaluation metrics have notable limitations. CIDEr (Papineni et al., 2002) has been unable to adapt to the diverse expression of captions from current LVLMs. The current rule-based extraction framework of CHAIR (Rohrbach et al., 2018) fails to accurately evaluate detailed captions, where many object words appear not for existence but for decorative features or even non-existence. Addressing this problem requires a deep understanding of

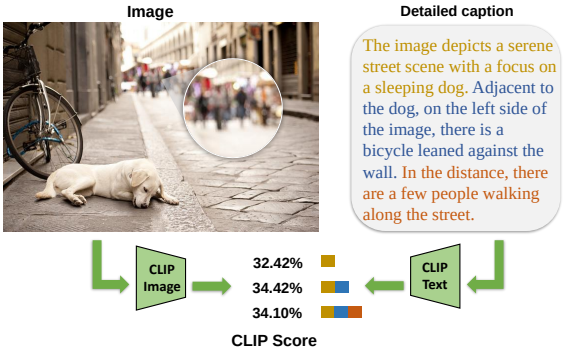

Figure 6: CLIP Scores between the image and captions with different detail levels. Each color square corresponds to a sentence of the same color. The full detailed caption has a lower CLIP score than the second one.

the caption text rather than rigid judgment. Additionally, language ground truths, as lossy compressions of visual information, restrict the evaluation of more detailed captions. To improve applicability, we need to focus on the real ground truth—the image—and its fine-grained comprehension.

These lead us to the CLIP score. However, the CLIP score has weaknesses. First, it only measures the consistency between the vision and language inputs but does not indicate whether the caption is undetailed or over-hallucinated. Moreover, the CLIP encoders focus on the main objects and overlook details, lacking a deep and fine-grained digest. Figure 6 shows an example where the CLIP image encoder omits the information in the background, leading to a lower CLIP score for the most detailed caption. In response, we propose a novel set of metrics: CLIP-Precision, CLIP-Recall, and CLIP-F1.

For an image-caption pair $(\boldsymbol{v}, \boldsymbol{y})$, assume we have dense partitions of both vision and language. For the image, partitions are cropped sub-images of different objects or environments, including the full image. For the caption, partitions are sentences or phrases describing different objects or their characteristics, including the full caption. Given these image and caption partitions, we utilize the CLIP image and text encoders to encode them, obtaining two sets of embeddings, both in various granularity levels. Let $\mathcal{V}^{\text{c}} = \{\boldsymbol{v}_i^{\text{c}}\}_1^{n_v}$ and $\mathcal{Y}^{\text{c}} = \{\boldsymbol{y}_i^{\text{c}}\}_1^{n_y}$ denote the set of embeddings from image and caption partitions, respectively, where n_v and n_y represent the set sizes. We then compute the proportion that one embedding set can be linearly represented by the other as CLIP-Recall or CLIP-Precision and combine them to produce CLIP-F1.

Consider an embedding $\boldsymbol{a} \in \mathbb{R}^{n_e}$ and a set $\mathcal{B} = \{\boldsymbol{b}_i\}_1^{n_b}$, where n_e is the number of dimensions, n_b

Model		MiniGPT-4			LLaVA-1.5			mPLUG-Owl2		
Method	k	CLIP-Recall	CLIP-Precision	CLIP-F1	CLIP-Recall	CLIP-Precision	CLIP-F1	CLIP-Recall	CLIP-Precision	CLIP-F1
Greedy	3	63.33	64.62	63.83	63.59	70.43	66.73	63.29	68.66	65.78
Beam	3	62.17	65.58	63.69	63.81	71.23	67.20	63.18	69.34	66.02
VCD	3	63.23	64.47	63.70	63.50	70.91	66.89	63.05	69.09	65.83
Opera	3	61.91	**65.60**	63.56	63.63	70.91	66.96	62.84	69.74	66.01
DBD (Ours)	3	**65.54**	64.46	**64.89**	**64.00**	**71.38**	**67.22**	**63.41**	**70.08**	**66.45**
Greedy	5	65.48	65.67	65.42	65.05	71.94	68.20	64.92	69.80	67.17
Beam	5	64.15	**66.50**	65.24	65.20	71.96	68.29	64.77	70.45	67.38
VCD	5	65.36	65.53	65.29	65.58	72.21	68.61	64.65	70.21	67.20
Opera	5	63.95	66.05	65.13	65.15	71.66	68.13	64.34	70.87	67.33
DBD (Ours)	5	**67.82**	65.63	**66.52**	**66.20**	**72.50**	**68.92**	**65.09**	**71.25**	**67.89**
Greedy	10	67.74	68.35	67.85	65.85	74.99	69.98	65.99	72.55	68.99
Beam	10	66.24	69.43	67.62	66.23	74.85	70.14	**66.01**	73.29	69.34
VCD	10	67.60	68.20	67.70	66.10	75.11	70.17	65.75	73.07	69.09
Opera	10	66.11	**69.46**	67.55	65.97	74.56	69.87	65.24	73.71	69.09
DBD (Ours)	10	**70.75**	67.94	**69.18**	**66.65**	**75.54**	**70.31**	65.54	**74.35**	**69.51**

Table 1: Proposed CLIP metric sets evaluation results on the Visual Genome dataset of LVLMS with different methods. Bold indicates the best results of all methods. k is the parameter for the CLIP metric set. For all k, higher CLIP-Recall indicates more details in captions; lower CLIP-Precision represents more hallucinations.

is the size of $\mathcal{B}$. The proportion of top$k$-similar linear representation for $\boldsymbol{a}$ by $\mathcal{B}$ is defined as the l_2 norm ratio of the best linear combination of top$k$-similar embeddings in $\mathcal{B}$ to the embedding $\boldsymbol{a}$,

$$\text{PLR}_k(\boldsymbol{a}; \mathcal{B}) := \frac{||B_k\gamma^*||_2}{||\boldsymbol{a}||_2},$$
$$\text{s.t. } \gamma^* = \arg\min_{\gamma} ||B_k\gamma - \boldsymbol{a}||_2,$$

where k is a parameter, $B_k \in \mathbb{R}^{n_e \times \min(k, n_b)}$ is a matrix whose columns are the embeddings $\boldsymbol{b}_i \in \mathcal{B}$ with the highest similarities to $\boldsymbol{a}$, and $||\cdot||_2$ is the l_2 norm. Notably, when we set $k = 1$, or there is only one embedding in $\mathcal{B}$, i.e., $n_b = 1$, the proportion of linear representation generalizes *cosine similarity*. We discuss the effect of k in Appendix A.3.1.

To obtain the recall, for each $\boldsymbol{v}_i^{\text{c}} \in \mathcal{V}^{\text{c}}$, we collect from $\mathcal{Y}^{\text{c}}$ caption partition embeddings with the top$k$ similarities to $\boldsymbol{v}_i^{\text{c}}$ and compute the proportion of linear representation. Then, CLIP-Recall is computed as the average PLR by Y_k^{c} across $\mathcal{V}^{\text{c}}$, rescaled by $\omega = 2.5$ (following Hessel et al. (2022)):

$$\text{CLIP-Recall} = \frac{\omega}{n_v} \sum_{i=1}^{n_v} \text{PLR}_k(\boldsymbol{v}_i^{\text{c}}; \mathcal{Y}^{\text{c}}).$$

Similarly, we define CLIP-Precision as the average proportion of top$k$-similar linear representation by the image partitions across the caption partitions:

$$\text{CLIP-Precision} = \frac{\omega}{n_v} \sum_{i=1}^{n_v} \text{PLR}_k(\boldsymbol{y}_i^{\text{c}}; \mathcal{V}^{\text{c}}).$$

CLIP-Recall shows how much information in the image can be explained by the caption in various granularity levels, and CLIP-Precision measures the reverse. A higher CLIP-Recall indicates a more detailed caption, while a lower CLIP-Precision indicates more hallucinations. Finally, we combine CLIP-Recall and CLIP-Precision to generally measure how well the caption aligns with the image:

$$\text{CLIP-F1} := \frac{2\text{CLIP-Recall} \cdot \text{CLIP-Precision}}{\text{CLIP-Recall} + \text{CLIP-Precision}}.$$

5 Experiments

5.1 CLIP Metric Set Evaluation

LVLMs and dataset. We evaluate on three state-of-the-art LVLMs, LLaVA-1.5 (Liu et al., 2023c), mPLUG-Owl2 (Ye et al., 2023a), and Minigpt-4-v2 (Chen et al., 2023a). Our experiments are conducted on the Visual Genome dataset (Krishna et al., 2017), from which we randomly pick 500 images from the validation set using the same random seed across different methods and LVLMs. We repeat the experiments 5 times with different random seeds and report the averaged results.

Baselines. To validate our method, we compare it against classic decoding baselines such as greedy decoding and beam search, as well as state-of-the-art approaches to mitigate object hallucination in image captioning, including VCD (Leng et al., 2023) and Opera (Huang et al., 2024). Details on baseline settings are provided in Appendix A.3.2.

Metric implementation details. We employ our proposed metrics—CLIP-Recall, CLIP-Precision, and CLIP-F1—for evaluation. We utilize the ViT-L/14 CLIP image and text encoder for these metrics. The Visual Genome dataset annotates dense regions and objects within each image with specific bounding boxes. We utilize these bounding boxes of regions and objects as the vision partitions of the images. Each image is also treated as a special partition in its entirety. For language partitions, we segment the captions into sentences based on punctuation marks like periods, question marks,

Model	α_t	CLIP-Recall	CLIP-Precision	CLIP-F1
MiniGPT-4	2	70.65	67.76	69.00
	3	70.51	67.42	68.78
	4	**70.66**	**67.82**	**69.04**
LLaVA-1.5	2	66.49	**75.29**	70.47
	3	66.33	75.24	70.36
	4	66.47	75.17	70.40
	5	**66.63**	**75.29**	**70.54**
mPLUG-Owl2	2	65.45	74.03	69.32
	3	65.66	74.11	69.47
	4	**65.70**	74.07	69.48
	5	65.65	**74.22**	**69.51**

Table 2: Proposed CLIP metric sets evaluation results on DBD with different α_t. Higher α_t represents a stronger influence from the differential score on the decoding.

and exclamation points and include the full caption as well. We compute the CLIP metric set with different parameters k and report all results.

Method implementation details. In the search phase, we set the differential weight α_t as 10 in the first 3 decoding step, i.e., $t \leq 3$, reducing it to 5 subsequently. We use 5 differentiated beams with a top K setting of 6. In the selection phase, we set a differential selection weight $\bar{\alpha}$ as 5 and the number of final unit facts $\bar{n}$ as 10. For more implementation details, please refer to the Appendix A.2.3.

Results. We present the evaluation results of our proposed CLIP metric set for our method, DBD, and various baselines across different LVLMs, as compiled in Table 1. The results consistently show that our DBD method surpasses all baseline methods. Compared to Greedy, Beam Search, and VCD, our method demonstrates superior CLIP-Recall and CLIP-Precision, indicating that DBD effectively enhances detail in LVLM-based image captioning without increasing the hallucination rate. While Opera occasionally achieves better results on CLIP-Precision, this is paid for by its generally lower scores on CLIP-Recall and CLIP-F1. Detailed caption comparison examples from the Visual Genome dataset are provided in Appendix A.3.2.

Abalation study. We report the ablation study on differential score weight α_t in Table 2. We set $k = 10$. The results illustrate that while variations in α_t impact performance, DBD remains robust and not overly sensitive to this hyperparameter.

5.2 Human Evaluation

We conducted a human evaluation study with 30 participants and 253 sampled questions to assess the performance of our method relative to baselines and validate the alignment between our proposed evaluation metrics and human preferences. The results, summarized in Table 3, not only showcase

	More Details	**Fewer Hallucinations**	**Overall**
Total Questions	80	74	99
Percentage (%)	55.00	58.11	62.63

Table 3: User Study Results. The study focused on three aspects: which captions contain more details, which contain fewer hallucinations, and which are better overall.

Method	Greedy	Beam	VCD	Opera	DBD
MiniGPT4	3.2	4.5	8.5	29.7	6.1
LLaVA-1.5	2.2	3.1	4.5	20.6	6.2
mPLUG-Owl2	4.3	5.1	8.8	34.3	7.3

Table 4: Inference time (s/sample) of our method, DBD, and baselines when using different LVLMs.

the effectiveness of our method but also confirm the consistency between our evaluation metrics and human judgments. More details about our user study can be found in Appendix A.3.3.

5.3 Inference Cost

Although DBD involves multiple beams, the final summarization pass, and differentiated score computation, the inference cost of our method, DBD, is not as high as imagined. The main reason is that, unlike beam decoding, the differentiated beams decode different unit facts, which are much shorter than a full caption. This allows us to set a small max token length in the parallel search phase. Particularly, when the max token length of final outputs is set as 256, we set the max token length of unit facts as 64. Moreover, the early stop strategy, when the candidate set reaches the predefined size, further reduces the inference cost. For instance, the longest unit fact using LLaVA-1.5 only has 49.9 tokens on average. We provide the per-sample inference time of our method and baselines when running on the H100 GPU in Table 4.

6 Conclusion

In this paper, we address the debate on whether more detailed descriptions in LVLM-based image captioning necessarily lead to increased object hallucinations. We propose a novel decoding method, Differentiated Beam Decoding (DBD), accompanied by a new set of metrics: CLIP-Recall, -Precision, and -F1. Extensive experiments on the Visual Genome dataset validate that our method produces more detailed and accurate image captions while effectively maintaining a low hallucination level. Future research can extend our methods to further improve the robustness and reliability of LVLM-based applications across diverse scenarios.

Limitations

While our method (DBD) effectively mitigates the problem of extra interference from pre-generated texts, it does not address other potential sources of hallucinations. However, the unique parallel decoding structure of DBD offers the potential for integration with most existing hallucination-mitigating methods, a synergy that remains to be explored. Additionally, our proposed CLIP metric set relies on given image partitions. This requirement constrained our experiments to Visual Genome, which, despite its comprehensiveness, may not capture the full diversity of real-world scenarios.

References

Mohammad Alsharid, Yifan Cai, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. 2022. Gaze-assisted automatic captioning of fetal ultrasound videos using three-way multi-modal deep neural networks. *Medical Image Analysis*, 82:102630.

Mohammad Alsharid, Rasheed El-Bouri, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. 2021. A course-focused dual curriculum for image captioning. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 716–720.

Shuang Bai and Shan An. 2018. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2023c. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. *Preprint*, arXiv:2310.01957.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *Preprint*, arXiv:2402.15300.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16, pages 18135–18143.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning. *Preprint*, arXiv:2104.08718.

Mingzhe Hu, Shaoyan Pan, Yuheng Li, and Xiaofeng Yang. 2023. Advancing medical imaging with language models: A journey from n-grams to chatgpt. *Preprint*, arXiv:2304.04920.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *Preprint*, arXiv:2311.17911.

Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. 2023. Adapt: Action-aware driving caption transformer. *Preprint*, arXiv:2302.00673.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023c. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023d. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. 2022. Vision-and-language pretrained models: A survey. *arXiv preprint arXiv:2204.07356*.

Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. 2023. Evaluation and mitigation of agnosia in multimodal large language models. *arXiv preprint arXiv:2309.04041*.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, 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.

Dongmin Park, Zhaofang Qian, Guangxing Han, and Ser-Nam Lim. 2024. Mitigating dialogue hallucination for large multi-modal models via adversarial instruction tuning. *arXiv preprint arXiv:2403.10492*.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Haoran Wang, Yue Zhang, and Xiaosheng Yu. 2020. An overview of image caption generation methods. *Computational intelligence and neuroscience*, 2020(1):3062706.

Teng Wang, Jinrui Zhang, Junjie Fei, Yixiao Ge, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, Shanshan Zhao, Ying Shan, et al. 2023. Caption anything: Interactive image description with diverse multimodal controls. *arXiv preprint arXiv:2305.02677*.

Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. 2023. Pink: Unveiling the power of referential comprehension for multi-modal llms. *arXiv preprint arXiv:2310.00582*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023a. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *Preprint*, arXiv:2311.04257.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023b. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

A.1 Detail-Hallucination Debate

Co-occurrence of Detail and Hallucination. To fully support the co-occurrence side's arguments, we show that LVLMs tend to describe details in the later part of captions by a simple experiment. Concretely, we first generate captions of 500 images from the Visual Genome dataset using different LVLMs. Then, we match the object words in the captions with the object annotations in the dataset and compute the object size using the corresponding bounding box annotations. We report how the average object size changes depending on where it appears in the caption. As presented in the Figure 7, the average size of objects consistently decreases as the sentence progresses, validating the statement.

One may find the co-occurrence of details and hallucinations similar to the natural phenomenon that "the more you say, the more mistakes you make." However, they are different. The former is saying that when describing details, it's more likely to produce hallucinations, and the latter is still tenable even if the hallucination frequency is common across all positions in the description.

Inflated Hallucination Presence. We provide more examples of failures in the current hallucination evaluation framework in Figure 8 and Figure 9. Images are from the MSCOCO validation set. The captions are generated by LLaVA-1.5 using our proposed decoding method, DBD.

A.2 Methodology: DBD

A.2.1 Preliminaries

LVLMs. The architecture of the LVLM $\mathcal{M}$ incorporates a visual encoder, a cross-modality aligner, and a Large Language Model (LLM) core structure. Initially, the input undergoes processing by the visual encoder, followed by the cross-modality aligner to obtain a visual embedding in the textual space. Subsequently, the LLM core utilizes this visual-textual embedding, along with the text prompt and previously generated tokens, to predict the probability of the next token.

A.2.2 Discussion

Beam Search. Our approach is grounded in Beam Search (BS), which aims to select an output sequence that maximizes the total log-likelihood, not just the highest probability next token at each step:

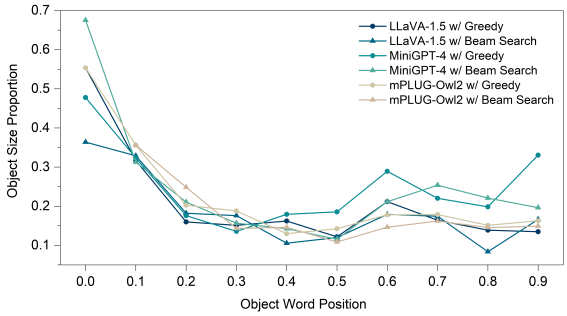

Figure 7: Object word position in the caption v.s. Object size proportion to the image. As the caption progresses, LVLMs are prone to describe smaller objects.

$$\boldsymbol{y}_{\text{Beam}} = \arg\max_{\boldsymbol{y} \in S^{\bar{L}}} \sum_{t=1}^{\bar{L}} \log p_{\mathcal{M}}(y_t \mid \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}_{<t}),$$

where $\bar{L}$ denotes the maximum output length and $\log p_{\mathcal{M}}(\boldsymbol{y} \mid \boldsymbol{v}, \boldsymbol{x})$ represents the sentence-level log-likelihood. BS maintains N_{beam} candidate sequences, each with a corresponding beam score, i.e., log-likelihood. In each step, BS calculates the top K probable next tokens for each candidate sequence, resulting in N beam K sub-candidates and their new log-likelihood scores. The top N_{beam} sequences with the highest scores are then selected to continue to the next step.

Relationship with Beam Search.To eliminate the extra interference from pre-generated text in typical strategies, an intuitive approach is to process several separate output candidates simultaneously so that there is no precedence. While this sounds similar to Beam Search (BS), the candidates in conventional BS usually converge to similar full descriptions and focus on the same aspects of the image. Moreover, within each candidate, the generation of later texts is still influenced by earlier ones. Our method, however, encourages the exploration of distinct narrative directions in parallel and concludes when a single unit fact is completed.

A.2.3 Implementation Details

Direct instruction for random aspects. The inherent language priors of LVLMs often manifest a preference for describing certain aspects of an image over others, typically favoring a progression from central to peripheral elements or from general to specific details. This bias can impede the generation of differentiated facts, particularly when using a standard prompt such as *"Please describe the image"*. To circumvent this issue, we leverage the LVLM's robust generalization capability by mod-

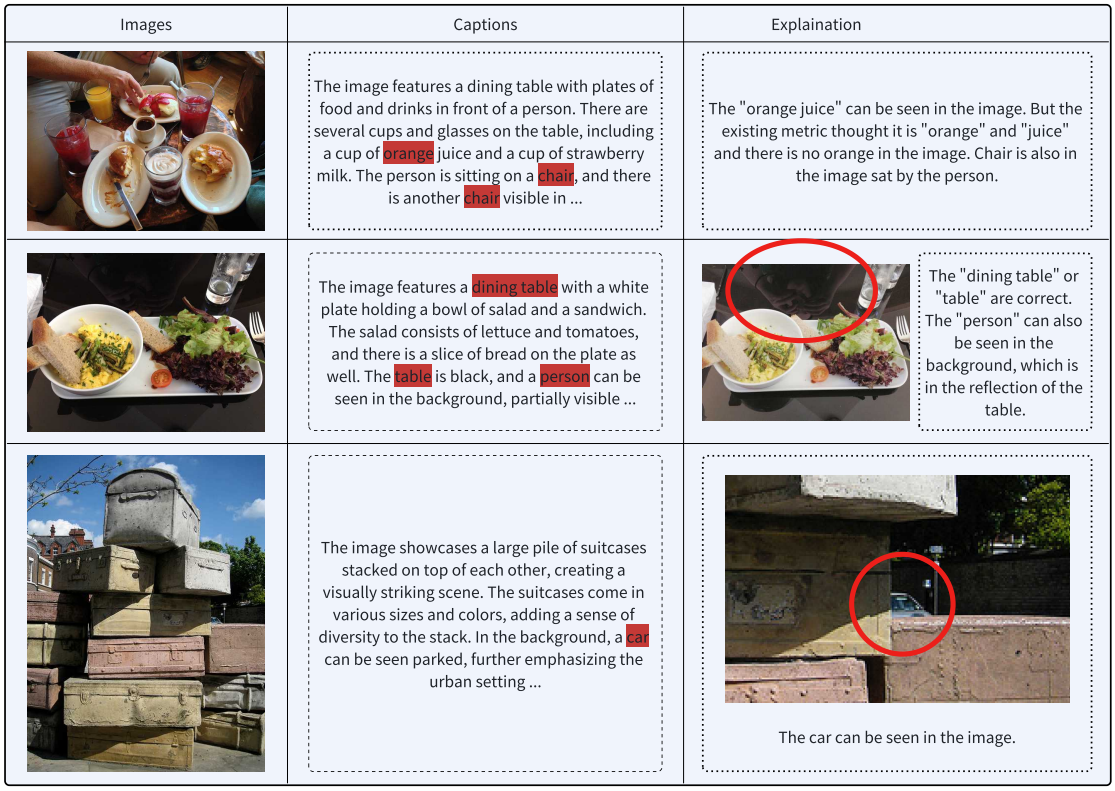

Figure 8: Examples illustrating misjudgments in image captioning. The left column contains images; the middle column shows automatically generated captions with highlighted sections in red that are considered incorrect by existing metrics (hallucination words). However, our model's captions are actually correct. The right column provides explanations and clarifications, demonstrating the accuracy of the highlighted sections and the flaws of the previous metric.

ifying our prompt during the differentiated beam decoding phase.

Specifically, we enhance the prompt to clarify the final goal of image captioning while also directly instructing the LVLM to focus on a randomly selected aspect of the image. This strategy encourages the model to diverge from its typical pattern of description, facilitating the generation of varied and independent factual content. For LLaVA-1.5 and mPLUG-Owl2, we use *"Please generate a random fact of the image. You can describe the main object, the background, the environment, or any other detail. Please make sure the choice of the fact is random. Do not only focus on the people or the main object."* For MiniGPT-4, we add another *"Give me the fact directly without other words"* at the end of the prompt to prevent MiniGPT-4 from generating too much nonsense. By directing the LVLM to select aspects randomly, we promote a more equitable consideration of all parts of the image, thereby supporting the production of a comprehensive and balanced description.

In the summarization phase, we prompt the *"These are the image and the facts of the image. Please summarize them and generate a detailed description of the image based on the facts and the image"* for LLaVA-1.5 and mPLUG-Owl2. Similarly, we add a *"Give me the fact directly without other words"* at the end.

A.3 Evaluation

A.3.1 CLIP Metric Set

Hyperparameter k. The value of the parameter k represents how many of the most relevant embeddings can be used to fit an embedding from another modal. The design of the parameter k is based on the observation that a description of an object or region may not be fully captured in just one sentence. A detailed caption may need to cover different aspects of the same area or object, leading to the use of multiple sentences. This requires allowing for a more accurate approximation of visual facts. Similarly, the evaluation of an overall description of a larger region may require the coordination of

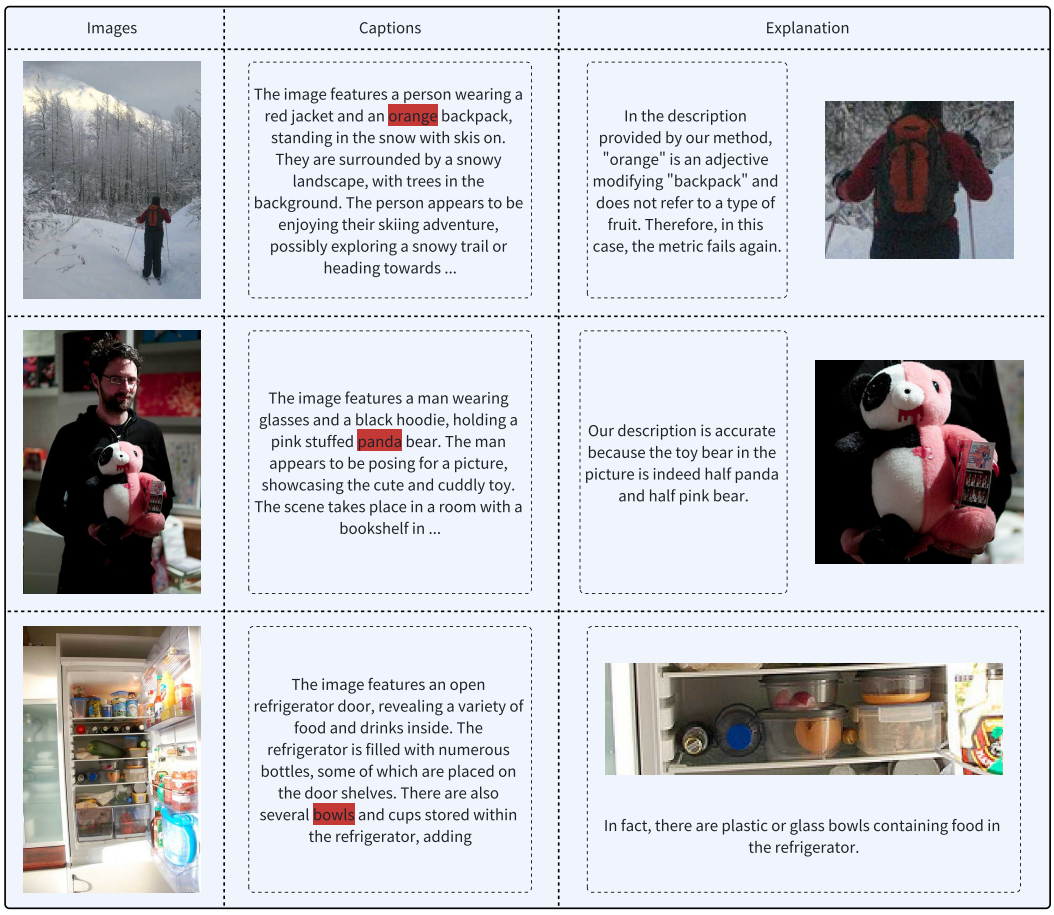

Figure 9: More examples illustrating misjudgments in image captioning.

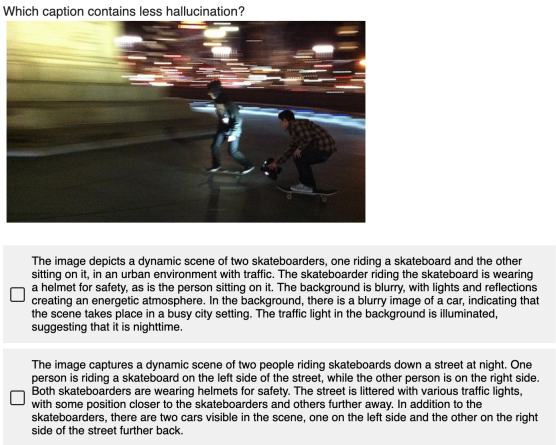

Figure 10: Human Evaluation/User Study Interface.

multiple dense image partitions. A higher value of k would provide more flexibility in combining different partitions. However, setting k too high may lead to interference from irrelevant partitions.

A.3.2 Experiments

Baseline settings. We provide the hyperparameter you used in the experiments for baselines. For greedy search, there is no hyperparameter. For beam search, we set the number of beams as 5 and the top$k$ as 5. For VCD (Leng et al., 2023), we set $\alpha = 1$, $\beta = 0.1$, and $t_{\text{noise}} = 500$. For Opera (Huang et al., 2024), we set the number of beams as 5, the top$k$ as 5, the scale factor as 50, the threshold as 15, the number of attention candidates as 5, and the penalty weight as 1. We prompt all baselines with "Please describe this image in detail."

Caption examples. We provide examples of descriptions generated by our method on images from the Visual Genome dataset in Figure 11.

A.3.3 Human Evaluation Interface

Participants are presented with images and two corresponding captions generated by different models, and the interface used is as shown in Figure 10. They are asked to determine which caption contains more details (*Which caption contains more details from the image?*), fewer hallucinations (*Which caption contains less hallucination?*), and overall preference (*Which caption is better overall?*).

Figure 11: Comparison of image captioning results using different models and decoding methods: LLAVA-1.5 with Greedy and Beam Search, MiniGPT-4 with Greedy and UnitFact, and mPLUG-Owl2 with Greedy and UnitFact. The captions illustrate the differences in the models' abilities to accurately describe the content of the images, with UnitFact generally providing more detailed and contextually rich descriptions compared to Greedy and Beam Search strategies.