A Comprehensive Analysis for Visual Object Hallucination in Large Vision-Language Models

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) demonstrate remarkable capabilities in multimodal tasks, but visual object hallucination remains a persistent issue. It refers to scenarios where models generate inaccurate visual object-related information based on the query input, potentially leading to misinformation and concerns about safety and reliability. Previous works focus on the evaluation and mitigation of visual hallucinations, but the underlying causes have not been comprehensively investigated. In this paper, we analyze each component of LLaVA-like LVLMs—the large language model, the vision backbone, and the projector, to identify potential sources of error and their impact. Based on our observations, we propose methods to mitigate hallucination for each problematic component. Additionally, we developed two hallucination benchmarks: QA-VisualGenome, which emphasizes attribute and relation hallucinations, and QA-FB15k, which focuses on cognition-based hallucinations.

1 Introduction

Large Language Models (LLMs), such as GPT-3 (Brown, 2020) and ChatGPT (OpenAI, 2022), have showcased remarkable proficiency in language tasks, yet they encounter significant challenges when it comes to processing multimodal inputs. This limitation has driven a shift in research towards Large Vision-Language Models (LVLMs) (Liu et al., 2023e; Ye et al., 2023; Sun et al., 2023b), which integrate advanced LLMs (Touvron et al., 2023; Chiang et al., 2023) with Vision Foundation Models (VFMs) (Dosovitskiy et al., 2021; Bommasani et al., 2021) to enhance multimodal understanding. LVLMs have demonstrated impressive capabilities across various tasks that require visual and textual integration, including Visual Question Answering (Antol et al., 2015), Image Captioning (Lin et al., 2014), and Visual Entailment (Zhang et al., 2025).

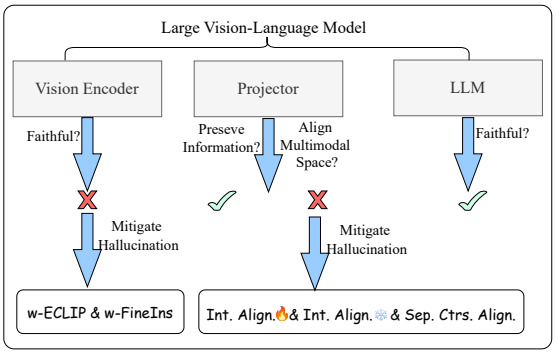

Figure 1: An overview of our paper. We first investigate the sources of hallucination from a component-level perspective within the LVLM architecture. Based on the identified causes, we then design targeted methods to mitigate hallucinations effectively.

Despite these advances, visual hallucination remains a persistent issue in LVLMs (Rohrbach et al., 2018; Liu et al., 2023b,a; Yin et al., 2023; Zhang et al., 2024b). This phenomenon occurs when models generate inaccurate or misleading information unrelated to the actual visual input, potentially leading to misinformation and raising concerns about safety and reliability in real-world applications (Li et al., 2023e). Visual object hallucination, including object existence, attribute, and relation, has garnered significant attention due to its widespread occurrence in images. Current works on visual object hallucination mainly focus on evaluation and mitigation. For example, Li et al. (2023e) extends CHAIR (Rohrbach et al., 2018) and proposes POPE, a polling-based query technique for probing object-level hallucination. For hallucination mitigation, Sun et al. (2023a) introduce new alignment algorithm called Factually Augmented RLHF that augments the reward model with additional factual information such as image captions and ground-truth multi-choice options, which alleviates the reward hacking phenomenon in RLHF and further improves the performance.

While existing works have achieved notable suc-

cess in visual object hallucination, they lack a comprehensive component-level analysis of the model architecture to pinpoint where and how hallucinations occur. In this work, we focus on visual object-related hallucination and LLaVA-like LVLMs, which typically consist of three modules: the large language model (LLM), the vision backbone, and the projector. Errors in any of these modules can lead to issues in the overall performance or functionality of the model. Therefore, we conduct an independent analysis of each component to identify potential sources of error and their impact. From our study, we have the following findings. 1) The LLM in LVLM is able to generate faithful content when captions of images are provided as input. 2) Hallucinations exist in the perception process of the vision backbone. 3) Projector is able to preserve visual features, but has trouble aligning between visual and textual spaces.

Based on our observations, we propose methods for the two problematic components to mitigate their hallucination issue. To improve the **vision backbone**, we propose to finetune CLIP with fine-grained data and fine-grained perception-based visual instruction tuning, and find that both of them can reduce hallucination caused by the vision backbone. For the **projector**, we propose a contrastive alignment objective with three variations, which can all be integrated into the original training pipeline with minimal additional costs.

To conduct a comprehensive hallucination evaluation, we develop a fine-grained hallucination benchmark named QA-VisualGenome, which is built upon the Visual Genome dataset (Krishna et al., 2017). Unlike existing object-oriented hallucination benchmarks (*e.g.*, POPE), QA-VisualGenome emphasizes the detailed attribute and relationship hallucinations. Furthermore, existing hallucination benchmarks primarily focus on perception-based hallucinations for general objects, neglecting cognition-based hallucinations such as the names of people and famous buildings. To address this gap, we construct a cognition-based hallucination benchmark named QA-FB15K, which is based on the FB-15K dataset (Bordes et al., 2013), a multimodal knowledge graph with textual entities, image entities, and textual relations. QA-FB15K presents challenges for models in leveraging world knowledge to solve the questions.

Our main content is shown in Figure 1. Our contributions can be summarized as follows: 1) We analyze the hallucination caused by each component in LVLMs and provide component-wise takeaway messages. 2) Based on our observation, we propose several methods to improve each hallucinated component. 3) We construct a fine-grained hallucination benchmark based on Visual Genome and a cognition-based hallucination benchmark based on FB15k for evaluation. 4) We extensively evaluate our proposed methods on various benchmarks, and provide in-depth analysis[1].

2 Hallucination Analysis

LVLMs consist of three components: language decoder $\mathcal{D}$, projector vision encoder $\mathcal{V}$, and $\mathcal{P}$. We first introduce the datasets for evaluation and then provide in-depth analysis for each component.

2.1 Settings

We select two benchmarks to benchmark the performance of each component. 1) **POPE** (Li et al., 2023e). POPE is a benchmark designed for evaluating object existence hallucinations in LVLMs, incorporating three sampling methods for generating negative samples: random, popular, and adversarial. In the random setting, objects not present in the image are randomly selected. In the popular setting, negative samples are drawn from a pool of frequently occurring objects. In the adversarial setting, the sampling focuses on objects that frequently co-occur with present objects but do not exist in the image. 2) **QA-VisualGenome**. To further investigate the hallucination issue on relations and attributes of objects, we construct a fine-grained evaluation benchmark based on the VisualGenome dataset (Krishna et al., 2017), which collects dense annotations of attributes and relationships of objects for each image. Specifically, we design two types of Yes-or-No questions to evaluate models: attributes and relations. For example, an attribute question could be "*Is the dog red in the image?*" A relational question would ask, "*Is the dog standing on the table?*". Similar to previous work (Wang et al., 2020), we exclude uncommon relations and attributes. We randomly select relations/attributes to generate negative samples.

2.2 Language Decoder

Conjecture 1. LLM in LVLM is able to generate faithful content when image captions are

[1]All benchmark datasets, code, and models will be released.

Table 1: Performance (%) of LLMs across different datasets when visual information is provided in textual format. *LLaVA*: image+text query as input on original LLaVA model; *Vicuna*: caption+text query as input on Vicuna-1.5; $Vicuna_{LLaVA}$: caption+text query as input on the Vicuna model in LLaVA (LLM undergone visual instruction tuning).

Model	POPE						QA-VisualGenome			
	Random		Popular		Adversarial		Attribute		Relation	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LLaVA-7B	87.42	86.36	86.63	85.25	85.13	83.88	64.67	66.60	67.57	74.81
Vicuna-7B	92.67	92.09	92.67	92.09	93.00	92.47	57.23	69.83	79.50	**80.79**
Vicuna-7B$_{LLaVA}$	100.00	**100.00**	100.00	**100.00**	99.67	**99.67**	68.29	**75.92**	63.2	73.06
LLaVA-13B	91.33	91.72	88.33	89.16	84.33	85.97	55.99	68.86	56.40	69.38
Vicuna-13B	87.90	89.15	95.00	95.24	90.00	90.91	87.90	**89.15**	87.90	**89.25**
Vicuna-13B$_{LLaVA}$	99.67	**99.67**	99.67	**99.60**	99.33	**99.33**	75.41	80.10	84.30	84.29

provided as input. To validate this conjecture, we use the POPE dataset to evaluate the performance of LLMs. Instead of providing images to the LVLMs, we only input text descriptions of the images. For POPE, we obtain objects from the MSCOCO (Lin et al., 2014) dataset and feed the LVLM with objects in the image and the textual query from POPE to generate the response. For QA-VisualGenome, we feed the LVLM with objects, object attributes, and relations presented in the image to replace visual information. This helped assess the model's ability to hallucinate when provided with accurate textual descriptions of the image. In addition, we also test the original Vicuna as a baseline.

We show the performance of LLMs in Table 1. From the results, we found that the performance will be improved largely if we provide the correct visual information in a textual format. This indicates the current main reason for hallucination is caused by a vision encoder or projector. Specifically, the model could achieve an accuracy of 99.67% when provided with complete object descriptions for the random setting of POPE, which shows the LLM is robust when given the correct information about the whole image. In addition, we also found that the LLM after the pertaining and instruction tuning of LLaVA performs better than the original LLM. LLaVA fine-tuning likely enhances the model's object recognition, memory of object-specific features, instruction-following ability, and contextual understanding of visual descriptions, enabling it to accurately identify common objects within text descriptions even without actual images.

2.3 Vision Encoder

Conjecture 2. There are hallucinations in the perception process of the vision encoder. To verify this factor, we conducted experiments using CLIP on a text-image matching task. Specifically, we designed a template of the form "There is a/an {object} in the image," where {object} corresponds to various objects in the input images. For each image, we assigned one ground-truth object and a hallucinated object for the template. We use accuracy as the evaluation metric. We show all the experimental results in Table 2. Overall, we found that the performance of CLIP on the text-matching task is not good. For example, the performance of CLIP on the text-image matching task is 83.33% accuracy on the random setting of POPE, indicating the presence of hallucinations within the vision encoder's perception process.

Table 2: Performance of CLIP in the text-image matching across different datasets measured by Accuracy (%).

POPE			QA-VisualGenome	
Random	Popular	Adversarial	Attribute	Relation
83.33	87.30	86.00	61.57	60.22

Another interesting phenomenon is that the accuracy of CLIP in recognizing objects is worse than LLaVA, even the LLaVA adopts CLIP as the vision encoder. Specifically, the accuracy of LLaVA is 91.33% on the random setting of POPE, but CLIP only achieves 83.33% accuracy. This indicates that the hallucination caused by CLIP can be alleviated to a certain extent after the pre-training feature alignment and instruction tuning. The potential reason may be that LLaVA's training uses diverse questions aligned with specific image features, optimizing for generative loss. This fine-grained alignment helps the model better understand and describe visual content with greater accuracy and detail.

2.4 Projector

We analyze the projector module from two perspectives corresponding to its two roles in the LVLM: *preserving visual information* and *aligning visual*

and textual spaces.

Conjecture 3. The projector should not result in significant visual information loss. We formalize the hypothesis using the notion of *V-information* (Hewitt et al., 2021). Let $\Phi_{pre}(X)$ and $\Phi_{post}(X)$ represent the pre-projector and post-projector representations, respectively. We compare the V-information between these representations and a target property Y (e.g., a classification label).

We define the V-information for pre- and post-projector representations as

$$I_{\mathcal{V}}(\Phi_{pre}(X) \to Y) = H_{\mathcal{V}}(Y) - H_{\mathcal{V}}(Y|\Phi_{pre}(X))$$

$$I_{\mathcal{V}}(\Phi_{post}(X) \to Y) = H_{\mathcal{V}}(Y) - H_{\mathcal{V}}(Y|\Phi_{post}(X))$$

where $H_{\mathcal{V}}$ is the *V-entropy* (Hewitt et al., 2021). $H_{\mathcal{V}}(Y)$ is the entropy of Y, which reflects the inherent uncertainty of Y without any conditioning on the representations. $H_{\mathcal{V}}(Y|\Phi(X))$ represents the uncertainty we have in predicting Y after observing the representation $\Phi(X)$, using functions from the family $\mathcal{V}$. It is formally defined as:

$$H_{\mathcal{V}}(Y|\Phi(X)) = \inf_{f \in \mathcal{V}} \mathbb{E}_{\Phi(X),Y}\left[-\log f(\Phi(X))(Y)\right]$$

This expression measures the best performance that a function f from the function family $\mathcal{V}$ can achieve when predicting Y given the representation $\Phi(X)$. The lower this value, the more predictive power the representation $\Phi(X)$ has regarding Y.

The goal is to determine whether information loss occurs in the projection layer. If the projection layer introduces no information loss, then the V-information of the pre-projector and post-projector representations should be approximately equal:

$$I_{\mathcal{V}}(\Phi_{pre}(X) \to Y) = I_{\mathcal{V}}(\Phi_{post}(X) \to Y)$$

We compare the V-information accessible from both the pre-projector and post-projector representations. The performance of a probe (e.g., classifier) trained on $\Phi_{pre}(X)$ and $\Phi_{\text{post}}(X)$ provides an empirical estimate of these quantities:

$$\mathit{Perf}_{pre} = \max_{\theta} \mathbb{E}[\log P(Y|f_{\theta}^{pre}(\Phi_{pre}(X)))]$$

$$\mathit{Perf}_{post} = \max_{\theta} \mathbb{E}[\log P(Y|f_{\theta}^{post}(\Phi_{post}(X)))]$$

To determine if information loss occurs, we compute the difference in performance:

$$\Delta\mathit{Perf} = \mathit{Perf}_{pre} - \mathit{Perf}_{post}$$

If $\Delta\mathit{Perf} = 0$, this implies that no information loss has occurred and the information available in $\Phi_{pre}(X)$ is fully retained in $\Phi_{post}(X)$. However, if $\Delta\mathit{Perf} > 0$, this indicates that the post-projector representation has lost some information present in the pre-projector representation, leading to a decrease in predictive power for Y.

With the hypothesis grounded to V-information, we conduct a probing experiment on LLaVA-7B to verify it. We linear-probe the pre- and post-projector feature with image classification tasks on CIFAR10 (Krizhevsky et al., 2009), CIFAR100 (Krizhevsky et al.) and ImageNet (Deng et al., 2009). Results in Table 3 shows that for the 13B LLaVA model, performance percentage drop of post-projection features is less than 2%, indicating that the visual features are well preserved by the projectors in both models.

Table 3: Performance of linear probing using pre- and post-projector image features on CIFAR10, CIFAR100 and ImageNet. Accuracy% is used as the metric.

Dataset	*LLaVA-13B*	
	Perf_{pre}	Perf_{post}
CIFAR10	96.27	96.15 -0.12%
CIFAR100	81.78	81.02 -0.93%
ImageNet	71.97	70.83 -1.58%

Conjecture 4. The projector should align the visual and textual spaces. As its name suggests, the projector should be able to project the source (visual) space to the target (textual) space. To probe the alignment between two spaces, we collect caption data from MSCOCO (Lin et al., 2014), LLaVA-Caption (Liu et al., 2023d), ALLaVA (Chen et al., 2024a) and compute the similarity between a projected image feature and the textual embedding of its caption. The rationale of using cosine similarity is that, based on the findings in Section 2.2, a large performance boost is observed if we replace an image with its caption. Therefore, if the projected image feature is similar enough to its caption embedding (*i.e.* cosine similarity=1), then an LVLM should gain similar performance to the case where an image is replaced by its caption as input.

Results in Table 4 show that the cosine similarities of the two features are fairly low, indicating nearly independent relationships. This finding is consistent with the existing work (Huang et al., 2024b; Li et al., 2025), which reveals that visual and textual representations are apart from each other in the embedding space. Therefore, the

Table 4: Cosine similarity between projected image features and the caption embedding. Captions are processed by Vicuna (Chiang et al., 2023) tokenizer.

Dataset	Token Length	Image Res.	Cos. Sim. 7B	Cos. Sim. 13B
MSCOCO	15.16	(575, 488)	0.03	0.04
LLaVA Caption	15.09	(412, 366)	0.03	0.04
ALLaVA	222.83	(1020, 923)	0.05	0.06

projector in LLaVA models may not function as an alignment module as well as expected, which could be one of the causes of hallucination for the entire model.

3 Mitigating Object Hallucination Caused by Different Modules

Based on the analysis in Section 2, we further devised different methods to mitigate the object hallucination in different components in LVLMs.

3.1 How to alleviate the hallucination caused by CLIP?

As previously noted, the vision backbone within LVLMs also contributes to hallucinations. The CLIP model, as the vision encoder of LLaVA, is trained on massive image-caption pairs from the internet with a contrastive loss objective. However, these captions are typically brief and noisy, and negative pairs often differ substantially from positive ones. Therefore, it is likely that the model can distinguish them without needing to capture the finer details in the images. Consequently, the model may achieve high accuracy while lacking a nuanced understanding of the visual content (Liu et al., 2024b). To address this issue, we propose two methods to reduce hallucination caused by the vision backbone, as shown in Figure 2.

Tuning CLIP with fine-grained data A direct method to improve CLIP is to post-train CLIP with more fine-grained samples. This is because the CLIP is trained with massive images paired with brief captions. In this method, we leverage GPT-4 (OpenAI, 2022) to generate negative examples, which are then used in a contrastive learning setup to improve the discriminative ability of CLIP.

Generate Negative Examples: Inspired by prior work indicating that LVLMs are more likely to generate hallucinatory responses for frequently occurring objects (Liu et al., 2024b), we devise two strategies: inserting hallucinatory objects and removing existing ones.

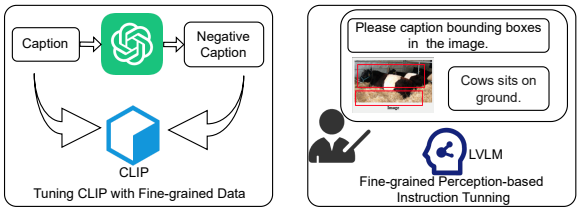

Figure 2: Tuning CLIP with fine-grained data (left) and fine-grained perception-based instruction tuning (right).

For the insertion strategy, we categorize objects in images into three types—random, popular, and adversarial—each containing three objects. Random objects are sampled randomly, popular objects are the top frequent objects in the whole dataset, and adversarial objects are the top frequent objects with the current objects. By inserting one to three objects from each category into the correct captions with the assistance of GPT-4, we create examples with varying levels of hallucinations (*i.e.,* negative samples). For the removal strategy, we randomly select one or two segmented objects from the caption and instruct GPT-4 to eliminate them from the caption.

Contrastive Learning: We use these generated negative examples in a contrastive learning framework where CLIP is trained to correctly distinguish between the positive and negative pairs. By exposing the model to these fine-grained differences, CLIP becomes better at understanding nuanced visual features.

First, let I represent an image embedding and T a text embedding. Let T^+ be the text vector that correctly matches I, and let T^- denote a collection of negative texts not semantically aligned with I. We also introduce β as a temperature parameter.

The fundamental image-to-text contrastive objective can be expressed as:

$$\mathcal{L}_{i2t} = -\log\left(\frac{\exp(I \cdot T^+/\beta)}{\sum_{T^* \in \{T^+, T^-\}} \exp(I \cdot T^*/\beta)}\right). \quad (1)$$

The symmetric term $\mathcal{L}_{t2i}$ can be constructed for text-to-image alignment. Combining them yields the image-text contrastive loss:

$$\mathcal{L}_{itc} = \frac{1}{2}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}). \quad (2)$$

Next, consider that we introduce an additional set of artificially generated negative texts $\{T^{neg}\}$. Incorporating these into the image-to-text objective

gives:

$$\mathcal{L}_{i2t} = -\log\left(\frac{\exp(I \cdot T^{+}/\beta)}{\sum_{T^{*}\in\{T^{+},T^{-},T^{neg}\}}\exp(I \cdot T^{*}/\beta)}\right). \quad (3)$$

To further refine the separation between correct matches and all classes of negative samples (both standard and synthetic), we introduce a margin-based term. Let τ_1 be the margin threshold enforcing that a positive pair's similarity should exceed that of any negative pair by at least τ_1:

$$\mathcal{L}_1 = \max\big(0,\ \tau_1 - (I \cdot T^{+}) + (I \cdot T^{\star})\big), \quad (4)$$

where $T^{\star} = \{T^{-}, T^{neg}\}$ is the union of standard and synthetic negatives.

Additionally, to encourage the model to distinguish synthetic negatives from standard negatives—thus capturing subtle semantic cues—we introduce another margin loss. Let τ_2 control the required margin between these two types of negative samples:

$$\mathcal{L}_2 = \max\big(0,\ \tau_2 - (I \cdot T^{neg}) + (I \cdot T^{-})\big). \quad (5)$$

Finally, assigning weighting factors λ_1 and λ_2 to the margin terms allows adaptive emphasis on these constraints. The complete objective function is:

$$\mathcal{L} = \mathcal{L}_{itc} + \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2. \quad (6)$$

This integrated loss framework guides the model to better discriminate correct image-text pairs from both standard and refined negative samples.

Fine-grained perception-based visual instruction tuning As we mentioned, CLIP may not capture the finer details in the visual representation from the vision encoder. Therefore, we attempt to enable the LLM to perceive the fine-grained information within the CLIP vision encoder. Meanwhile, the method of enhancing CLIP and then replacing it is time-consuming, as it requires additional steps for feature alignment and instruction tuning after replacing the vision encoder of LVLMs. As a result, we explore a more efficient approach by directly enabling the LLM to perceive the detailed visual features during visual instruction tuning.

To achieve this, we propose *fine-grained perception-based visual instruction tuning*. Specifically, we randomly select two bounding boxes from the image, and then use the object attributes corresponding to these bounding boxes and their relationships to generate the corresponding captions. We then create instruction tuning data (I_f, T_f, R_f), where T_f is the textual prompt: "Please caption the content in the bounding box", I_f is the image with bounding boxes, and R_f is the corresponding caption. This approach allows the model to perceive fine-grained information, such as region-level details, within the image.

3.2 How to reduce hallucination caused by the projector?

In Section 2.4, we reveal that hallucination introduced by the projector may be due to the inability of aligning visual and textual spaces, manifested by the low cosine similarity of caption embeddings and projected image features. Therefore, a straightforward remedy would be to explicitly bridge the image and caption representation during LLaVA's alignment stage.

3.2.1 Loss Objectives

Besides autoregressive image-text generation loss $\mathcal{L}_{itg} = -p(R|I,T)$, we introduce an in-batch contrastive alignment loss $\mathcal{L}_{itc}$ similar to Equation 2, where we maximize the similarity between a projected image feature and the corresponding text embedding for its caption. *We only focus on the alignment stage* and design three settings that involve the contrastive loss in different fashions.

Integrated Alignment Loss 🔥 The training process consists of two stages: *alignment* and *visual instruction tuning*. The contrastive loss is integrated to the *alignment* stage with a *learnable* (🔥) weight λ. The alignment objective is given by: $\min_{\mathcal{P},\lambda} \mathcal{L}_{itg} + \lambda\mathcal{L}_{itc}$. The visual instruction tuning stage is identical to LLaVA's.

Integrated Alignment Loss ❄️ All settings are the same as above except that the weight λ is *fixed* (❄️). The alignment objective is given by: $\min_{\mathcal{P}} \mathcal{L}_{itg} + \lambda\mathcal{L}_{itc}$.

Separate Contrastive Alignment Loss We prepend a *contrastive alignment stage* solely for the projector $\mathcal{P}$. Namely, the first stage objective is given by: $\min_{\mathcal{P}} \mathcal{L}_{itc}$. The second stage and third stage correspond to the original *autoregressive alignment* and *visual instruction tuning* stage.

4 Results and Analysis

We first introduce the benchmarks on which our methods to be evaluated, which are shown as follows. 1) Object-based benchmarks: testing the object perception of LVLMs. *POPE* and *POPE-NoCaps* (Liu et al., 2024b) are adopted, where the

Table 5: Performance of different methods across different benchmarks. The best results in each column are made **bold**. *w-ECLIP*: LLaVA with enhanced CLIP trained on fine-grained data; *w-FineIns*: LLaVA trained on fine-grained visual instruction tuning data.

Method	POPE						POPE-NoCaps						QA-VisualGenome			
	Random		Popular		Adversarial		Random		Popular		Adversarial		Attribute		Relation	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LLaVA-7B	85.40	86.36	86.63	85.25	85.13	83.88	84.80	82.97	79.40	78.30	74.77	74.69	64.67	66.60	67.57	74.81
w-ECLIP	**87.80**	**86.87**	**87.30**	**86.04**	**85.87**	**84.70**	85.27	83.50	81.00	79.69	75.77	75.46	67.67	68.79	67.00	74.11
w-FineIns	87.77	86.78	86.80	85.51	85.53	84.33	**85.53**	**84.00**	**81.73**	**80.61**	**76.50**	**76.37**	**69.01**	**70.12**	**69.75**	**76.17**

Table 6: Performance of different projector alignment methods across different benchmarks. The best results in each column are made **bold**. *Int. Align.*: Integrated Alignment Loss with trainable (🔥) / frozen(❄) weighting parameter; *Sep. Ctrs. Align.*: Separate Contrastive Alignment Loss.

Method	POPE						POPE-NoCaps						QA-VisualGenome			
	Random		Popular		Adversarial		Random		Popular		Adversarial		Attribute		Relation	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
LLaVA-7B	87.42	86.36	86.63	85.25	**85.13**	**83.88**	84.80	82.97	79.40	78.30	74.77	74.69	**64.67**	**66.60**	67.57	74.81
Int. Align. 🔥	88.21	87.41	86.70	85.65	84.27	83.46	**85.57**	**84.46**	77.27	77.58	72.23	73.91	60.95	61.97	66.67	74.60
Int. Align. ❄	88.04	87.20	86.67	85.56	84.50	83.60	84.90	83.28	79.37	78.47	74.57	74.76	63.84	65.21	66.73	74.26
Sep. Ctrs. Align.	**88.56**	**87.86**	**87.33**	**86.38**	84.57	**83.88**	**85.57**	84.24	**80.07**	**79.42**	**75.13**	**75.54**	64.26	64.77	**69.60**	**76.06**

latter is built on NoCaps (Agrawal et al., 2019) following a similar manner as in POPE. 2) Attribute- and relation-based benchmark: *QA-VisualGenome* is constructed and adopted (detailed in Sec. 2.1). We provide an in-depth analysis of our methods for improving the vision encoder and the projector. We call object-, attribute- and relation-based benchmarks as perception-based benchmarks.

For a fair comparison, we only use the LLaVA-Caption dataset for alignment. All experiments are conducted on 4*A100 GPUs. For the *alignment* stage, we set per-GPU batch size to 64, which is also the batch size contrastive alignment. We choose the well-known LLaVA-v1.5-7B model as our baseline. All three settings introduce no extra learnable parameters (except for the weighting parameter λ in **Integrated Alignment Loss** 🔥 setting). Under our setting, both the *original* and *integrated alignment stage* take 6 hours, and *visual instruction tuning stage* takes 24 hours. Notably, the prepended *contrastive alignment stage* takes only 12 minutes to train since only the vision encoder $\mathcal{V}$, projector $\mathcal{P}$ and the embedding layer of LLM $\mathcal{D}$ are involved in the forward process. For the two integrated loss settings, we empirically initialize λ with 5, make it learnable for 🔥 while keep it fixed for ❄. λ_1 and λ_2 are set to 1.

Can our methods reduce hallucination caused by the vision encoder? Table 5 presents the comprehensive experimental results of various settings across different testing benchmarks. From this table, several key observations can be drawn: **1) Our proposed *w-ECLIP* method demonstrates superior performance compared to LLaVA-7B on perception-based benchmarks**. This result underscores the effectiveness of our approach in reducing visual object hallucinations by enhancing the fine-grained perception capabilities of CLIP. **2) *w-FineIns* exhibits better performance than baseline on perception-based benchmarks**. This finding suggests that our fine-grained instruction data can augment the fine-grained perception abilities of LLaVA by leveraging region-level captions during training. **3) Compared to *w-FineIns*, *w-ECLIP* demonstrates comparable or even better performance on perception-based benchmarks**. Notably, *w-FineIns* offers efficiency advantages as it only requires the final training stage—instruction tuning—for the LVLM, simplifying the overall training process.

Can our methods reduce hallucination caused by the projector? We benchmark our methods in Table 6. For object-oriented benchmarks POPE and POPE-NoCaps, the model trained with *Separate Contrastive Alignment Loss* outperforms others on most splits of benchmarks, though the improvement over baseline seems marginal. For QA-VisualGenome benchmark, we only observe improvement on the "Relation" split with *Separate Contrastive Alignment Loss*, whereas slight performance drops are observed for others. These observations provide insights for the alignment process.

Table 7: Performance on QA-FB15K.

Method	Entity		Relation	
	Acc	F1	Acc	F1
LLaVA-7B	78.39	73.14	56.79	48.79
Int. Align. 🔥	**84.28**	**83.03**	59.16	**58.07**
Int. Align. ❄️	84.05	81.76	59.16	56.97
Sep. Ctrs. Align.	83.94	81.65	**59.39**	57.41
LLaVA-7B	78.39	73.14	56.79	48.70
w-ECLIP	77.60	71.47	56.79	45.58
w-FineIns	76.47	69.86	55.45	49.10

Firstly, **object hallucinations may not be directly related to alignment in LVLM**, where vision encoder is mostly responsible for the perception process. Secondly, **perception-based attribute and relation hallucination can hardly be mitigated by contrastive training of projector**. Similar to object hallucination, better visual representations may be needed as a remedy.

Can our method influence other hallucinations? To further investigate the influence of our method on other kinds of hallucination, we introduced the Cognition-based benchmark: necessitating world knowledge in LVLMs for problem solving. We construct a cognition-based benchmark *QA-FB15k* based on the knowledge graph FB15K (Bordes et al., 2013). We show the results in Table 7. Contrastive alignment objective is beneficial for cognition-based knowledge, as evidenced by the performance boost on QA-FB15K. By better aligning between vision encoder and LLM, the LVLM is able to leverage the ability of LLM to answer the question that requires world knowledge, which is typically stored in LLMs pretrained on mountains of data. Nevertheless, performance boosts are found on QA-FB15K for all three settings over baselines. Neither w-FineIns nor w-ECLIP shows any improvement on the cognition-based benchmark. This may be attributed to the fact that, unlike perception-based benchmarks, cognition-based benchmarks necessitate not only the ability to identify objects but also the comprehension and application of relevant associated knowledge. The two methods primarily focus on improving perception, may not cater for the knowledge-intensive requirements of cognition-based benchmarks.

More Analysis: In addition, we add more experimental results on the hallucination benchmark and general benchmark, ablation study, and performance comparison with more baselines in Appendix B, E, F, and D.

5 Related Work

Large Vision-Language Model. The multimodal learning field has recently pivoted its focus towards Large Vision-Language Models (LVLMs) (Awadalla et al., 2023; Li et al., 2023a). Current advanced LVLMs primarily comprise three essential components: a language encoder, a visual encoder, and a cross-modal alignment mechanism (Rohrbach et al., 2018). To achieve comprehensive visual understanding, LVLMs generally undergo a series of training stages (Gong et al., 2023; Zhu et al., 2023; Liu et al., 2023d,e; Ye et al., 2023; Dai et al., 2023; Liu et al., 2023e). Despite significant advancements, LVLMs still face challenges with hallucination, which significantly affects performance across various multimodal applications.

Hallucinations in Large Vision-language Models. Since hallucination issues and mitigation techniques have been extensively explored in text generation (Ji et al., 2023; Min et al., 2023), research on hallucinations in LVLMs (Dai et al., 2023; Liu et al., 2023e; Jing and Du, 2024) attracts more attention. To evaluate the hallucination in the LVLMs, several researchers propose metrics and benchmarks (Rohrbach et al., 2018; Li et al., 2023e; Lovenia et al., 2023; Lu et al., 2023; Jing et al., 2024). Recently, various methods have been proposed to mitigate hallucinations in LVLMs, leveraging a range of techniques including decoding strategies (Leng et al., 2023; Huang et al., 2023), post-processing methods (Zhou et al., 2023; Chang et al., 2024; Yin et al., 2023), the development of higher-quality datasets (Liu et al., 2023c; Li et al., 2023d), and modality alignment(Li et al., 2023c; Yu et al., 2023; Zhou et al., 2024; Jing and Du, 2024; Sun et al., 2023a; Gunjal et al., 2023). Despite the success of the existing works, there lacks a comprehensive study of what causes visual hallucinations in LVLMs.

6 Conclusion

Our study delves into the visual hallucination problem in LVLMs, identifying its sources within the model's components. By independently analyzing the LLM, vision backbone, and projector, we propose targeted mitigation strategies. We introduce fine-grained hallucination benchmarks, QA-VisualGenome and QA-FB15k, to comprehensively evaluate hallucinations. Our methods demonstrate effectiveness in reducing hallucinations, contributing to the reliability and accuracy of LVLMs.

Limitations

Our work primarily focuses on analyzing and improving hallucinations of general objects, such as tables and people, while neglecting the research topic of how to mitigate cognition-level hallucinations, such as the names of individuals and famous buildings.

References

Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956. IEEE.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *IEEE International Conference on Computer Vision*, pages 2425–2433. IEEE Computer Society.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *CoRR*, abs/2308.01390.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Yue Chang, Liqiang Jing, Xiaopeng Zhang, and Yue Zhang. 2024. A unified hallucination mitigation framework for large vision-language models. *CoRR*, abs/2409.16494.

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*.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024b. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXI*, volume 15139 of *Lecture Notes in Computer Science*, pages 19–35. Springer.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024c. Halc: Object hallucination reduction via adaptive focal-contrast decoding.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. vicuna: An open-source chatbot impressing gpt-4 with 90

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias

Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *CoRR*, abs/2305.04790.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D Manning. 2021. Conditional probing: measuring usable information beyond a baseline. *arXiv preprint arXiv:2109.09234*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *CoRR*, abs/2311.17911.

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

Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024b. Deciphering cross-modal alignment in large vision-language models with modality integration rate. *arXiv preprint arXiv:2410.07167*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Liqiang Jing and Xinya Du. 2024. FGAIF: aligning large vision-language models with fine-grained AI feedback. *CoRR*, abs/2404.05046.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research).

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. *CoRR*, abs/2311.16922.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *CoRR*, abs/2305.03726.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023c. Silkie: Preference distillation for large visual language models. *CoRR*, abs/2312.10665.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023d. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *CoRR*, abs/2306.04387.

Qing Li, Jiahui Geng, Derui Zhu, Zongxiong Chen, Kun Song, Lei Ma, and Fakhri Karray. 2025. Internal activation revision: Safeguarding vision language models without parameter update. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 27428–27436. AAAI Press.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023e. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *ECCV*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023c. Aligning large multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023e. Visual instruction tuning. *CoRR*, abs/2304.08485.

Yufang Liu, Tao Ji, Changzhi Sun, Yuanbin Wu, and Aimin Zhou. 2024b. Investigating and mitigating object hallucinations in pretrained vision-language (CLIP) models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 18288–18301. Association for Computational Linguistics.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models.

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. *CoRR*, abs/2309.04041.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *CoRR*, abs/2305.14251.

OpenAI. 2022. Chatgpt blog post.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *EMNLP*, pages 4035–4045. ACL.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023a. Aligning large multimodal models with factually augmented RLHF. *CoRR*, abs/2309.14525.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023b. Aligning large multimodal models with factually augmented RLHF. *CoRR*, abs/2309.14525.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, abs/2311.07397.

Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. 2020. Visual commonsense R-CNN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10757–10767. Computer Vision Foundation / IEEE.

Jinfeng Wei and Xiaofeng Zhang. 2024. DOPRA: decoding over-accumulation penalization and reallocation in specific weighting layer. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 7065–7074. ACM.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024. Mitigating object hallucination via concentric causal attention. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *CoRR*, abs/2310.16045.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2023. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *CoRR*, abs/2312.00849.

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an

EOS decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11766–11781. Association for Computational Linguistics.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2024a. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *CoRR*, abs/2411.09968.

Yue Zhang, Liqiang Jing, and Vibhav Gogate. 2025. Defeasible visual entailment: Benchmark, evaluator, and reward-driven optimization. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 25976–25984. AAAI Press.

Yue Zhang, Jingxuan Zuo, and Liqiang Jing. 2024b. Fine-grained and explainable factuality evaluation for multimodal summarization. *arXiv preprint arXiv:2402.11414*.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *CoRR*, abs/2402.11411.

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. *CoRR*, abs/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. *CoRR*, abs/2304.10592.

A Hallucinations in Different Components

We show the potential hallucinations of each component of LVLMs, and the corresponding mitigation methods in Table 8

B More Experiments on Hallucination Benchmark

We further add experiments on another hallucination benchmark, Amber. The experimental results of Table 9 show the effectiveness of our method.

C Case Study

We showed some hallucinated examples in Figure 3. We can see that the hallucination caused by CLIP can be further input to the LVLM, causing the hallucination in the LVLM.

Image

Question: Is there a bed in the image?

LLaVA output: There is a bed in the image.

Text 1: a bird is in the image.
CLIP probobility: 0.1848

Text 2: a bed is in the image.
CLIP probobility: 0.8154

Figure 3: The illustration of the hallucinated case for CLIP and LLaVA.

D Comparison with the Existing Hallucination Mitigation Method

To verify the effectiveness or our methods, we further add more baselines on POPE, as shown in

Component	Hallucination?	Mitigation
Vision Backbone	✓	w-ECLIP & w-FineIns
Projector	✓	Int. Align. 🔥 & Int. Align. ❄ & Sep. Ctrs. Align.
LLM	✗	N/A

Table 8: Illustration of potential hallucinations in the components of LVLMs, and the corresponding mitigation methods

Table 10. From this table, our methods show competitive performance with the best baseline (i.e., Less is more). This further demonstrates the effectiveness of our method.

E Experiment on General Benchmark

To verify the impact of the proposed method on general capabilities, we further conduct experiments on the general benchmark LLaVA-Bench (Liu et al., 2023e). The results of Table 11 show the effectiveness of our method.

F Ablation Study

In this section, we conduct ablation experiments to assess the contribution of each component in the loss function by individually removing the weights λ_1 and λ_2. The results are shown in the Table 12. These results demonstrate that both components play meaningful roles in enhancing model performance.

Dataset	LLaVA-7B	w-ECLIP	w-FineIns	Int. Align. 🔥	Int. Align. ❄	Sep. Ctrs. Align.
Existence	83	**93**	92	88	91	87
Attribute	64	**81**	**81**	75	78	76
Relation	65	69	**70**	57	62	59
All	71	73	**81**	73	77	74

Table 9: Performance on the Amber dataset across different model variants. Bold indicates best scores per row.

Method	F1 Score
DoLa (Chuang et al., 2024)	80.2
ITT (Li et al., 2023b)	83.7
VCD (Leng et al., 2023)	83.2
AGLA (An et al., 2025)	84.6
OPERA (Huang et al., 2024a)	85.2
DOPRA (Wei and Zhang, 2024)	85.6
HALC (Chen et al., 2024c)	83.9
FastV (Chen et al., 2024b)	81.3
Less is more (Yue et al., 2024)	86.0
CCA-LLAVA (Xing et al., 2024)	85.5
LRV (Liu et al., 2024a)	80.0
Amber (Wang et al., 2023)	81.6
EAH (Zhang et al., 2024a)	85.7
w-ECLIP	85.9
w-FineIns	85.5
Int. Align. 🔥	85.5
Int. Align. ❄	85.5
Sep. Ctrs. Align	**86.0**

Table 10: POPE F1 scores for baselines and proposed methods. Bold indicates the highest score.

Model	Conv	Detail	Complex	Full
LLaVA-7B	92	75	75	81
w-ECLIP	93	84	87	88
w-FineIns	94	86	86	89
Int. Align. 🔥	95	87	83	89
Int. Align. ❄	93	84	82	86
Sep. Ctrs. Align	99	85	87	90

Table 11: Model performance comparison on different categories and the full set on LLaVA-Bench.

Method	POPE					
	Random		Popular		Adversarial	
	Acc	F1	Acc	F1	Acc	F1
w-ECLIP	**87.80**	**86.87**	**87.30**	**86.04**	**85.87**	**84.70**
$\lambda_1 = 0$	87.50	86.38	86.93	85.84	85.62	83.97
$\lambda_2 = 0$	87.52	86.47	86.79	85.88	85.47	84.11

Table 12: Ablation study on the impact of loss function components λ_1 and λ_2 across different POPE test subsets.