What if...?: Thinking Counterfactual Keywords Helps to Mitigate Hallucination in Large Multi-modal Models

Junho Kim Yeonju Kim Yong Man Ro*

Integrated Vision and Language Lab, KAIST
{arkimjh, yeonju7.kim, ymro}@kaist.ac.kr

Abstract

This paper presents a way of enhancing the reliability of Large Multi-modal Models (LMMs) in addressing hallucination, where the models generate cross-modal inconsistent responses. Without additional training, we propose Counterfactual Inception, a novel method that implants counterfactual thinking into LMMs using self-generated counterfactual keywords. Our method is grounded in the concept of counterfactual thinking, a cognitive process where human considers alternative realities, enabling more extensive context exploration. Bridging the human cognition mechanism into LMMs, we aim for the models to engage with and generate responses that span a wider contextual scene understanding, mitigating hallucinatory outputs. We further introduce Plausibility Verification Process (PVP), a simple yet robust keyword constraint that effectively filters out suboptimal keywords to enable the consistent triggering of counterfactual thinking in the model responses. Comprehensive analyses across various LMMs, including both open-source and proprietary models, corroborate that counterfactual thinking significantly reduces hallucination and helps to broaden contextual understanding based on true visual clues.

1 Introduction

After witnessing the great success of Large Language Models (LLMs) products, such as ChatGPT [OpenAI, 2023a] and Gemini [Google, 2023], the emergence of Large Multi-modal Models (LMMs) naturally followed as the next step towards a unified, general-purpose AI system [OpenAI, 2024; xAI, 2024; Reid et al., 2024]. In the vision research area, various works [Li et al., 2022, 2023; Zhu et al., 2023] have actively resorted LLMs into the vision models due to their remarkable capability of off-the-shelf text generation. Especially when it comes to in-context learning [Brown et al., 2020; Alayrac et al., 2022], prompt engineering [Zhou et al., 2022; Bsharat et al., 2023], and chain-of-thought [Wei et al., 2022; Kojima et al., 2022; Zhang et al., 2023], vision models can exploit the generation power into the various vision tasks such as visual understanding and reasoning [Yu et al., 2022; Huang et al., 2024].

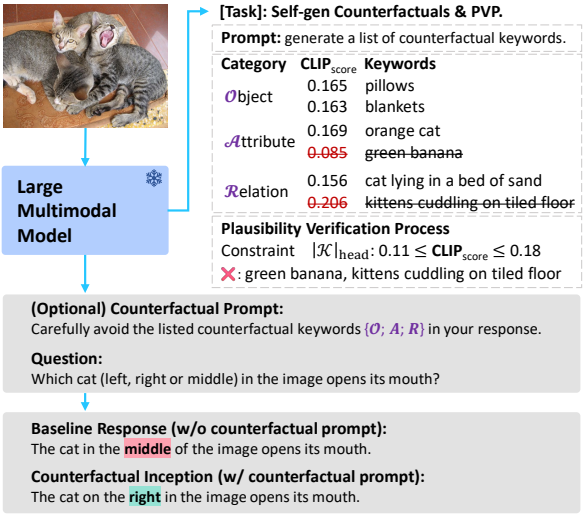


Figure 1: Counterfactual Inception: LMMs generate counterfactual keywords at the object, attribute, and relation levels, then integrate them with a counterfactual prompt to implant counterfactual thinking to the models. To filter out keywords that are either too similar or too deviated from the visual content, we adopt a robust constraint called PVP.

Although the recent breakthroughs of multi-modal instruction tuning approaches [Dai et al., 2023; Liu et al., 2023c] unlock enhanced visual proficiency by aligning model responses with human-specific instructions, LMMs still struggle with unexpected hallucination in their responses [Liu et al., 2023a; Zhou et al., 2024]. The hallucination in LMMs involve false premises, where the models generate incorrect, nonsensical, or unrelated responses for the visual contents. To al-

*Corresponding author.
https://ivy-lvlm.github.io/Counterfactual-Inception

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10672–10689
November 12-16, 2024

leviate the hallucination in LMMs, recent studies have been proposed in the context of curated instruction-tuning [Liu et al., 2023a; Wang et al., 2023], or integrating visual information using external solvers [Wang et al., 2024; Yin et al., 2023; Zhou et al., 2024]. However, they require additional training on the tailored instruction or labor-intensive resources to fine-tune the models [Sun et al., 2024; Yu et al., 2023]. To step out such limitations and reduce hallucination in a training-free manner, we present a novel way of eliciting an exceptionality capability from LMMs by engaging them to consider alternative counterfactuals.

In our daily life, we ponder *what if...?* scenarios at least once in awhile— these sorts of thoughts can be termed as *counterfactual* that is contrary to what actually happened [Menzies and Beebee, 2001; Epstude and Roese, 2008]. By thinking of how events might have unfolded differently if we had taken alternative actions (or even seemingly irrelevant thinking), we can enhance cognitive flexibility in the present and identify more about what happens now [Roese, 1997]. Motivated by such human tendency, we delve into the following question: *"Can we elicit counterfactual thinking from LMMs by imagining what-if scenarios and mitigate hallucination in their responses?"*.

Building on the concept of counterfactuals, we propose Counterfactual Inception, a novel method of implanting counterfactual thinking into LMMs using inconsistent keywords against given visual contents. In our work, we expose LMMs to self-generated counterfactual priors and examine their contextual flexibility in generating responses. Such approach not only allows LMMs to explore a wide range of potential answers but also promotes broader contextual exploration and the consideration of hypothetical narratives. Our findings demonstrate that this thinking enhances the model's ability to engage with and generate responses that spans a wider spectrum of visual understanding, effectively reducing hallucinatory outputs.

Specifically, as illustrated in Fig. 1, we instruct LMMs themselves to generate counterfactual keywords at the object-, attribute-, and relation-levels for the visual contents. These keywords are then incorporated into the conditional response generation for user queries with a counterfactual prompt. To consistently promote LMMs to engage in counterfactual thinking, the key challenge is on the optimal selection of counterfactual keywords in triggering the exceptional thought. Accordingly, we present Plausibility Verification Process (PVP), a robust constraint designed to filter out the sub-optimal keywords based on CLIP [Radford et al., 2021] alignment between the visual contents and their counterfactual keywords. Through extensive analyses on recent LMMs including open-source [Liu et al., 2023b; Dong et al., 2024; Liu et al., 2024b; Chen et al., 2024b] and proprietary models [Google, 2023; OpenAI, 2023c], we corroborate that Counterfactual Inception helps to alleviate hallucination in general across various benchmarks.

Our contributions can be summarized as follows: (i) we introduce Counterfactual Inception, a novel method that prompts counterfactual thinking into LMMs using deliberately deviated language keywords to mitigate hallucination, (ii) we present Plausible Verification Process (PVP), a robust constraint designed to refine the selection of counterfactual keywords, ensuring the optimal trigger of counterfactual thinking in LMMs. (iii) Through extensive experiments and analyses on various LMMs, including both open-source and proprietary models, we demonstrate that Counterfactual Inception effectively enhances reliability of model responses across diverse benchmarks.

2 Related Work

V+L: Large Multi-modal Models. The release of open-sourced LLMs [Touvron et al., 2023; Chiang et al., 2023] has spurred active research towards more generalized integration, especially vision-language (VL) modalities. By using the language models as linguistic channels, LMMs can integrate visual information into broader VL understanding tasks [Yang et al., 2022; Lu et al., 2023]. After the surge of VL learning [Li et al., 2021, 2022; Yu et al., 2022] facilitated cross-modal alignment, recent approach in LMMs is adopting visual instruction-tuning [Dai et al., 2023; Liu et al., 2023c; Dong et al., 2024; Chen et al., 2024b] on various datasets. LLaVA series [Liu et al., 2023c,b, 2024b] have paved the way for building multi-modality systems that can freely interact with users' instructions. Along with such paradigm, a wide range of advanced architectures and adaptations to specific domains [Lin et al., 2023; Li et al., 2024] have actively explored. Additionally, numerous proprietary LMMs are expanding their capabilities into multi-modal tasks, by releasing advanced products such as Gemini 1.5 [Reid et al., 2024], and GPT-4o [OpenAI, 2024], which allow users to interact

with the models through multi-modal channels.

Hallucination in Large Multi-modal Models. Despite the remarkable advancements of LMMs, the major issue of hallucination still persists in their responses. Hallucination refers to the phenomenon where generated texts are inconsistent with the visual contents, one of the long-standing challenges in image captioning [Rohrbach et al., 2018]. When it comes to LMMs, this problem can be worse due to their use of the expressive capabilities of LLMs, which enable more detailed and rich descriptions [Jing et al., 2023]. As their representation becomes abundant, the complexity of hallucinations also increases, leading to a multifaceted issue. This includes challenges: (i) the scarcity of large-scale image-text instruction pairs [Liu et al., 2023a], and (ii) the entropic gap between visual and textual data [Sun et al., 2024], which can be exacerbated during alignment pre-training.

Recent works have explored various ways to mitigate hallucination, including fine-tuning LMMs with robust instructions [Liu et al., 2023a; Wang et al., 2023], implementing multi-step LMM-aided reasoning [Wang et al., 2024; Yin et al., 2023; Zhou et al., 2024; Chen et al., 2024a], utilizing RLHF [Sun et al., 2024; Yu et al., 2023] for providing human feedback instructions, and deploying contrastive decoding in the inference phase of LMMs [Leng et al., 2023; Woo et al., 2024; Kim et al., 2024]. More recent hallucination survey compilation can be found in [Liu et al., 2024a; Bai et al., 2024b]. Our counterfactual method provides a novel approach to reducing hallucinations in LMMs by conditioning them on exceptional thought through counterfactual keywords. We emphasize that our method can achieve significant mitigation of hallucinatory responses without additional fine-tuning or human-resource instructions.

3 Proposed Method

3.1 Counterfactual Keyword Generation

Firstly, we can generally categorize the hallucinatory types into three distinct groups— nonexistent objects, incorrect object attributes, and erroneous object relations, as found in previous research [Liu et al., 2024a; Bai et al., 2024b]. To mitigate the hallucination in the model response, our approach is implanting counterfactual thinking into LLMs by harnessing counterfactual keywords. These keywords intentionally do not describe what is visibly apparent but instead provide hypothetical contexts.

Example of Counterfactual Inception:	
	
User	Is the jam on the bread made of strawberry?
GT	In fact, this photo only shows a cut and toasted bread slice. There is no jam on it.
Intern-VL 1.5	Based on the image, the jam on the bread has a deep red color, which is commonly associated with strawberry jam. However, without a clearer view or additional information, it is not possible to definitively confirm the flavor of the jam.
Counterfactual Inception	**Counterfactual Prompt:** Carefully avoid the listed Counterfactual Keywords in your response.
	Counterfactual Keywords: $\mathcal{O}$: cookie $\mathcal{A}$: a piece of burnt crusty french baguette $\mathcal{R}$: toaster on the table
Intern-VL 1.5	The image does not show any jam on the bread, so I cannot confirm if it is made of strawberry or any other fruit.

Table 1: Example of Counterfactual Inception using a baseline Intern-VL 1.5 [Chen et al., 2024b].

Importantly, they serve as primary anchors for the contextual exploration for better understanding of true visual clues. Therefore, we concretize counterfactual categories into trinary taxonomy, which can serve plausible alternatives for the visual contents:

- **Object Substitution:** replacing an object in the image with another that could logically occupy the same space but alters the scene's context.
- **Attribute Modification:** changing an object's color, size, or shape in a way that makes sense visually but leads to a different interpretation.
- **Relational Changes:** adjusting the spatial or interactional relationships between objects to suggest a different narrative within the scene.

Following tailored criteria ($\mathcal{O}$: object, $\mathcal{A}$: attribute, and $\mathcal{R}$: relation), we instruct LMMs themselves to generate three different categorical keywords for the given images, providing plausible but misleading interpretations of the visual contents. Here, obtaining counterfactual keyword is a challenging and complex task for LMMs. Accordingly, we first manually generate a few examples for in-context learning, then design a structured

prompt with these seed examples to generate keywords for the categories: $\mathcal{O}=\{o_i\}_{i=0}^{N_o}$, $\mathcal{A}=\{a_i\}_{i=0}^{N_a}$, and $\mathcal{R}=\{r_i\}_{i=0}^{N_r}$, where N_o, N_a, and N_r represent the different numbers of keywords in each category. We illustrate detailed keyword generation prompts in Table 6. Please see Appendix B.2 for the further explanation of keyword generation.

3.2 Counterfactual Inception

After generating the keywords, we implant the counterfactual keywords into LMMs as conditional prior information to guide model responses that disregard these inputs in the generation phase. Specifically, for the given LMMs M_θ, parameterized with θ, our objective is generating output sequences $y_{<t+1}=[y_1, y_2, \ldots, y_t]$ with given visual content v and textual query q. When incorporating self-generated counterfactual keywords to the models, we concatenate all of the keywords generated from a given image into a single list $k=[\mathcal{O};\mathcal{A};\mathcal{R}]\in\mathbb{R}^{|\mathcal{K}|}$, where $\mathcal{K}$ denotes whole counterfactual keywords set. After that, utilizing these keywords as conditional prior, we can formulate auto-regressive responses of LMMs as follows:

$$p_\theta(y|v,q,k)=\prod_{t=1}^{T} p_\theta(y_t|v,q,k,y_{<t}). \quad (1)$$

Note that our method can be adapted to existing LMMs in a training-free manner with a specific counterfactual prompt (see Table 7). As exemplified in Table 1, we prompt the models to carefully disregard the self-generated counterfactual keywords during their response generation for the user textual query (please see details in algorithm 1).

In other words, our method explicitly signal the models to consider alternative explanations anchoring from the self-generated counterfactual keywords. Consequently, our counterfactual approach not only promotes broader contextual understanding but also enhances reliability of the model response. It enables LMMs to focus on true visual clues within the context by incorporating counterfactual information into the response generation, which helps to mitigate hallucination.

3.3 $\mathcal{K}_{\text{head}}$: Plausibility Verification Process

Even when we instruct the models to generate keywords, they may not always fulfill our counterfactual intentions— for example, even with specific instruction, they might produce completely nonsensical keywords that are irrelevant to the visual content, or generate keywords that are closer to factual rather than counterfactual. Therefore, the key challenge lies in finding the optimal counterfactual keywords $k^*=\mathcal{K}_{\text{head}}(k)$ that trigger the counterfactual thinking. To analyze the keywords, we randomly sample 500 images from COCO [Chen et al., 2015] and extract counterfactual keywords from 6 baselines, totaling 3000 instances and approximately 10K ($\mathcal{O}$), 9.5K ($\mathcal{A}$), and 9.5K ($\mathcal{R}$) keywords in each category, respectively.

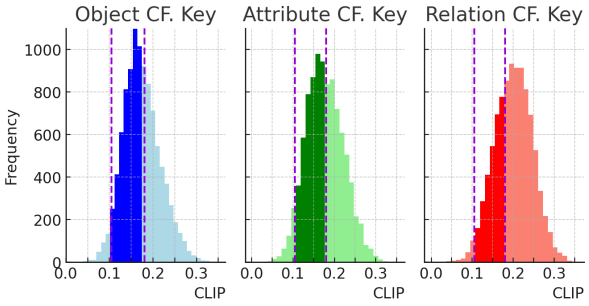


Figure 2: Frequency distribution for the counterfactual keywords. The dashed lines indicate truncation level. We have empirically observed that the keywords in the upper half of the distribution are closer to factual information rather than counterfactual, thus the lower half, excluding extreme low, is set as the criteria. See Fig. 6 for the keyword analysis.

To measure semantic alignment between the counterfactual keywords and visual contents, we employ CLIP [Radford et al., 2021] and delve into the cross-modal similarity for the text-image pairs. As in Fig. 2, the counterfactual keywords, while not directly descriptive, still touch upon concepts or contexts loosely related to the visual contents, leading to a wide range of medium to low scores. Following central limit theorem, the semantic space covered by the keywords has inherent symmetry around a mean value, with fewer keywords being extremely poorly or highly related, creating the bell curve typical of a normal distribution.

Regarding higher CLIP score suggests a better match— that is, the text more accurately or relevantly describes the image, we truncate the counterfactual keyword set based on the score, such that $\mathcal{K}_{\text{head}}(k)=\{k\in\mathcal{K}:\lambda_{\text{bot}}\leq\textbf{CLIP}(v,k)\leq\lambda_{\text{top}}\}$. As in the dashed lines in Fig 2, we empirically set the truncation hyperparameter to the lower half of the distribution, but not at the extreme low end, which aligns with the definition of a counterfactual keyword— meaningful, yet not direct, alternatives to the visible content. Further analysis in Sec. 4.4.

Model	#param	POPE				MMVP									
		Acc (↑)	Prec	Rec	F1 (↑)										Avg (↑)
Open-source Models															
LLaVA-1.5	13B	84.07	90.88	75.73	82.62	22.2	50.0	23.1	20.0	40.0	60.0	36.4	37.5	16.7	35.33
+ Ours		**85.03**	93.61	75.20	**83.40**	22.2	50.0	30.1	10.0	60.0	70.0	40.1	25.0	16.7	**39.33**
IXC2-VL	7B	84.13	83.12	85.67	84.37	11.1	53.3	30.8	50.0	35.0	60.0	27.3	37.5	16.7	36.00
+ Ours		**87.50**	94.61	79.53	**86.42**	22.2	60.0	42.3	40.0	25.0	70.0	36.4	50.0	50.0	**42.67**
LLaVA-NeXT	34B	**86.50**	83.86	90.40	87.01	16.7	60.0	38.5	30.0	35.0	80.0	40.9	37.5	0.0	40.67
+ Ours		85.63	79.35	96.33	**87.02**	33.3	63.3	46.2	40.0	45.0	60.0	40.9	25.0	0.0	**44.67**
InternVL 1.5	26B	85.83	82.83	90.40	86.45	27.8	76.7	46.2	30.0	45.0	80.0	36.4	25.0	33.3	48.00
+ Ours		**89.50**	92.11	86.40	**89.16**	33.3	73.3	61.5	40.0	50.0	60.0	36.4	25.0	50.0	**51.33**
Proprietary Models															
Gemini 1.5 Pro	N/A	80.70	85.78	73.60	79.22	27.8	53.3	38.5	40.0	55.0	40.0	45.5	62.5	66.7	46.00
+ Ours		**84.09**	77.78	95.45	**85.71**	55.6	56.7	34.6	40.0	45.0	50.0	50.0	50.0	66.7	**48.67**
GPT-4V	N/A	82.70	85.50	78.80	82.00	38.9	50.0	38.5	40.0	30.0	70.0	36.4	62.5	66.7	44.00
+ Ours		**85.50**	87.60	82.60	**85.07**	50.0	45.5	50.0	37.5	50.0	53.3	66.7	80.0	25.0	**48.67**

Table 2: Evaluation results on discriminative benchmarks. We focus on the most challenging category *adversarial* for POPE [Li et al.]. The each column symbol in MMVP [Tong et al., 2024] indicates 9 different visual patterns. We refer Appendix. A for subset details.

4 Experiments

4.1 Experimental Setup

Baselines & Implementation. We adopted recent high-performing 6 LMMs as our baseline models, which can be categorized into open-/closed-source: (i) open-source: LLaVA-1.5 (13B) [Liu et al., 2023b], InternLM-XComposer2 (7B) [Dong et al., 2024], LLaVA-NeXT (34B) [Liu et al., 2024b], InternVL 1.5 (26B) [Chen et al., 2024b] and (ii) proprietary models: Gemini 1.5 Pro [Reid et al., 2024] and GPT-4V [OpenAI, 2023c]

For generating counterfactual keyword set $\mathcal{K}$ from each model, we equally used same prompt format in Table 6, but with different guidelines and seed examples. To configure the settings for PVP, CLIP-ViT-L [Radford et al., 2021] is employed to measure CLIP score (cosine similarity) for the visual contents and the generated counterfactual keyword pairs. We set CLIP score truncation to 0.11 for lower and 0.18 for upper boundary.

Benchmarks and Evaluation Metrics. To assess hallucination in LMMs, benchmarks can be sorted into two types: (i) hallucination discrimination, which involves selecting the correct answers from multiple choices, and (ii) non-hallucinatory generation, testing the broader range of hallucinations in model responses, measured by either rule-based or GPT-aided methods [OpenAI, 2023b]. In our experiments, key evaluation benchmarks include POPE [Li et al.] and MMVP [Tong et al., 2024] for hallucination discrimination, and CHAIR [Rohrbach et al., 2018] and MMHal-Bench [Sun et al., 2024] for non-hallucinatory generation (Please see details in Appendix A):

- **POPE** uses 9K image-question pairs from COCO dataset to detect object hallucinations. We exclusively focus on the most challenging, *adversarial* setting. Evaluation metrics are accuracy, precision, recall, and F1-score.
- **MMVP** measures accuracy for CLIP-blind pairs, which have similar CLIP score but vary visually (300 instances & 9 visual patterns). Each pattern has curated questions with two response options and scores only if the models identify both pairs.
- **CHAIR** evaluates the proportion of hallucinatory objects in the model responses relative to the total number of objects in the true image caption. It consists of two metric variations: per-sentence and per-instance proportion.
- **MMHal-Bench** assesses descriptive score and hallucination severity in the model responses using GPT-4 with distinct eight question types. The metric ranges from 0 to 7 for the overall score, and the hallucination rate (%).

4.2 Counterfactual Keyword Statistics

As in Sec. 3.1, we first instruct the LMMs themselves to perform the counterfactual keyword generation task and adopt PVP constraint to filter out sub-optimal keywords. For 6 baselines and 4 benchmarks we have summarized the keywords statistics in Fig. 3. The solid color indicates the frequency after adjusting PVP constraint.

We can observe several interesting findings in the statistics: (i) similar to human perception [Lin et al., 2021], we can observe LMMs tend to struggle with performing counterfactual thinking in the order of object-, attribute-, and relation-level imagination.

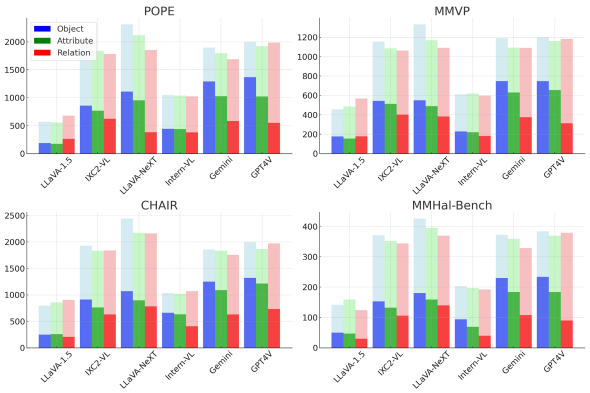


Figure 3: The statistical results for the number of counterfactual keywords for 6 baselines and 4 benchmarks in each three category. Note that the brighter colors in each bar indicates raw keyword count, and the solid colors are the count after adjusting PVP constraint.

This difficulty is clearly shown in the filtered ratios using PVP for each keyword category— note that the most filtered category is relation. (ii) following the scaling law, the more outperforming models that exploiting larger LLMs shows a better capability of extracting keywords. Especially for proprietary models, they show less than 40% filtered ratio in object- and attribute-level keyword categories, unlike open-source models, which have a filtered ratio of over 50%. This results in overall lower average CLIP scores for the keywords generated by both Gemini and GPT-4V compared to the open-sourced models, as in Table 8. More detailed statistics are in Fig. 7 and Appendix B.4.

4.3 Experimental Results

Discriminative Benchmarks. The evaluation for discriminative benchmarks is summarized in Table 2. As in the table, we can observe that overall performance has been improved, compared to the baselines after adopting our methods. Especially, as analyzed in [Liu et al., 2023a], the composition of POPE focuses solely on questioning the existence of objects, rather than their absence (*e.g.,* "Is there `{something}` in the image?"). The combinatorial results of a high accuracy and F1 score indicate that our method can boost the existing LMMs to effectively mitigate hallucination by cautiously confirming *yes* for the existence of objects (*i.e.,* the model does not often *make up* objects).

We further compare our method with 6 LMM baselines in MMVP benchmark, which comprehensively assess CLIP-blind pairs for 9 distinct visual patterns. As shown in the Table 2, the results indicate significant improvements in average accuracy after adjusting Counterfactual Inception— increasing from 5.8% up to 18.53%. These improvements show that the counterfactual thinking is indeed helpful to reassess the visual context for the given images without further fine-tuning, leading to reliable responses that capture more relevant facts and complex visual patterns.

Model	#param	CHAIR		MMHal-Bench	
		C_S (↓)	C_I (↓)	All (↑)	Hal (↓)
Open-source Models					
LLaVA-1.5	13B	26.4	11.12	2.39	52.1
+ Ours		**22.4**	**10.94**	**2.54**	**42.7**
IXC2-VL	7B	24.4	9.75	3.17	29.2
+ Ours		**20.2**	**8.30**	**3.38**	**25.0**
LLaVA-NeXT	34B	19.6	10.10	3.30	34.0
+ Ours		**16.6**	**7.81**	**3.42**	**32.0**
InternVL 1.5	26B	18.2	9.00	3.15	33.3
+ Ours		**17.8**	**7.93**	**3.42**	**26.0**
Proprietary Models					
Gemini 1.5 Pro	N/A	23.4	**12.01**	3.62	31.0
+ Ours		**22.4**	12.76	**4.30**	**13.5**
GPT-4V	N/A	20.0	9.23	3.44	28.1
+ Ours		**17.8**	**8.67**	**3.47**	**20.8**

Table 3: The evaluation results on generative benchmarks. C_S and C_I indicates CHAIR metric for sentence- and instance-level, respectively. In MMHal-Bench, "All" indicates overall scores evaluated by GPT-4 and "Hal" denotes the hallucination rate (%) in the model responses.

Generative Benchmarks. Beyond the discriminative benchmarks, which primarily evaluate multiple choice questions, we assess LMM baselines to identify their non-hallucinatory generation capabilities by measuring the proportion of hallucinated contents in their responses. As presented in Table 3, our method enhances the overall performance on both CHAIR and MMHal-Bench benchmarks. For CHAIR evaluation, we randomly sample 500 images from COCO 2014 validation set and prompt ("Please describe this image in detail.") to the models with max generation length of 64. As in the table, for the both per-sentence (C_S) and per-instance (C_I) results demonstrate consistent improvements in the tasks of long and short description generation across LMM baselines in general.

For the results of MMHal-Bench using GPT-aided evaluation, we clearly observe not only performance gains in the overall score but also a remarkably reduced hallucination ratio. In particular, Gemini 1.5 Pro exhibits a significant hallucination reduction in their responses, with improvements of more than 50%. From the generative results

	Models	PVP	POPE (dis)		MMHal-B (gen)	
			Acc (↑)	F1 (↑)	All (↑)	Hal (↓)
Baseline	LLaVA-1.5	-	84.07	82.62	2.39	52.08
	IXC2-VL		84.13	84.37	3.17	29.17
+ $\mathcal{O}$	LLaVA-1.5	✗	83.47	81.37	2.41	46.88
	IXC2-VL		84.57	83.39	2.93	30.00
+ $\mathcal{O}$	LLaVA-1.5	✓	84.43	82.70	2.48	45.00
	IXC2-VL		86.53	85.29	3.21	27.00
+ $\mathcal{O};\mathcal{A};\mathcal{R}$	LLaVA-1.5	✗	83.57	81.64	2.42	46.00
	IXC2-VL		86.13	84.89	2.79	36.46
+ $\mathcal{O};\mathcal{A};\mathcal{R}$	LLaVA-1.5	✓	85.03	83.40	2.54	42.71
	IXC2-VL		87.50	86.42	3.38	25.00

Table 4: The results of ablation study for the effectiveness of PVP constraint and the conjunction of keyword categories. $\mathcal{O}$ indicates the result of only utilizing object-level keywords.

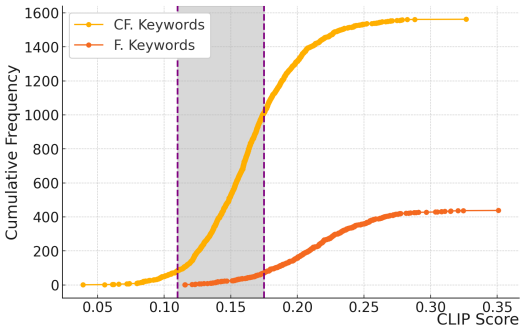


Figure 4: The cumulative frequency distribution along the scores for COCO dataset with 6 baselines. The dashed lines indicates PVP constraint area.

above, by introducing counterfactuals to LMMs, we demonstrate that our method encourages the model to explore alternative paths, thereby enhancing contextual understanding based on true visual clues and reducing hallucinatory responses.

4.4 Analysis on Counterfactual Inception

Ablation Study. We mainly conduct ablation studies on the following two components: (i) the effectiveness of PVP constraint, which is designed to truncate the self-generated keywords that are either too similar or too deviated and (ii) the combinatorial results of using object-, attribute-, and relation-level counterfactual keywords. For the ablation studies, we use two baselines (LLaVA-1.5 and IXC2-VL) along with POPE (discriminative) and mmHal-Bench (generative) benchmarks.

First, as shown in Table 4, the existence of PVP constraint can significantly boost benchmark performances, indicating that the selection of optimal keywords is an important factor for counterfactual thinking. This indicates that disregarding too similar (closer to factual) or too deviated keywords potentially provokes ill-posed response generation and leads to cross-modal inconsistency. Through this ablation, we demonstrate that PVP, which leverages a simple yet effective truncation method based on the alignment score between visual contents and keywords, is a necessary step for integrating counterfactual keywords into LMMs without additional training. Further discussion is in Appendix C.2.

Next, as in Sec. 3.1, we mainly generate counterfactual keywords at three different levels of granularity— object, attribute, or relation. We analyze how the attribute- and relation-level keywords can further enhance performance by using object-level keywords ($\mathcal{O}$) as the primary anchors for conceptualizing counterfactuals. By comparing the results of $+\mathcal{O}$ and $+\mathcal{O};\mathcal{A};\mathcal{R}$ with PVP constraint adjusted, we recognize that the conjunction of keywords indeed helps to broaden context awareness, which results in performance improvements and mitigates hallucinatory responses.

Validity on Counterfactual Keywords. We explore the validity of generated counterfactual keywords and the use of PVP constraint by analyzing their distribution across CLIP scores. First, since no ground truth labels for the self-generated keywords, we randomly sampled 100 images from COCO 2014 validation set and manually determine whether the keywords were closer to counterfactual or factual for the given images (binary task)— total 2K generated keywords integrated from whole 6 baselines. After that, as illustrated in Fig. 4, we visualize the cumulative frequency of each sample based on their CLIP score and analyze distribution with the gray colored PVP constraint area.

The thresholds of PVP constraint are depicted as purple dashed lines for distinguishing optimal counterfactual keywords. In PVP constraint area, we can observe that a large number of yellow scatter points, categorized as counterfactual keywords, are included in the gray zone with a steep slope. In addition, the orange distribution of factual keywords are mostly located above the upper threshold. In summary, we highlight the robustness of our refinement method in identifying optimal counterfactual keywords. Note that extreme cases (either too similar or too deviated) are sparsely distributed at both extremes and filtered out through PVP constraint.

Closer Look at Counterfactual Keywords. As an additional analysis, we explore the counterfactual keywords that frequently occurred in each of 6 baselines for the same 500 images sampled from

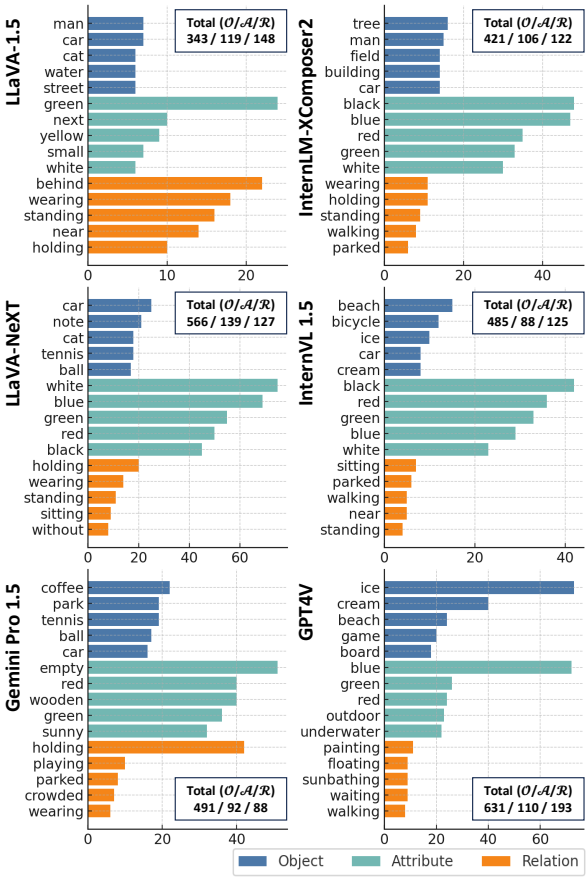


Figure 5: The graphical results of Top-5 words occurrence using morphological analysis (NLTK) in counterfactual keywords. Each legend box indicates total number words in object, attribute, and relation keyword category, respectively.

COCO 2014, which can reveal word-level distribution and potential bias when generating the keywords. To do that, we tokenize the counterfactual keywords for each category: $\mathcal{O}$, $\mathcal{A}$, and $\mathcal{R}$ with PVP constraint. Then, we conduct a morphological analysis for each category using the following criteria: $\mathcal{O}$ for nouns, $\mathcal{A}$ for adjectives, and $\mathcal{R}$ for adverbs and verbs. In Fig. 6, we visualize the top-5 morpheme words for each category. As in the figure, we can observe that $\mathcal{A}$ keywords tend to focus on colors when modifying attributes, while both $\mathcal{O}$ and $\mathcal{R}$ are relatively evenly distributed in general, especially considering the low count of top-1 words and total categorical counts. Interestingly, we find that GPT4V shows a notable bias towards "ice" in its generation of counterfactual keywords ($\mathcal{O}$)— ice cream, iced tea, iced donuts, etc,. Such bias may the frequently occurred words in its training data, reflecting a specific weakness of the model's ability to generate diverse alternatives. Also this indicates the potential availability

What colors are the traffic signs at the top, in the middle, and at the bottom?

Baseline:
The traffic signs at the top and in the middle are blue with white text, while the sign at the bottom is white with black text.

Counterfactual Keywords:
windmill tower, flag of a city, curved traffic light pole, street sign for beach, traffic light with different colors
Counterfactual Inception:
The traffic signs at the top and bottom are white, while the one in the center is blue.

What is the gender of the person riding the motorcycle in the image?

Baseline:
The person riding the motorcycle in the image is a woman.

Counterfactual Keywords:
parked on a gravel path in the snow, vintage car, bike in a shop display, motorcycle on sidewalk
Counterfactual Inception:
The image does not show a person riding the motorcycle, so it is not possible to determine the gender of the rider.

Which sheep is in front of the other sheep?

Baseline:
The sheep in front of the other sheep is the one that is standing in the mud.

Counterfactual Keywords:
cows and pigs, goats, flock in an abandoned factory, giant lambs' ears
Counterfactual Inception:
The sheep in the front is the one with a predominantly white coat and a blue tag in its ear. It is standing in front of the other sheep, which have a mix of black and white wool.

Figure 6: Case study on MMHal-Bench using the highest-performing model (InternVL 1.5). The hallucinatory responses are marked as red, and the refined responses are blue using ours.

of counterfactual keywords as revealing generative vulnerabilities in the alternative responses.

Case Study of Counterfactual Inception. The case studies are depicted in Fig. 6 for the image-question pairs on MMHal-Bench, where it evaluate the degree of hallucination in the generated model responses. As shown in the figure, our method mitigates hallucinatory responses and answers grounded on the true visual clues in the image (not solely based on the biases). We highlight that this is mainly due to the counterfactual keywords— plausible but misleading visual interpretations, which expand visual understanding by using these keywords as the primary anchor, thereby enabling broader contextual exploration based on alternative visual contents. We include additional qualitative results and failure cases in Appendix C.

5 Conclusion

In this work, we propose a novel method of reducing hallucination in LMMs, Counterfactual Inception. By integrating counterfactual thinking to the models through self-generated keywords, our approach improves the reliability of model responses.

The introduction of Plausibility Verification Process (PVP) further ensures the precision of selecting counterfactual keywords to implant counterfactual thinking. Our extensive analyses across various models and benchmarks corroborate that our approach can effectively trigger exceptional thought to the models without additional training and mitigate hallucination in their responses.

6 Limitation and Future Scope

Our study introduces Counterfactual Inception, implanting counterfactual thinking into LMMs and demonstrates that conditioning on counterfactual keywords is helpful to mitigate hallucinatory response generation. Despite our new findings, our work reveals several limitations to discuss and future research direction for further exploration.

Firstly, even if we have examined the recent outperforming baselines with varying model sizes including both open-source and closed-source, due to academic budget and computational power, our work restricted to investigate how the model sizes can affect the capability of implanting counterfactual thinking and the degree of hallucination in their responses. This leaves an open question to figure out the impacts of counterfactual thinking across smaller and larger size of LMMs.

Furthermore, our framework requires additional computational costs due to the self-generation task of counterfactual keywords. As computational analysis, we compare the token throughput (token/s) and latency (ms/token) on 8 NVIDIA RTX A6000 GPUs as in Table 5 (randomly sample 50 examples on COCO dataset with two baselines). Here, even at the cost of slightly increased inference time, we emphasize the importance of our approach, which significantly mitigates the hallucinatory responses for real-world applications, as well as ensuring quick responses.

Lastly, while we introduced a simple yet effective PVP constraint to filter out counterfactual keywords, its optimality can be enhanced with a more rigorous filtering mechanism. As we investigated in Sec. 4.4, selecting optimal counterfactual keywords significantly affects hallucinatory generation. As discussed in Appendix C.2, incorrectly assigned counterfactual keywords can provoke ill-posed response generation, such as parroting keywords— this tendency is exacerbated in smaller models. This suggests a further need to explore more effective methods for identifying optimal counterfactual keywords as a future direction.

Model	# param	Throughput (token/s)↑	Latency (ms/token)↓
LLaVA-1.5	13B	11.08	91.03
+ Ours		7.60	134.08
InternVL 1.5	26B	9.40	106.44
+ Ours		5.93	169.91

Table 5: Throughput (token/s) and latency (ms/token) are computationally analyzed. Note that we calculate the total computation involved in keyword generation and model responses.

Acknowledgments

This work was partially supported by two funds: IITP grant funded by the Korea government (MSIT) (RS-2022-II220984) and Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD230017TD). Additionally, it was supported by the KISTI National Supercomputing Center with supercomputing resources including technical support (KSC-2024-CRE-0160).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Wang, et al. 2024a. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *arXiv preprint arXiv:2403.18058*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024b. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. 2023. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Jinjie Gu, and Huajun Chen. 2024a. Unified hallucination detection for multimodal large language models. *arXiv preprint arXiv:2402.03190*.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

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%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Kai Epstude and Neal J Roese. 2008. The functional theory of counterfactual thinking. *Personality and social psychology review*, 12(2):168–192.

Google. 2023. Gemini.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. 2024. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36.

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

Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. 2024. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *arXiv preprint arXiv:2406.01920*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

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*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*. PMLR.

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.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Yin-ting Lin, Garry Kong, and Daryl Fougnie. 2021. Object-based selection in visual working memory. *Psychonomic Bulletin & Review*, 28:1961–1971.

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

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. In *Advances in Neural Information Processing Systems*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023d. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2023. UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In *International Conference on Learning Representations*.

Peter Menzies and Helen Beebee. 2001. Counterfactual theories of causation.

OpenAI. 2023a. ChatGPT. https://openai.com/blog/chatgpt/.

OpenAI. 2023b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2023c. GPT-4V(ision) System Card.

OpenAI. 2024. Hello gpt-4o.

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

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Neal J Roese. 1997. Counterfactual thinking. *Psychological bulletin*, 121(1):133.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Teknium. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Amos Tversky, Daniel Kahneman, and Paul Slovic. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. 2023. Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. 2024. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms. *arXiv preprint arXiv:2405.17821*.

xAI. 2024. Grok-1.5 vision preview.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer.

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. *arXiv preprint arXiv:2310.16045*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023. Rlhf-v: Towards

trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *International Conference on Learning Representations*.

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 Benchmark and Metric

We additionally explain the benchmarks details for better understanding of their data statistics and metrics to evaludate hallcination.

A.1 Discriminative Benchmark

POPE [Li et al.] (Polling-based Object Probing Evaluation) is designed to detect object hallucinations using 9K image-question pairs. The questions are about the presence of objects (*e.g.,* `"Is there a person in the image?"`) and are categorized into three sampling settings based on the selection method of nonexistent objects: *random*, *popular*, and *adversarial*. In the random setting, nonexistent objects are chosen randomly. In the popular setting, objects are selected from a pool of those most frequently occurring, whereas in the adversarial setting, objects that often co-occur but are absent in the image are chosen. In our experiment, we focus exclusively on *adversarial* setting, as it is the most challenging setting than the others and better represents the complex hallucination aspects of real-world adaptation. The evaluation metrics used are accuracy, precision, recall, and F1-score.

MMVP [Tong et al., 2024] (Multi-modal Visual Patterns) aims to identify *CLIP-blind pairs* that are considered similar by CLIP but have distinct visual semantics. It contains 150 pairs with 300 questions across 9 visual patterns: Orientation and Direction (🧭), Presence of Specific Features (🔍), State and Condition (🔄), Quantity and Count (🔢), Positional and Relational Context (📍), Color and Appearance (🎨), Structural and Physical Characteristics (⚙️), Text (🅰️), Viewpoint and Perspective (📷). The questions are carefully designed to ask the details that CLIP vision encoder ignores and provides two options to select (*e.g.,* `"Where is the yellow animal's head lying in this image? (a)Floor (b)Carpet"`). Accuracy is used as the evaluation metric for each of the 9 visual patterns, and only when the models correctly predict both pairs is the accuracy considered.

A.2 Generative Benchmark

CHAIR [Rohrbach et al., 2018] (Caption Hallucination Assessment with Image Relevance) is a benchmark for evaluating image and caption consistency from the language generation. It calculates the degree of word cardinality intersection between the responses generated by the model and the actual image captions. It uses two variations of the metric, per-sentence (C_S) and per-instance (C_I), to evaluate whether the responses include hallucinated objects:

$$C_S = \frac{|\{\text{sentences w/ hallucinatory object}\}|}{|\{\text{all sentences}\}|}, \quad C_I = \frac{|\{\text{hallucinatory objects}\}|}{|\{\text{all objects mentioned}\}|}. \tag{2}$$

For CHAIR evaluation, we randomly sampled 500 images from COCO 2014 validation and generate model responses with the max length of 64.

MMHal-Bench [Sun et al., 2024] focuses on the evaluation of the degree of hallucination, which is different from the previous LMM benchmarks [Liu et al., 2023d], with GPT-4. The question, response, category names of the image content, and human-generated answer are provided as input to GPT-4. Then, GPT-4 measures the severity of hallucination in a range of 0 to 7. The higher score denotes less hallucination. The questions can be sorted into 8 types: object attribute, adversarial object, comparison, counting, spatial relation, environment, holistic description, and others.

B Details of Counterfactual Inception

B.1 Algorithm

The better understand of full method, we specified the detailed algorithm of Counterfactul Inception in algorithm 1.

Algorithm 1 Counterfactual Inception

Require: Input image v, user query q, LMM M_θ, keyword generation prompt p in Table. 6

Initialize keyword lists $\mathcal{O}, \mathcal{A}, \mathcal{R}$
for c $\in \{\mathcal{O}, \mathcal{A}, \mathcal{R}\}$ **do** ▷ Keyword gen & PVP
  $k \leftarrow M_\theta$.generate(v, p_c)
  $k_{pvp} \leftarrow \{k \in |\mathcal{K}| : \lambda_{bot} \leq$ **CLIP**$(v, k) \leq \lambda_{top}\}$
  Append k_{pvp} to category list.
end for
$k^* \leftarrow [\mathcal{O}; \mathcal{A}; \mathcal{R}]$ ▷ Concatenate all keywords
while $t < T$ **do** ▷ Implanting keywords
  $\text{logit}_{M_\theta} \leftarrow M_\theta(v, q, k, y_{<t})$
  $y_t = \text{argmax}(\text{Softmax}(\text{logit}_{M_\theta}))$
  Set $t \leftarrow t + 1$
end while
return $y_{<t+1}$ ▷ Return generated responses

B.2 Keyword Generation

We have utilized counterfactual keywords to implant counterfactual thinking into LMMs. Due to space limits in the main manuscript, the detailed methodology for generating these keywords

Counterfactual Keywords Generation Prompt:
###Instruction###
Generate a list of counterfactual keywords for the provided image. These keywords should propose plausible yet intentionally misleading alternatives to the actual visual content of the image. Ensure that the changes are visually conceivable and logically consistent within the context of the scene.

###Guidelines###
(option. $\mathcal{O}$**)** Object Substitution: Replace an object in the image with another that could logically occupy the same space but alters the scene's context or meaning.
(option. $\mathcal{A}$**)** Attribute Modification: Change an object's color, size, or shape in a way that makes sense visually but leads to a different interpretation.
(option. $\mathcal{R}$**)** Relational Changes: Adjust the spatial or interactional relationships between objects to suggest a different narrative or dynamic within the scene.

##Examples##
(Image 1): The photo features a tuxedo cat sitting inside the drum of a front-loading washing machine. The cat's distinctive white and black fur, white bib, and paws are visible against its dark body. It stares directly at the camera with bright eyes. The washing machine has various control knobs and buttons, and the area is cluttered with items like laundry detergent. The ambient, warm lighting adds a homely feel.
(option. $\mathcal{O}$**)**: small dog, laundry basket, robot vacuum, soccer ball
(option. $\mathcal{A}$**)**: orange cat, glowing dryer, vintage suitcase, oversized watch
(option. $\mathcal{R}$**)**: cat outside the dryer, dryer in a store display, cat playing with socks

###Your Answer###
List as many counterfactual keywords as possible for the image following the guidelines.
[Counterfactual Keywords]:

Table 6: Instruction prompt for generating counterfactual keywords. To generate different category of counterfactual keywords: object-, attribute-, or relation-level, the instruction has three options to choose $\mathcal{O}$, $\mathcal{A}$, or $\mathcal{R}$.

Counterfactual Prompt:
Carefully avoid the listed Counterfactual Keywords in your response.

Counterfactual Keywords: {`cf_keywords`}.
Question: {`question`}

Table 7: Counterfactual prompt to integrate the generated counterfactual keywords with user queries. Note that red text indicates placeholders for the keywords and user questions.

is elaborated in this section. In Sec. 3.1 of the main manuscript, we categorized counterfactual keywords in three different taxonomy: object substitution $\mathcal{O}$, attribute modification $\mathcal{A}$, and relational changes $\mathcal{R}$. In generating the counterfactual keywords directly from the LMMs, we discovered that a simple instruction such as "Generate counterfactual keywords that mismatch for the given image" cannot fulfill our initial counterfactual intention. This is because the counterfactual thinking requires models to possess complex reasoning capabilities that capture exceptional clues in both visual and linguistic contexts.

Referring to comprehensive prompt engineering [Bsharat et al., 2023], we found that adopting in-context learning is an effective way of generating plausible yet misleading counterfactual keywords for visual content. We hypothesize that this is achievable due to the diverse pre-training on the language models inside LMMs, which includes a wide array of hypothetical and counterfactual scenarios found in various texts such as literature and speculative fiction.

Accordingly, we first instruct GPT4V [OpenAI, 2023c] to generate seed examples that are not grounded in the true visual clues, from the perspectives of three different views— object, attribute, and relation. Then, we manually modify the seed examples to meet our counterfactual design. Consequently, as illustrated in Table 6, we introduce a structured prompt to generate counterfactual keywords in three different granularity with selecting options: $\mathcal{O}$, $\mathcal{A}$, and $\mathcal{R}$.

B.3 Counterfactual Prompt

After obtaining counterfactual keywords, we apply a simple rule-based text pre-processing to filter out non-informative characters such as punctuation marks, stop words, noise words. Subsequently, we designed a specific prompt to integrate the counterfactual keywords with user queries with placeholders, which is then forwarded to the models. As shown in Table 7, we sophisticatedly designed a counterfactual prompt to guide the models in disregarding the extracted counterfactual key-

Model	PVP	POPE				MMVP				COCO				MMHal-Bench			
		$\mathcal{O}$	$\mathcal{A}$	$\mathcal{R}$	Score	$\mathcal{O}$	$\mathcal{A}$	$\mathcal{R}$	Score	$\mathcal{O}$	$\mathcal{A}$	$\mathcal{R}$	Score	$\mathcal{O}$	$\mathcal{A}$	$\mathcal{R}$	Score
LLaVA-1.5	✗	571	557	678	0.205	457	487	568	0.199	796	858	907	0.204	142	159	124	0.201
	✓	190	175	262	0.154	177	155	179	0.154	249	258	205	0.153	50	47	30	0.153
IXC2-VL	✗	1963	1836	1783	0.189	1156	1087	1062	0.189	1928	1830	1838	0.188	371	352	344	0.191
	✓	858	768	623	0.152	543	513	403	0.154	913	764	629	0.152	153	132	106	0.152
LLaVA-NeXT	✗	2312	2120	1856	0.191	1333	1170	1092	0.192	2441	2172	2159	0.190	454	400	356	0.197
	✓	1109	954	383	0.154	550	489	383	0.154	1070	897	781	0.153	180	159	140	0.154
InternVL 1.5	✗	1050	1039	1024	0.194	611	619	598	0.189	1034	1020	1071	0.191	203	197	192	0.193
	✓	445	439	380	0.154	230	221	182	0.152	662	634	407	0.151	94	69	40	0.153
Gemini 1.5	✗	1897	1795	1687	0.173	1191	1093	1090	0.178	1859	1832	1753	0.172	372	359	329	0.172
	✓	1291	1028	582	0.151	749	630	377	0.151	1250	1090	632	0.150	230	184	108	0.150
GPT4V	✗	2000	1922	1988	0.178	1200	1160	1182	0.181	1995	1865	1972	0.169	384	370	379	0.181
	✓	1369	1021	549	0.153	748	656	314	0.151	1320	1211	732	0.150	234	184	90	0.150

Table 8: Details of counterfactual keywords statistics and average CLIP score along keyword category.

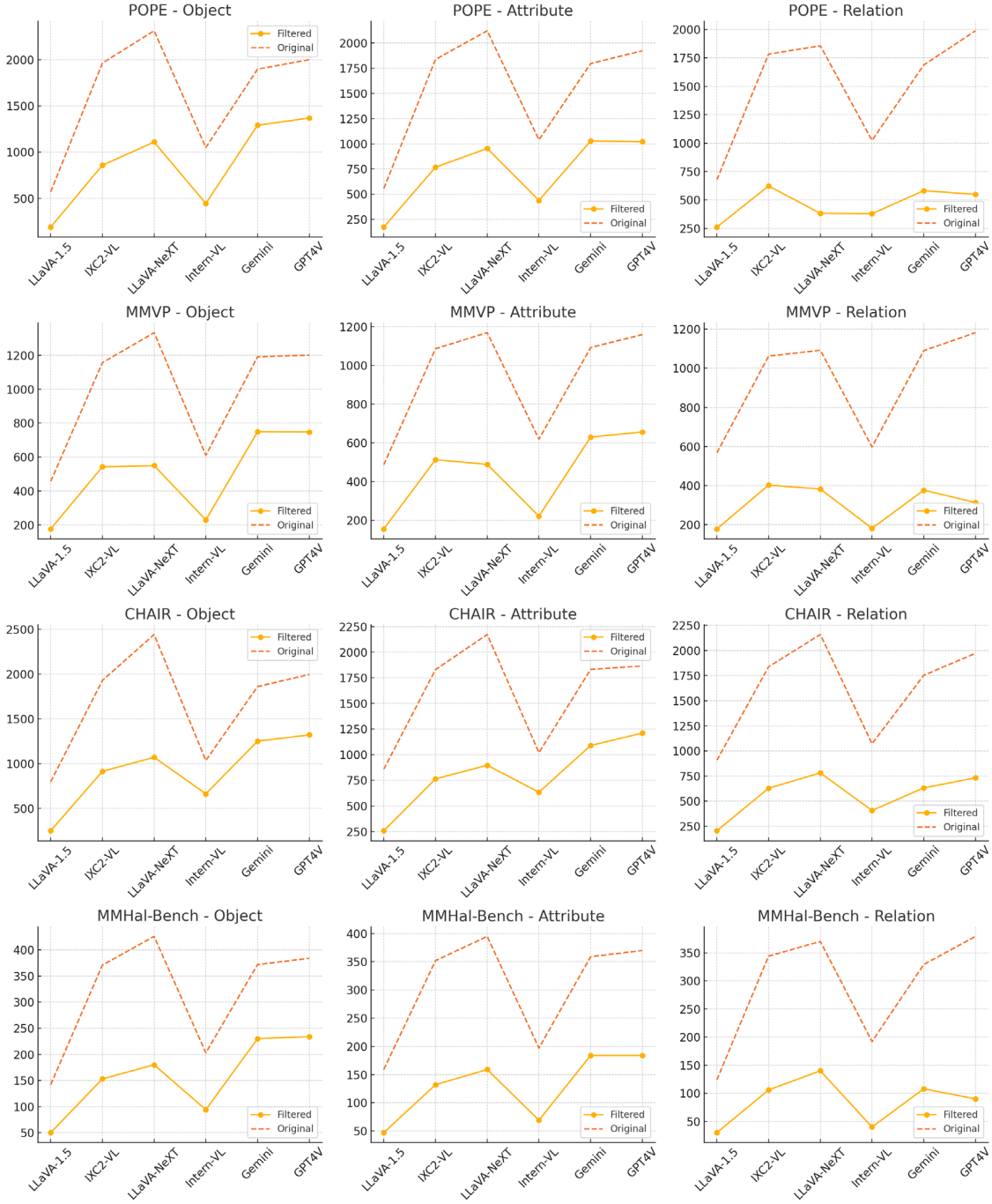


Figure 7: Detailed analysis on the categorical counterfactual keyword distribution.

Q1. Is there a cup in the image?
Q2. Is there a bowl in the image?

InternVL 1.5:
A1. Yes, there is a cup in the image. It is located near the sandwich on the plate.
A2. Yes, there is a bowl in the image. It is located near the sandwich on the plate.

Counterfactual Keywords:
green salad bowl
Counterfactual Inception:
A1. No, there is no cup in the image.
A2. No, there is no bowl in the image.

Q1. Is there a car in the image?
Q2. Is there a truck in the image?

GPT4V:
A1. No, there is no car in the image. The image shows the back of a bus with an advertisement featuring an animated character. The bus is on a road, and there are some trees and signs visible in the background, but no cars are visible in this particular shot.
A2. Yes, there is a truck in the image. It appears to be a commercial truck with an advertisement on the back featuring an animated character. The truck is on a road with other vehicles and buildings in the background.

Counterfactual Keywords:
desert road, blue sky, bicycle
Counterfactual Inception:
A1. Yes, there is a car in the image.
A2. No, there is no truck in the image

Figure 8: Additional case study on POPE dataset. The hallucinatory responses are marked as red, and the refined responses are blue using ours.

words when generating responses to user queries. We pinpoint that simply implanting the counterfactual prompt with the counterfactual keywords enables the models to mitigate hallucinatory responses without additional training.

B.4 Details of Keyword Statistics

In addition to Sec. 4.2, we further explore the details of self-generated counterfactual keywords statistics for object, attribute, and relation category. One findings, we can observe as in Table 8, is that the more outperforming LMM baselines show lower average CLIP scores, which indicates better association for the alternatives for the visual clues. Among open-sourced models, we found that InternVL 1.5, which achieved competent performances compared to proprietary multi-modal models, generates relatively a limited number of counterfactual keywords for the given counterfactual instruction. Our assumption of this tendency is on the combined results of its fine-tuning stage, which utilizes text-only data sources such as OpenHermes 2.5 [Teknium, 2023], Alpaca-GPT4 [Taori et al., 2023], ShareGPT [Zheng et al., 2024], and COIG-CQIA [Bai et al., 2024a], and its deeper cross-modal alignment layers, which may leads to focus on the actual clues within the visual context.

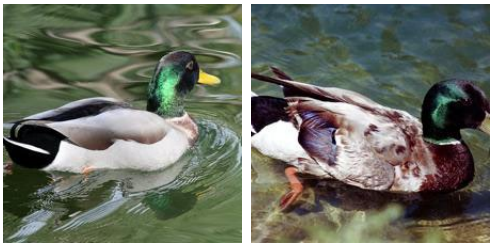
C Qualitative Assessment

C.1 Additional Case Study

In our additional case study, we focus on providing further instances demonstrating the effectiveness of our approach, Counterfactual Inception, across various benchmarks. We evaluated our method on discriminative benchmarks such as POPE [Li et al.] and MMVP [Tong et al., 2024], generative benchmark MMHal-Bench [Sun et al., 2024].

As in Fig. 8 and Fig. 9, we illustrate qualitative results for POPE and MMVP datasets, both are discriminative benchmarks where models select answers from the multiple options provided. The utilized models used in this qualitative study are LLaVA-NeXT, InternVL 1.5, and GPT-4V, all of them are the most outperforming multi-modal models in open-source and close-source, respectively. Importantly, we highlight that after conditioning on the given plausible but misleading counterfactual keywords, the baselines demonstrate a better understanding of the true visual clues, enabling a broader contextual exploration that helps to mitigate hallucinatory responses.

In Fig. 11, we visualize case studies of MMHal-Bench, which is a generative benchmark, to illustrate the effectiveness of Counterfactual Inception



Is the duck's entire beak visible in the picture?
(a) Yes
(b) No

LLaVA-NeXT:
(left) (b) No / (Right) (b) No

Counterfactual Keywords:
orange fish, piranha, dog with yellow collar
Counterfactual Inception:
(left) (a) Yes / (Right) (b) No

Is the lock locked or unlocked?
(a) Locked
(b) Unlocked

InternVL 1.5:
(left) (b) Unlocked / (Right) (b) Unlocked

Counterfactual Keywords:
red paint chipped away, combination dial pad
Counterfactual Inception:
(left) (a) Locked / (Right) (b) Unlocked

Are the wheels of the school bus visible in the image?
(a) Yes
(b) No

GPT4V:
(left) (b) No / (Right) (b) No

Counterfactual Keywords:
neon lights, convertible car, fire truck, delivery van
Counterfactual Inception:
(left) (a) Yes / (Right) (b) No

Figure 9: Additional case study on MMVP dataset. The hallucinatory responses are marked as red, and the refined responses are blue using ours.

Q1. What type of fruit is this?
Q2. How many uncut fruits are in the image?

LLaVA 1.5:
A1. The image shows a group of unripe lychee fruits.
A2. There are four uncut fruits in the image

Counterfactual Keywords:
Kiwi fruit peeled open revealing flesh, fruit bowl
Counterfactual Inception:
A1. The image shows a group of unripe fruit, which are likely to be kiwi fruit, sitting on a wooden table.
A2. There are three uncut fruits in the image.

Figure 10: Failure cases on in-the-wild dataset. The hallucinatory responses are marked as red, and the refined responses are blue using ours.

in mitigating descriptive hallucination and improving generative ability. The results reveal that the original baselines generate ambiguous or inconsistent responses not grounded on the visual contents, as if the model recognizes non-existent objects. These comprehensive case studies demonstrate that our approach not only enables LMMs to clearly understand the visual context but also significantly enhances their reliability in identifying and describing actual elements present in the visual content, thereby providing more reliable and contextually appropriate responses.

C.2 Failure Case

Here, we investigate failure cases to understand the limitations of counterfactual thinking as in Fig. 10. Through the analysis, we identified that small models (LLaVA 1.5-13B) sometimes parrots counterfactual keywords in its generated sentences, rather than effectively constructing counterfactual scenarios using these keywords. We hypothesize that this tendency could be linked to the lack of exceptional thought in small models, which potentially leads to the anchoring effect [Tversky et al., 1982], a cognitive bias where initial information disproportionately influences subsequent responses. Although we have proposed a simple and effective PVP constraint to mitigate such negative potential in advance, developing more advanced constraints could be another future research to enhance the counterfactual thinking capabilities of LMMs.

How is the yellow boat positioned in relation to the white yacht in the image?

LLaVA-NeXT:
In the image, the yellow boat is positioned to the right of the white yacht. The yellow boat appears to be moving towards the right side of the frame, while the white yacht is stationary or moving in the opposite direction.

Counterfactual Keywords:
cloudy sky, man holding phone, red, man, the image shows a man driving, man looking at water.

Counterfactual Inception:
The yellow boat is positioned in front of the white yacht in the image.

What is playing on the monitor?

InternVL 1.5:
The monitor is not displaying any content, so I cannot determine what is playing on it.

Counterfactual Keywords:
computer keyboard accessory, white laptop

Counterfactual Inception:
The image does not provide any information about what is playing on the monitor, as the monitor is not visible in the photo.

How many people are there in the image?

GPT4V:
There are three people visible in the image. They appear to be posing together for a photo.

Counterfactual Keywords:
blue background, congratulations, summer 2021, summer 2014, beach photo, winter 2014, family at the beach, birthday card, happy new year.

Counterfactual Inception:
There are four people in the image.

Figure 11: Additional case study for MMHal-Bench dataset. The hallucinatory responses are marked as red, and the refined responses are blue using ours.