CODE: Contrasting Self-generated Description to Combat Hallucination in Large Multi-modal Models

Junho Kim* Hyun Jun Kim* Yeon Ju Kim Yong Man Ro†

Integrated Vision and Language Lab, KAIST, South Korea
{arkimjh, kimhj709, yeonju7.kim, ymro}@kaist.ac.kr

Abstract

Large Multi-modal Models (LMMs) have recently demonstrated remarkable abilities in visual context understanding and coherent response generation. However, alongside these advancements, the issue of hallucinations has emerged as a significant challenge, producing erroneous responses that are unrelated to the visual contents. In this paper, we introduce a novel contrastive-based decoding method, COuntering DEscription Contrastive Decoding (CODE), which leverages self-generated descriptions as contrasting references during the decoding phase of LMMs to address hallucination issues. CODE utilizes the comprehensive descriptions from model itself as visual counterpart to correct and improve response alignment with actual visual content. By dynamically adjusting the information flow and distribution of next-token predictions in the LMM's vocabulary, CODE enhances the coherence and informativeness of generated responses. Extensive experiments demonstrate that our method significantly reduces hallucinations and improves cross-modal consistency across various benchmarks and cutting-edge LMMs. Our method provides a simple yet effective decoding strategy that can be integrated to existing LMM frameworks without additional training.

1 Introduction

With recent advancements of Large Language Models (LLMs) [18, 4, 54, 11, 46], Large Multi-modal Models (LMMs), sometimes referred as Large Vision-Language Models [24, 63, 32, 31], have been drawn great attention for their natural multi-modal interaction with users through back-and-forth conversations. Leveraging their robust generation capabilities, various pioneering tasks in pre-LMM era such as image captioning [7, 58, 27], visual question answering [1, 2], object detection [17], etc., have been integrated into a single task rather than treated as sub-tasks and achieved significant milestones [21, 59, 44]. However, at the same time, the hallucination issue [49, 67] has become one of the emerging problems when adopting LMMs into real-world applications due to their potential spurious generation in critical areas.

Here, unlike hallucination studies in LLMs [22] mainly focusing on factuality hallucination originated from the language knowledge, the hallucination problem in LMMs refers cross-modal inconsistency between the given visual contents and the generated responses for the user instructions. After the seminal works [13, 42, 68] giving eyes to LLMs to understand visual contents with visual instruction tuning, numerous cutting-edge LMMs [61, 9, 41, 36] actively have been proposed. Albeit the scaling laws following more stronger versatile vision models [37, 29, 33], higher resolution [41, 8], deeper alignment layers [5, 45, 8], larger model sizes, etc., LMMs still suffer from generating responses that seem plausible but are factually incorrect for the given visual contents.

*Equal contribution. † Corresponding author.
Code is available at https://ivy-lvlm.github.io/CODE

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-4246

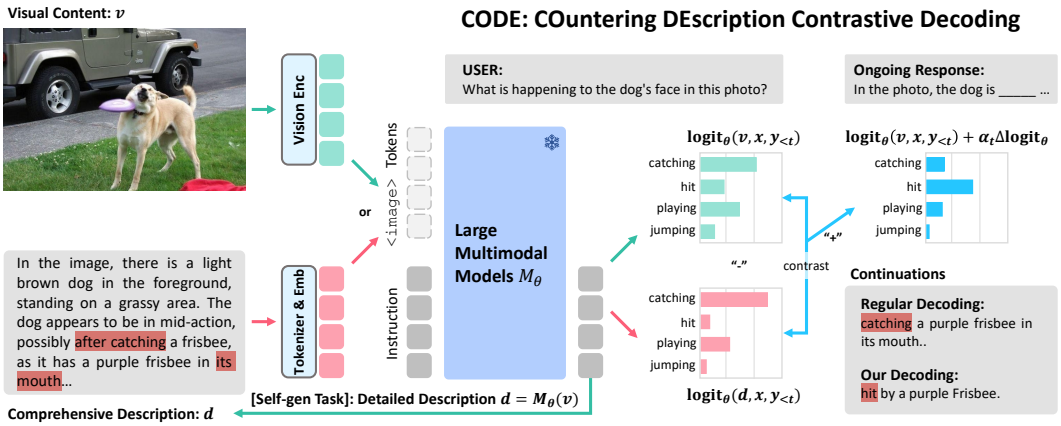


Figure 1: The overall decoding procedure of CODE. After LMMs generate detailed description for the visual content by themselves (see instruction in Appendix. A), the model recursively outputs logits from each v and d. By contrasting between two log-likelihoods, CODE produces more contextual and correct responses that match the given visual content suppressing inconsistent words (*catching*→*hit*).

The origin of LMM hallucination is an intertwined problem for their inherent training paradigm, which involves alignment projection matching during the pre-training, followed by fine-tuning with the limited instruction-following data. Several approaches, aimed for mitigating hallucinatory effects, have addressed the inconsistent issues in the context of data-associated solution [39, 56], scaling model architectures [65], or additional RL-based training [64, 52]. Among them, reactive methods [23, 14, 26] intervene the decoding phase of LMMs' inference and alleviate undesired responses. Motivated by Li *et al.* [34] that have proposed contrastive decoding (CD) method between expert and amateur language models, recent CD-based approaches in LMMs have proposed several ways of contrasting model responses from visual inputs with their counterparts (*e.g.,* visual contamination [30], image-biased models [69], or fine-grained visual information [10]).

Our research question begins with *"How effectively do contemporary LMMs capture visual evidences in their descriptive responses, and what information must be curbed to produce informative and consistent responses?"*. As illustrated in Fig 1 (bottom-left), when asking LMMs to generate a comprehensive description for visual content, the output seemingly generates detailed description effectively, but a closer examination often reveals missed fined-grained information or hallucinatory instances in the responses. By recursively referring these incomplete descriptions generated by the models themselves, we aim to restrict the incorrect information flow during the generation phase and enhance the alignment of the model responses grounded in true visual evidences.

In this paper, we introduce a novel training-free contrastive decoding method, COuntering DEscription Contrastvie Decoding (CODE), designed to use self-generated descriptions as a contrastive reference to mitigate hallucination issues in LMMs. The core idea of our proposed method is on harnessing the self-generated descriptions which possibly encompass both factual evidence and hallucinatory information from visual contents as a look-up reference for response correction. Specifically, within our contrastive framework as illustrated in Fig. 1, the comprehensive descriptions from model itself alternatively propagate to visual input tokens and contrast the discrepancy with the logits from actual visual contents to enhance next-token prediction. In addition, we introduce a dynamic restriction strategy that enables adaptive control of information flow during the auto-regressive decoding phase, taking into account both token-level predictions and their distribution within the vocabulary set.

By conducting extensive experiments and analyses on prevailing cutting-edge LMMs [40, 51, 16, 41, 62, 8], we corroborate the effectiveness of our method in reducing hallucination and enhancing the coherence and informativeness in various benchmarks [35, 42, 52, 53, 57]. Our decoding method can be seamlessly integrated into existing LMMs by simply substituting the image tokens with self-generated descriptions in a training-free manner.

Our contribution can be summarized into three-fold as follows:

- We introduce COuntering DEscription Contrastive Decoding (CODE), a training-free decoding strategy that employs self-generated descriptions to minimize hallucinations in LMMs.

https://doi.org/10.52202/079017-4246

By contrasting logit information from descriptions with actual visual contents, CODE enhances visual consistency and coherence in the model responses.

- Our approach incorporates dynamic restriction strategies within the contrastive decoding phase. It selectively regulates the information flow by adjusting token-level predictions based on their distribution in the vocabulary, thus ensuring more contextual responses.
- We validate the effectiveness of our decoding method across various benchmarks using cutting-edge LMMs. The results demonstrate that CODE significantly reduces hallucination while enhancing the relevance and informativeness in the responses.

2 Related Work

Vision+LLM: Large Multi-modal Models After the emergence of large-scaled LLMs [4, 54, 11] that can interact with users with question-answer chat, various of vision+LLM studies— *i.e.,* LMMs, have been proposed to integrate the robust linguistic capability into the visual understanding and reasoning in diverse vision-language task. As earlier works such as LLaVA [42], Instruct-BLIP [13], and MiniGPT-4 [68], which utilized visual instruction tuning, have bridged two modalities of vision and language through fine-tuning with a learnable query, exemplified by Q-Former [31] or projection layer-based alignments [40]. To enhance cross-modal consistency in vision-language representation, recent works have proposed several solutions to address the underlying weaknesses in both modalities: (i) utilizing higher-resolution visual inputs [41, 9], (ii) deploying Mixture-of-Expert (MoE) concepts integrating versatile vision models [37, 29, 33], (iii) improving weak alignment interface [5, 45, 8], or (iv) adopting larger LLMs to scale up the language model prior [62].

Hallucination Issue, Harming Cross-modal Consistency Despite of the endeavor developments of LMMs, they cannot be free from cross-modal inconsistency between the visual contents and their generated responses, so-called hallucination [49]. This not only leads to performance degradation but also provokes an over-reliance issue, resulting in incorrect model responses that are not grounded in true visual evidence. This critical concern regarding response trustworthiness and model reliability hinders the adoption of LMMs in real-world applications. To mitigate the hallucination problem, diverse works have been proposed employing additional training on curated datasets [52, 55] or reinforcement learning under feedback systems [64, 66]. Among them, by intervening during the response generation, decoding-based approaches [34, 12] are introduced to encourage models to represent more precise responses. We refer to readers for more comprehensive survey papers [43, 3] addressing hallucination in LMMs. Our work is in line with CD-based approaches that utilize logit discrepancy from counterpart outputs to enhance coherence. Unlike the previous works [30, 23, 10] that focus on twisting visual information, we utilize self-generated description as contrasting visual counterpart and correct hallucinatory responses based on the model understanding.

3 Proposed Method

Problem Setup and Preliminaries. Let M_θ denote a vision-language LMM parameterized by θ that auto-regressively generates responses for the given visual contents v and input textual query x. Then the model maps the logit distribution to the next token prediction output $y_t \in \mathbb{R}^{|\mathcal{V}|}$ at time step t in the vocabulary set $\mathcal{V}$ such that $y_t \sim p_\theta(y_t|v, x, y_{<t}) \propto \text{logit}_\theta(y_t|v, x, y_{<t})$, where $y_{<t}$ indicates all previously generated tokens. During the response generation, we can deploy several decoding strategies to choose the next word using either deterministic search (*e.g.,* greedy, beam search) or stochastic sampling (*e.g.,* top-k, Nucleus search [20]).

After the seminal works [34, 12] in natural language processing have introduced Contrastive Decoding (CD) mechanism, which considering information disparities between expert and amateur models for more coherence and informativeness, various works have deployed this strategy into LMMs by twisting visual contents [30] or model information [69] for the contrastive approach. The next-token probability p_{cd} from CD can be generally formulated as follows:

$$p_{\text{cd}}(y_t \mid y_{<t}) = \text{Softmax}\left[(1+\alpha)\text{logit}_\theta(y_t \mid v, x, y_{<t}) - \alpha\text{logit}_{\hat{\theta}}(y_t \mid \bar{v}, x, y_{<t})\right], \tag{1}$$

where $\bar{v}$ and $\hat{\theta}$ indicates visual counterparts and sub-optimal amateur model, respectively— note that $\theta{=}\hat{\theta}$ can be regarded as self-correction. Intuitively, the objective of CD is amplifying model outputs

 https://doi.org/10.52202/079017-4246

by reflecting information deviation between top candidate log-probabilities. Therefore, the selection of logit counterparts for referring is the key challenge for high-quality and consistent responses during the contrastive decoding frameworks.

3.1 Comprehensive Image Description as Visual Counterpart

As the visual counterpart, we deploy comprehensive image descriptions generated from the model as contrasting reference, which are inevitably less informative than the visual contents themselves. Our motivation is on the innate difference of information density between vision and language [19]. While vision information exhibits relatively less redundancy for spatial signals— *e.g.,* human can visually recognize objects with a few masked patches on images, languages contain high-entropy information, resulting in spans that are more semantic and information-dense than visual signals. That is, it is difficult to infer blanked-out words— *e.g.,* "I went to the store to buy some ____". Building upon the property of each modality, we first delve into the self-generated model responses with specific instruction for the visual contents to elicit the encompassed visual evidences in the representation space and assess its potential subject role as a visual counterpart for contrastive decoding.

As illustrated in Fig. 1, we input a query instruction into the model to generate a comprehensive visual description for the given visual content (please see the detailed instruction x_0 for the self-generated description in Appendix. A). Then, we exploit the generated description as recursive visual inputs, replacing the position of image tokens in the model input sequence (*e.g.,* `<image>` token in LLaVA series [42, 40, 41]). Ideally, if the model sufficiently covers whole visual evidences to answer any vision-related questions, the generated response should provide competent answers with solely utilizing the description-only embeddings as an alternative of visual embeddings. However, as shown in Fig 2, it is obvious that the results from description-only show sub-optimal performance to answer questions due to insufficient information in capturing visual evidences. Consequently, we use the comprehensive description for the visual contents, which is generated by model itself— but partially incorrect or hallucinatory to capture visual evidences, as a contrasting visual counterpart to enhance response coherence during the decoding phase, which also in line with the amateur model selection philosophy of contrastive decoding [34].

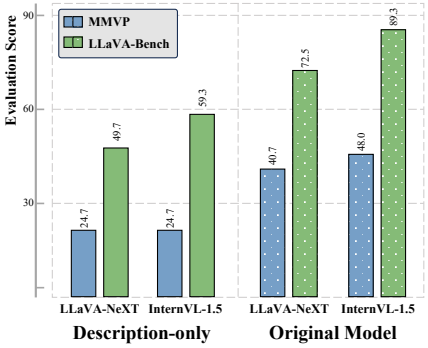


Figure 2: The comparison is based on two benchmarks (MMVP [53]: multiple choice / LLaVA-Bench [42]: description-level). The plain and dotted bars indicate the results for the models that use self-generated descriptions as visual input replacements and original model with actual visual contents, respectively.

3.2 COuntering DEscription Contrastive Decoding

Based on our analysis in sec. 3.1, we can obtain a pair of the visual content and its comprehensive description (v, d), such that d corresponds to $M_\theta(y|v, x_0)$. By contrasting the logit variation between the paired information into the model response generation, we can formulate the next-word prediction using our proposed method, COuntering DEscription Contrastive Decoding (CODE):

$$p_{\text{code}}(y_t \mid y_{<t}) = \text{Softmax}\left[(1+\alpha_t)\text{logit}_\theta(y_t \mid v, x, y_{<t}) - \alpha_t \text{logit}_\theta(y_t \mid d, x, y_{<t})\right]. \quad (2)$$

Here, unlike the previous approaches [34, 30, 12] that restrict the logit variations with fixed α value as in Eqn. 1, we present a dynamic restriction α_t for the logit variations by comparing the information between visual contents and their comprehensive descriptions. Revisiting the role of α, it determines whether to promote or curb information from logit variation, thus directly influencing next-token generation— higher value results in more aggressive adjustment for the variations. However, when confronting that both v and d yield similar logit score on the correct token, the variation gets closer to zero, thereby the next-token prediction can be unexpectedly reversed if other tokens get rewarded than the correct token with a fixed α on a token-by-token basis. Although this aligns with the initial intent of CD, a more robust selector is necessary to effectively restrict the logit information flow.

Accordingly, our method predicts next-token not only at the individual token-level but also considering its distribution across the entire vocabulary set, enabling dynamic control of the information flow.

To measure the relative entropy between the token distributions from visual contents P_t^v and its comprehensive description P_t^d at time step t, we deploy Bounded Divergence ($\mathcal{D}_{\text{bd}}$) [6], which is a type of statistical distance that ensures symmetric and bounded measure:

$$\mathcal{D}_{\text{bd}}(P\|Q) = \frac{1}{2}\sum_{i=1}^{n}(p_i + q_i)\log_2(|p_i - q_i|^k + 1), \tag{3}$$

where $\mathcal{D}_{\text{bd}}(P\|Q) \geq 0$ and equals 0, if and only if $p{=}q$, and k denotes a smoothing parameter. Here, the upper-bound of the divergence apparently exists, such that $\mathcal{D}_{\text{bd}}(P\|Q) \leq \sum_{i=1}^{n} p_i \log_2 2{=}1$, due to the following condition $|p_i{-}q_i| \leq 1$.

We define the dynamic restriction α_t as $1{-}\mathcal{D}_{\text{bd}}(P_t^v\|P_t^d)$, where it enables a token-wise feedback control that adjusts the information weighting with respect to the closeness of the two distributions. The major role of the restriction term is maintaining a balance in the logit variation for the observed prediction disparities between the v and d distributions. That is, when the distributions are close enough (*i.e.,* $P_t^v \approx P_t^d$), the value of $\mathcal{D}_{\text{bd}}(P_t^v\|P_t^d)$ approaches zero, indicating minimal divergence. That is, α_t approaches 1, allowing for higher amplification of logit variations in predicting the next-token outputs. This adjustment reflects the increased reliability of predictions when the two distributions from v and d are closely aligned. On the other hand, for the dissimilar distributions, α_t decreases towards zero, compelling the model to restrict information flow from the variation. This reduction limits the potential for introducing erroneous or less probable predictions by focusing more on visual information, thereby maintaining coherence in the output when the model's understanding of the visual content significantly deviates from its textual description.

3.3 Adaptive Information Constraint

One major challenge in contrastive-based decoding is the scenario where implausible tokens are rewarded, even when predictions are made with low confidence. This issue can also arise in our method, particularly when token distributions derived from textual descriptions provide more confidence than the visual content, ironically undermining the most predictive tokens. To address it, Li *et al.* [34] have introduced an adaptive plausibility constraint, which filters out less plausible tokens by truncating them based on the maximum token confidence from the expert model. While this approach simply penalizes false positive tokens in the candidate pool, it may also have unintended side effects by prematurely applying a cutoff threshold to lower-confidence tokens. Specifically, early threshold settings can sometimes eliminate the possibility of identifying correct token predictions, which might otherwise be dismissed in a pool considered to contain mostly false negatives.

Improving the previous constraint [34, 30], we present adaptive information constraint ($\mathcal{V}_{\text{head}}$) designed to dynamically retain tokens that may be informative despite their lower confidence. By comparing prediction distributions between P_t^v and P_t^d, we filter out less relevant tokens from the candidate pool as follows:

$$\mathcal{V}_{\text{head}}(y_{<t}) = \{y_t \in \mathcal{V} : p_\theta(y_t \mid v, x, y_{<t}) \geq \beta_t \max_w p_\theta(w \mid v, x, y_{<t})\}, \tag{4}$$

where β_t dynamically regulate the token candidate pool utilizing the divergence term in Eqn. 3, defined as $\beta_t = \mathcal{D}_{\text{bd}}(P_t^v\|P_t^d)$. This strategy can expand the token searching pool when the next-token prediction, derived from both visual content and comprehensive description, shows a similar distribution yet uncertainty in selecting the candidate token (*i.e.,* false negatives). Finally, we only consider the next-token prediction within $\mathcal{V}_{\text{head}}(y_{<t})$, and for the tokens satisfying $y_t \notin \mathcal{V}_{\text{head}}(y_{<t})$, we set their logits to $-\infty$ to filter out from the candidate pool. Please see comprehensive Algorithm. 1 in Appendix. B.

4 Experiments

Experimental Setup To validate the efficacy of our method over various LMM families and sizes, we implemented our method on contemporary LMMs: LLaVA-1.5 (13B) [40], Emu2-Chat (14B) [51], InternLM-XComposer2 (7B) [16], LLaVA-NeXT (34B) [41], Yi-VL (34B) [62], and InternVL 1.5 (26B) [8]. We compared our method with five baseline decoding strategies. For the regular decoding strategies, we used greedy decoding, Nucleus sampling [20], and beam search decoding. Additionally, we selected OPERA [23] and VCD [30] for contrastive decoding method, which designed to mitigate

 https://doi.org/10.52202/079017-4246

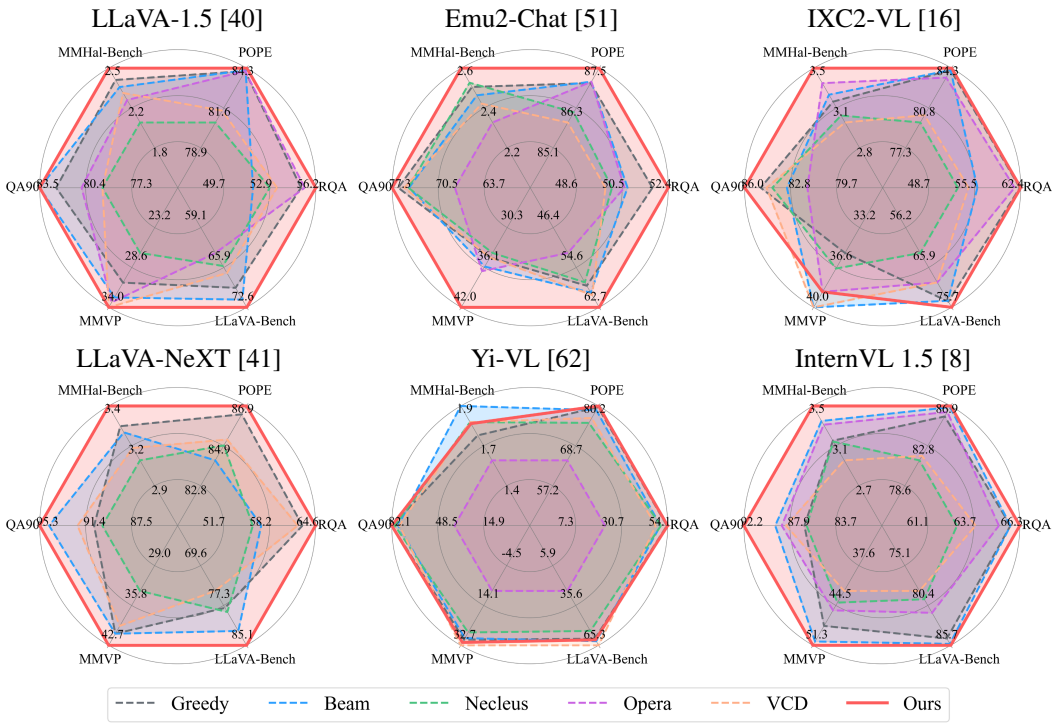


Figure 3: Overview of experimental results on 6 baseline LMMs, 6 decoding method, and 6 hallucination benchmarks in spider chart format.

hallucinations with contrastive frameworks. We used the default parameter settings for all methods, where top-p value 0.95 and temperature 1.0 for Nucleus sampling, the number of window size for searching is 5 (*i.e.,* num-beams 5) for both beam search decoding and OPERA, and CD-$\alpha = 1$, CD-$\beta = 0.1$ for VCD, and $k = 0.3$ for our method. Note that OPERA inference requires too much memory especially for LLaVA-NeXT (34B), so that we excluded OPERA results for this model.

Benchmarks and Evaluation Metrics The benchmarks for evaluating hallucinations in LMMs can be broadly categorized into discriminative and generative streams. The discriminative type assesses hallucinations by evaluating the predicted answer among given options (*e.g.,* multiple choice or yes/no question), while generative benchmarks typically employ more advanced language models (*e.g.,* GPT-aided evaluation) to rate the subject model descriptions. Within this taxonomy, we carefully select 6 benchmarks to test baselines. Please see Appendix. C for benchmark details.

As discriminative benchmarks, we utilize mainly three datasets for detailed evaluation. Specifically, **POPE** [35] is a commonly used benchmark for detecting object hallucination by converting object annotations sourced from MSCOCO [7]. Under the three different subsets: random, popular, and adversarial, the metric for POPE measures binary classification performance for simple yes/no questions. **MMVP** [53] aims to evaluate the understanding of visual details for 9 different visual patterns using paired classification accuracy. Due to its evaluation design, which involves comparing two similar CLIP-blind image pairs, MMVP requires LMMs to capture subtle visual differences. **RealworldQA** [57] is the most recent dataset tailored to assess the capability of LMMs in basic real-world spatial understanding, using the accuracy metric within multiple-choice questions.

We use three benchmarks for generative benchmarks, extending the evaluation scope to include open-ended captioning tasks beyond merely assessing classification within given answer options. Generally, ChatGPT [46] is used to score the quality of the model-generated sentences. The metric for both **LLaVA-QA90** [42] and **LLaVA-Bench (In-the-Wild)** [42] is score ratio, where model responses rated from GPT-4 [47] are divided by GPT-4 answers such that $\sum |\text{model-score}| / \sum |\text{GT-score}|$, where all scores are rated by GPT-4. It has three types of questions: conversation, detailed description, and complex reasoning. **MMHal-Bench** [52] evaluates the degree of hallucination for the 8 various question types: object attribute, adversarial object, comparison, counting, spatial relation,

Table 1: The hallucination evaluation results for the discriminative benchmarks [35, 53]. Each emoji in MMVP colum cell indicates 9 different visual patterns (details in Appendix. C).

Model	#Param	Method	POPE			MMVP									
			Acc↑	Prec	F1↑	🧭	🔍	🔄	🎚️	📍	🎨	⚙️	A	📷	Avg↑
LLaVA-1.5	13B	Greedy	84.07	90.88	82.62	30.77	27.27	0.00	12.50	10.00	53.33	16.67	50.00	40.00	30.67
		Beam	84.13	90.89	82.70	19.23	27.27	11.11	25.00	10.00	60.00	16.67	70.00	35.00	32.67
		Nucleus	80.60	84.46	79.45	26.92	27.27	22.22	12.50	20.00	33.33	0.00	60.00	20.00	26.67
		Opera	84.07	91.01	82.59	42.31	36.36	11.11	25.00	10.00	56.67	16.67	70.00	35.00	33.33
		VCD	81.40	85.00	80.39	34.62	18.18	22.23	37.50	50.00	43.33	33.33	40.00	35.00	**34.00**
		Ours	**84.27**	90.99	**82.86**	19.23	31.82	11.11	25.00	20.00	53.33	16.67	80.00	40.00	**34.00**
Emu2-Chat	14B	Greedy	87.03	94.48	85.85	30.77	36.36	11.11	12.50	20.00	63.33	16.67	70.00	20.00	34.67
		Beam	87.07	93.99	85.96	38.46	27.27	11.11	25.00	20.00	66.67	16.67	60.00	25.00	36.00
		Nucleus	86.17	92.22	85.10	38.46	31.82	16.67	0.00	40.00	43.33	33.33	60.00	30.00	34.00
		Opera	87.07	94.06	85.95	38.46	31.82	11.11	25.00	20.00	63.33	16.67	70.00	25.00	36.67
		VCD	85.87	91.58	84.82	34.62	36.36	7.78	0.00	0.00	56.67	0.00	70.00	20.00	34.67
		Ours	**87.47**	92.77	**86.64**	42.31	40.91	7.78	0.00	30.00	63.33	50.00	70.00	30.00	**42.00**
IXC2-VL	7B	Greedy	84.13	83.12	**84.37**	34.62	31.82	11.11	37.50	50.00	43.33	16.67	70.00	30.00	35.33
		Beam	84.13	83.12	**84.37**	34.62	27.27	11.11	37.50	50.00	6.33	33.33	70.00	35.00	**40.00**
		Nucleus	79.53	77.41	80.30	34.62	27.27	16.67	62.50	30.00	53.33	0.00	80.00	25.00	36.67
		Opera	83.47	82.85	83.62	42.31	31.82	16.67	37.50	50.00	53.33	16.67	70.00	25.00	38.67
		VCD	80.17	78.23	80.83	34.62	31.82	38.89	50.00	40.00	40.00	16.67	90.00	35.00	**40.00**
		Ours	**84.30**	86.26	83.86	34.62	27.27	11.11	37.50	6.00	56.67	16.67	70.00	35.00	38.67
LLaVA-NeXT	34B	Greedy	86.50	83.86	87.01	38.46	40.91	16.67	37.50	30.00	60.00	0.00	80.00	35.00	40.67
		Beam	84.13	90.89	82.70	38.46	31.81	22.22	37.50	50.00	60.00	0.00	80.00	30.00	40.67
		Nucleus	84.90	82.78	85.37	34.62	22.73	27.78	25.00	20.00	43.33	0.00	50.00	45.00	33.33
		Opera	-	-	-	-	-	-	-	-	-	-	-	-	-
		VCD	85.20	83.00	85.68	42.31	22.73	22.22	37.50	50.00	46.67	16.67	80.00	40.00	39.33
		Ours	**86.93**	83.82	**87.51**	34.62	36.36	33.33	25.00	50.00	70.00	0.00	70.00	30.00	**42.67**
Yi-VL	34B	Greedy	79.83	87.64	77.50	23.08	8.18	16.67	37.50	30.00	46.67	0.00	70.00	25.00	30.00
		Beam	78.97	89.46	75.74	26.92	18.18	16.67	37.50	30.00	43.33	16.67	70.00	15.00	29.33
		Nucleus	75.30	78.86	73.68	23.08	18.18	16.67	37.50	20.00	30.00	33.33	60.00	25.00	26.67
		Opera	64.50	60.20	70.67	7.69	4.55	0.00	12.50	0.00	16.67	0.00	0.00	10.00	7.33
		VCD	76.70	80.69	75.08	23.08	36.36	3.89	25.00	30.00	33.33	16.67	60.00	30.00	**32.67**
		Ours	**80.17**	87.37	**78.05**	23.08	18.18	27.78	62.50	30.00	35.67	0.00	70.00	30.00	31.33
InternVL	26B	Greedy	85.83	82.83	86.45	42.31	36.36	27.78	25.00	30.00	80.00	33.33	80.00	45.00	48.00
		Beam	86.80	87.45	86.68	38.46	45.45	22.22	37.50	50.00	83.33	50.00	70.00	45.00	50.67
		Nucleus	81.23	79.45	81.76	50.00	31.82	27.78	12.50	60.00	70.00	33.33	60.00	25.00	44.00
		Opera	86.30	84.44	86.66	42.31	27.27	16.67	25.00	30.00	76.67	50.00	70.00	50.00	45.33
		VCD	81.70	80.08	82.18	30.77	36.36	11.11	12.50	50.00	66.67	50.00	50.00	55.00	42.00
		Ours	**86.93**	83.70	**87.53**	42.31	50.00	44.44	12.50	30.00	83.33	50.00	70.00	40.00	**51.33**

environment, holistic description, and others. GPT-4 measures the severity of hallucination in a range of 0 to 7 and the higher score denotes less hallucination.

4.1 Evaluation Results

Overview We summarize our comprehensive experimental results across six LMMs, six decoding methods, on six benchmarks in a spider chart format for visibility of the improvements in Fig. 3. As in the figure, CODE generally shows competent performance and consistent results among different measurements and benchmarks. We delve into each result in detail in the subsections below.

Results on Discriminative Benchmarks We conduct hallucination evaluation across 6 LMM baselines on two discriminative benchmarks. POPE [35] focuses on object-level hallucinations, where its prompts consist of asking about the existence of objects, such as *"Is there* `<something>` *in the image?"*. Here, we exclude trivial splits and test with the most challenging subset: *adversarial* split, which requires models to identify highly-relevant objects to correctly answer yes/no. As shown in Table 1, both the accuracy and F1 score from CODE lead to the best performance in 6 LLM baselines, among 6 decoding methods, which indicates that our decoding method can properly address object hallucination within simple binary question format.

As a more sophisticated visual assessment, we compare our method on MMVP benchmark [53]. Beyond object hallucination, the benchmark is divided into 9 systematic visual patterns within CLIP-blind pairs. These pairs consist of two images that clearly display visual differences, yet CLIP [48] struggles to discriminate between them. The test subject models only receive scores when they correctly identify both images in the pair, requiring precise visual understanding. On the right part of Table. 1, we can observe a remarkable gain in average, indicating that CODE effectively enhances visual consistency through the contrastive mechanism using self-generated descriptions.

Results on GPT-aided Benchmarks Although the discriminative datasets are intuitive benchmarks for evaluating the degree of object-level hallucination, they are limited in measuring a deeper under-

https://doi.org/10.52202/079017-4246

Table 2: GPT-aided evaluation results among 6 LMMs and decoding methods on generative benchmarks (LLaVA-QA90 [42]: score ratio for GPT answer / MMHal-Bench [52]: score rated by GPT).

Model	#Param	Decoding	LLaVA-QA90				MMHal-Bench									
			Conv	Detail	Comp	Overall↑	Attr	Adv	Comp	Count	Rel	Env	Hol	Other	Overall↑	Hal↓
LLaVA-1.5	13B	Greedy	79.5	77.1	90.3	82.4	3.17	1.92	2.42	2.25	1.75	3.50	1.92	2.17	2.39	52.08
		Beam	81.8	77.7	90.7	**83.5**	2.92	1.67	2.92	1.83	1.50	2.92	2.33	2.58	2.33	53.13
		Nucleus	70.4	78.3	89.9	79.3	2.08	0.75	1.83	2.33	1.92	3.42	1.67	2.25	2.03	60.42
		Opera	82.4	73.5	85.4	80.7	3.08	0.75	2.58	1.83	1.92	3.08	2.08	2.42	2.22	55.00
		VCD	72.1	76.4	89.5	79.3	2.83	1.58	2.33	2.42	1.67	2.67	1.83	2.92	2.28	54.00
		Ours	81.8	77.7	90.7	**83.5**	3.00	1.75	2.25	2.17	1.58	3.00	3.25	2.92	**2.49**	**51.00**
Emu2-Chat	14B	Greedy	81.1	59.0	88.1	76.4	3.17	1.08	2.50	2.58	2.17	2.83	2.83	2.83	2.50	39.58
		Beam	81.7	54.0	87.9	74.8	3.00	0.58	1.83	2.75	2.58	3.33	3.17	2.42	2.46	39.58
		Nucleus	78.6	57.9	87.3	74.9	2.50	0.67	2.83	2.83	1.92	3.25	3.00	3.17	2.52	**37.50**
		Opera	82.6	54.5	66.2	68.0	2.75	1.08	1.75	2.75	2.08	2.67	2.92	2.58	2.33	43.75
		VCD	80.7	56.8	87.6	75.5	2.92	0.75	1.83	3.00	2.08	2.92	2.92	2.92	2.42	42.00
		Ours	82.1	59.6	89.5	**77.3**	2.92	1.08	2.42	2.83	2.25	3.33	3.00	2.92	**2.59**	**37.50**
IXC2-VL	7B	Greedy	83.0	81.3	90.3	84.9	3.58	3.08	3.17	2.17	4.08	3.50	2.75	3.08	3.17	29.17
		Beam	79.1	80.2	90.4	83.1	3.50	3.25	2.83	2.25	4.33	3.83	2.75	3.08	3.23	28.13
		Nucleus	86.5	76.5	88.3	84.1	3.25	2.75	2.92	2.92	3.25	3.42	2.92	3.00	3.05	33.00
		Opera	78.8	79.5	86.9	81.7	3.67	3.25	2.92	2.42	4.25	4.17	3.00	3.00	3.33	27.08
		VCD	79.5	82.1	92.0	84.5	3.58	2.42	3.08	2.50	3.33	3.92	2.33	2.75	2.99	32.00
		Ours	85.3	79.4	92.9	**86.0**	3.67	3.25	3.17	3.00	3.67	3.92	3.33	3.67	**3.46**	**25.00**
LLaVA-NeXT	34B	Greedy	80.0	95.7	97.9	90.7	2.75	3.92	3.50	2.75	2.83	4.08	3.08	3.50	3.30	**34.00**
		Beam	87.7	97.2	99.2	94.5	3.33	3.83	2.83	2.58	3.17	3.92	2.67	3.75	3.26	35.42
		Nucleus	83.2	88.7	98.3	90.0	2.50	3.92	3.50	1.33	3.92	4.42	2.33	2.75	3.08	40.63
		Opera	-	-	-	-	-	-	-	-	-	-	-	-	-	-
		VCD	86.4	93.8	96.6	92.1	3.25	3.75	3.08	3.08	2.17	3.92	2.50	3.58	3.16	39.58
		Ours	89.2	99.1	98.3	**95.3**	2.92	4.08	3.92	2.17	2.92	4.67	2.92	3.83	**3.43**	**34.00**
Yi-VL	34B	Greedy	80.2	76.4	88.9	81.9	2.67	0.00	3.08	1.00	2.00	2.25	1.83	1.17	1.75	65.63
		Beam	70.7	75.7	85.9	77.5	2.92	0.42	3.33	0.92	2.25	2.33	1.83	1.58	**1.95**	**60.42**
		Nucleus	76.2	81.2	82.7	79.9	2.33	1.42	2.58	1.08	2.50	2.42	1.33	1.08	1.84	62.50
		Opera	63.6	29.1	16.0	36.3	2.50	1.25	1.83	0.83	1.00	1.83	1.92	1.50	1.58	61.50
		VCD	66.9	76.7	87.3	76.7	2.58	0.33	3.00	1.58	2.42	2.00	1.92	0.92	1.84	62.50
		Ours	73.9	83.6	89.4	**82.1**	2.75	0.58	3.00	1.00	2.50	2.50	1.33	1.00	1.83	65.63
InternVL	26B	Greedy	79.2	93.2	89.1	86.6	3.92	2.58	2.67	2.42	3.17	3.75	3.25	3.42	3.15	33.33
		Beam	83.7	93.3	91.6	89.3	3.83	2.67	3.83	2.50	4.00	3.75	2.75	3.58	3.36	31.25
		Nucleus	75.1	95.1	91.1	86.4	2.92	3.25	2.75	3.50	2.75	3.33	3.33	3.25	3.14	37.50
		Opera	81.0	94.6	91.9	88.7	3.58	2.83	3.00	2.92	4.17	3.83	2.92	3.33	3.32	32.29
		VCD	83.9	92.6	89.5	88.3	3.67	2.17	3.50	2.50	2.67	3.92	2.08	3.00	2.94	42.00
		Ours	83.1	103.3	91.9	**92.2**	4.25	2.92	3.5	2.92	4.17	3.83	2.92	3.67	**3.52**	**30.21**

standing of whether the given LMM responses encompass contextual or sequential hallucination. To probe the expanded effectiveness of our method beyond simple multiple choice tests, we utilize two generative benchmarks that can verify model responses at the sentence level. As in Table. 2, CODE generally outperforms the overall score of LLaVA-QA90 [42], showing up to $+13.7\%$ improvements than other CD methods. Additionally, we compare our models in MMHal-Bench [52] specialized to evaluate hallucination effects sourced from more challenging image-question pairs— OpenImages [50]. As in the result, our method generally not only improves overall average score with consistent results among 6 other baseline LMMs, but also effectively mitigates the hallucination ratio. Through combinatorial results from both discriminative and generative benchmarks, we corroborate the robustness and efficacy of our proposed method in addressing both object-level and contextual hallucinations, thereby ensuring more reliable LMM responses.

4.2 Analyses on CODE

Ablation Study Previous CD-based methods [34, 30, 12] require heuristic choices to control the degree of amplification for logit variation α and penalizing parameter β that filter out the implausible next-tokens in adaptive plausibility constraint. To tackle it, we proposed two regulation methods that can dynamically control information flow in CODE and token candidate pool, respectively: (i) dynamic restriction (DR), α_t in Eqn. 2 and (ii) adaptive information constraint (AIC), β_t in Eqn. 4. To validate the effectiveness of such adaptive regulations built in CODE, we conducted ablation study on them.

Table 3: Ablation study on α_t (DR) and β_t (AIC) for (abbreviated) LV1.5 [40], LV-N [41], IVL1.5 [8] on two benchmarks [53, 42]. We report overall scores for the benchmarks. ✗ indicates fixed hyper-parameter for α and β.

DR	AIC	MMVP			LLaVA-Bench		
		LV1.5	LV-N	IVL1.5	LV1.5	LV-N	IVL1.5
✗	✗	28.67	32.67	45.33	66.4	76.0	81.0
✗	✓	32.00	41.33	48.00	67.9	81.6	83.5
✓	✗	29.33	37.33	48.67	69.0	79.7	82.4
✓	✓	**34.00**	**42.67**	**51.33**	**72.6**	**85.1**	**85.7**

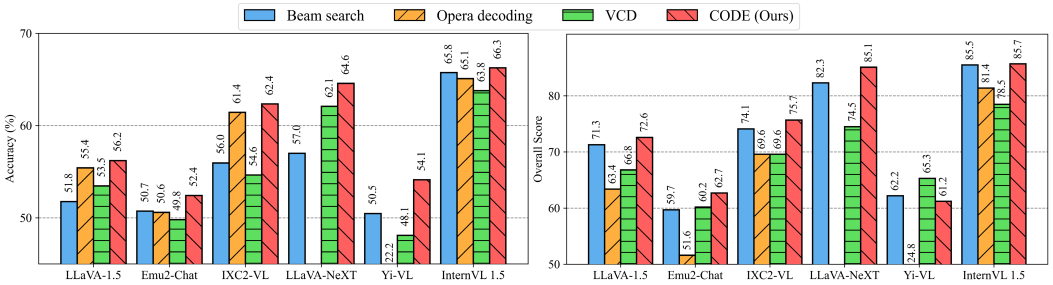
We implement baselines with same CODE framework using self-generated descriptions as visual counterparts, but with default settings of α=1.0 and β=0.1. As in Table. 3, either use of DR or AIC can enhance the benchmarks than the fixed α and β. Our CODE implementation that utilizes both DR and AIC to dynamically restrict information flow exhibits the best results among the baselines.

Question: What is the brand of the yogurt flavored with blueberry?

GT Answer: The brand of the blueberry-flavored yogurt is Fage.

Original Tokens	The	brand	in	the	image	is	from	the	brand	"	Yoplait	…
CODE Tokens	The	yogurt	in	the	image	is	from	the	brand	"	Fage	…
Visual content logit_v	20.81	21.34	20.83	27.03	22.47	22.84	18.42	19.78	20.97	18.63	**15.34**	…
	20.81	21.13	20.83	27.03	22.47	22.84	18.42	19.78	20.97	18.63	15.02	…
Description-only logit_d	16.98	20.14	18.69	25.55	21.00	23.00	16.41	19.23	19.59	17.53	**15.27**	…
	16.98	15.82	18.69	25.55	21.00	23.00	16.41	19.23	19.59	17.53	13.02	…
CODE $\text{logit}_{\text{code}}$	22.01	21.85	22.17	28.34	23.40	22.72	19.78	20.26	22.06	19.42	15.39	…
	22.01	23.38	22.17	28.34	23.40	22.72	19.78	20.26	22.06	19.42	**16.66**	…

Figure 4: An example of token-level case study for CODE. Each row indicates the logit score from visual content logit_v, comprehensive description logit_d, CODE applied $\text{logit}_{\text{code}}$, respectively.



(a) RealworldQA [57] (b) LLaVA-Bench (In-the-Wild) [42]

Figure 5: Additional experiments on In-the-Wild benchmarks. Note that, unlike other datasets, OPERA [23] fails to generate consistent responses in real-world datasets using Yi-VL [62].

Computational Analysis We utilize comprehensive descriptions from models as additional information for contrasting with visual contents, thereby leading to computational loads similar to other decoding methods. To analyze the computation, we compare the token throughput (token/s) and decoding latency (ms/token) with other CD-based methods on 8 NVIDIA RTX A6000 GPUs as in Table. 4.

Table 4: Computational analysis on decoding throughput and latency among CD-based methods. We compare three different model sizes.

	Throughput (token/s)↑			Latency (ms/token)↓		
	VCD	OPERA	CODE	VCD	OPERA	CODE
7B [16]	5.62	1.23	3.66	177.99	809.73	272.92
14B [51]	4.04	1.04	2.82	247.6	960.14	354.09
34B [41]	3.61	oom	2.81	277.27	oom	355.81

Token-level Case Study As illustrated in Fig. 4, to verify whether the proposed CODE effectively mitigates object hallucination, we analyze the output logit values of LMM [41] at the token-level case study, with a greedy search as the baseline. The first two rows in the table indicate the original greedy decoded tokens which are elected based on high logit_v from the visual content and CODE output tokens, respectively. As in the figure, we can observe that visual hallucination occurs at the *"Yoplait"* token highlighted in red. For relatively easy tokens at the beginning of sentence, $\text{logit}_{\text{code}}$ produces identical decisions maintaining consistency with logit_v, which indicates the amplification of logit variation is effectively adjusted due to similar prediction distributions from visual contents and description-only information. However, at the hallucination-occurred time step, logit scores are deviated between the two information, resulting in a more confusing state to identify between GT token *"Fage"* and hallucinatory *"Yoplait"*. In our framework, *"Fage"* is relatively more amplified from 15.02 to 16.66 than *"Yoplait"*, which changes from 15.34 to 15.30. By simultaneously considering both token-level and distributional prediction over the vocabulary, CODE changes the wrong next-token output to correct one, mitigating hallucination. For more case studies, please refer Appendix. D.

Additional Experiments on In-the-Wild Contemporary open-sourced LMMs are fine-tuned with various combinations of vision-language datasets [7, 28, 25], mostly composed of COCO-sourced visual images [38] and their curated instruction. Although the existing hallucination benchmarks intentionally convert question queries to assess model robustness against inconsistency, the visual contents in benchmarks are limited to in-distribution COCO images. To validate our method in more challenging and real-world scenarios, we compared baselines on LLaVA-Bench (In-the-wild) [42] and RealworldQA [57] as in Fig. 5 and achieved competent performance (case studies in Appendix. F).

https://doi.org/10.52202/079017-4246

5 Discussion and Limitation

Albeit the computational analysis in Table. 4, as one of limitations, our contrastive decoding method requires additional computational resources than the use of vanilla decoding. However, considering an essential ongoing research topics and developments [60, 15] aimed at mitigating the negative effects of hallucination problems in both LLMs and LMMs, our work contributes important societal impacts towards more real-world applicability and robust AI system.

6 Conclusion

We present COuntering DEscription Contrastive Decoding (CODE), a novel and training-free decoding method to mitigate hallucination in Large Multi-modal Models. By utilizing self-generated descriptions as corrective references during the decoding phase, CODE dynamically adjusts the information flow for next-token predictions, enhancing the coherence and informativeness of responses while reducing the cross-modal inconsistency. Extensive experiments demonstrate that CODE effectively decreases hallucinations across various benchmarks and contemporary LMMs, significantly improving contextual relevance and response alignment with visual contents.

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).

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

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

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. *arXiv preprint arXiv:2312.06742*, 2023.

[6] Min Chen and Mateu Sbert. On the upper bound of the kullback-leibler divergence and cross entropy. *arXiv preprint arXiv:1911.08334*, 2019.

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

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

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[10] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024.

[11] 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. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[12] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024.
[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.
[14] Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*, 2024.
[15] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.
[16] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.
[17] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
[18] Google. Bard. `https://bard.google.com/`, 2023.
[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
[20] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.
[21] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1439–1449, 2021.
[22] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.
[23] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*, 2023.
[24] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36, 2024.
[25] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.
[26] Junho Kim, Yeon Ju Kim, and Yong Man Ro. What if...?: Counterfactual inception to mitigate hallucination effects in large multimodal models. *arXiv preprint arXiv:2403.13513*, 2024.
[27] Yeonju Kim, Junho Kim, Byung-Kwan Lee, Sebin Shin, and Yong Man Ro. Mitigating dataset bias in image captioning through clip confounder-free captioning network. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1720–1724. IEEE, 2023.
[28] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
[29] Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. Moai: Mixture of all intelligence for large language and vision models. *arXiv preprint arXiv:2403.07508*, 2024.
[30] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023.
[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*. PMLR, 2023.
[32] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

 https://doi.org/10.52202/079017-4246

[33] Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *arXiv preprint arXiv:2405.05949*, 2024.

[34] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312. Association for Computational Linguistics, July 2023.

[35] 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*, pages 292–305, Singapore, Dec. 2023. Association for Computational Linguistics.

[36] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

[37] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024.

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

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

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

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

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

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

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

[45] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024.

[46] OpenAI. ChatGPT. `https://openai.com/blog/chatgpt/`, 2023.

[47] OpenAI. Gpt-4 technical report, 2023.

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

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

[50] H Rom, Neil Alldrin, J Uijlings, I Krasin, et al. The open images dataset v4. *Int J Comput Vis.*, 128:1956–1981, 2020.

[51] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. 2023.

[52] 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. Aligning large multimodal models with factually augmented RLHF, 2024.

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

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

[55] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

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

[57] xAI. Grok-1.5 vision preview., 2024.

[58] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

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

[60] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[61] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[62] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024.

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

[64] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023.

[65] Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, pages arXiv–2310, 2023.

[66] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

[67] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*, 2023.

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

[69] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*, 2024.

 https://doi.org/10.52202/079017-4246

A Instruction for Comprehensive Description

In generating a comprehensive description for the given visual content, we aim to obtain as detailed a description as possible, ensuring that the response fully spans the visual representation space, even though it may not be entirely feasible as discussed in 3.1. The specific prompt instruction used to describe the image contents in detail is described in Table. 5.

Comprehensive Description Prompt:
Provide a detailed description of the image, covering all visible elements and their interactions, so as to thoroughly answer any potential questions about the image.

Table 5: A simple instruction prompt for generating comprehensive description.

B Detailed Algorithm for CODE

We describe the complete details of CODE implementation for better understanding in Algorithm. 1.

Algorithm 1 COuntering DEscription Contrastive Decoding

Require: LVLM M_θ, Visual Content v, Text Query x, Comprehensive Description Prompt x_0, Target Token Length T

$d \leftarrow M_\theta(v, x_0)$ ▷ Step 1: Generate comprehensive description d for the given visual content
Initialize $t \leftarrow 1$
while $t < T$ **do** ▷ Step2: CODE decoding
  $p_{\text{v}} = \text{Softmax}\left[\text{logit}_\theta(y_t \mid v, x, y_{<t})\right]$ ▷ Compute p_v
  $p_{\text{d}} = \text{Softmax}\left[\text{logit}_\theta(y_t \mid d, x, y_{<t})\right]$ ▷ Compute p_d
  $\mathcal{D}_{\text{bd}}(p_{\text{v}} \| p_{\text{d}}) = \frac{1}{2}\sum_{i=1}^{n}(p_{v,i} + p_{d,i})\log_2(|p_{v,i} - p_{d,i}|^k + 1)$ ▷ Bounded Divergence $\mathcal{D}_{\text{bd}}$
  Set $\alpha_t \leftarrow 1 - \mathcal{D}_{\text{bd}}(p_{v,t} \| p_{d,t})$ ▷ Set α_t for Dynamic Restriction
  Set $\beta_t \leftarrow \mathcal{D}_{\text{bd}}(p_{v,t} \| p_{d,t})$ ▷ Set β_t for Adaptive Information Constraint
  $\mathcal{V}_{\text{head}}(y_{<t}) = \{y_t \in \mathcal{V} : p_\theta(y_t \mid v, x, y_{<t}) \geq \beta_t \max_w p_\theta(w \mid v, x, y_{<t})\}$ ▷ Set $\mathcal{V}_{\text{head}}$
  $p_{\text{code}} = \text{Softmax}[(1 + \alpha_t)\text{logit}_\theta(y_t \mid v, x, y_{<t}) - \alpha_t \text{logit}_\theta(y_t \mid d, x, y_{<t})]$
  $p_{\text{code}}(y_t \mid y_{<t}) = 0, \text{if } y_t \notin \mathcal{V}_{\text{head}}$ ▷ Apply Adaptive Information Constraint
  $y_t = \text{argmax}(p_{\text{code}})$ ▷ Select Token
  Set $t \leftarrow t + 1$
end while

C Benchmark Details

Discriminative Benchmarks

- **POPE** [35] is a widely used benchmark designed for evaluating object-level hallucination, which can be split into the three subset categories based on how to select object replacements: (i) random, randomly sampled objects (ii) popular, top-k frequent objects not existing in the image, and (iii) adversarial, top-k objects that have high co-occurrence. The number of images is 500 and each image has 6 questions along with the subsets, making a total of 9000 images. In this work, we only consider adversarial split, which is the most challenging subset in POPE benchmark.
- **MMVP** [53] includes 300 images with 9 different visual patterns that CLIP model [48] struggles to identify the visual differences (CLIP-paired images): 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 (📷). It follows multiple selection tests, but uses GPT-4 to map the model response to the answer options.
- **RealworldQA** [57] is recently introduced benchmarks for evaluating basic real-world understanding for multi-modal models. It consists of total 765 anonymized outdoor (mostly taken from vehicles) and indoor images with multiple selection questions.

Generative Benchmarks

- **LLaVA-QA90 & LLaVA-Bench (In-the-Wild)** [42] consist of three subset response types for each image: (i) Conversation, which is conversation format between the user and assistant answering the vision-related questions for the given images, (ii) Detailed description, which requires detailed description for the given image scene, and (iii) Complex reasoning, which involves in-depth reasoning questions for the image. The former benchmarks sourced from COCO images (total 30 images with 90 questions), while the latter benchmarks are gathered from web for challenging domain situations (total 24 images with 60 questions).
- **MMHal-Bench** [52] is specially focused on penalizing hallucinations. It has total 96 image-question pairs composed with 8 question categories for 12 objects: Object attribute (Attr), Adversarial object (Adv), Comparison (Comp), Counting (Count), Spatial relation (Rel), Environment (Env), Holistic description (Hol), and Others (Other). As like in the above LLaVA-Bench benchmarks, MMHal-Bench also utilize GPT-4 to analyze and rate the model responses and score in a range of 0 to 7.

D Additional Token-level Case Study

As additional token-level case studies, we explore how the logit information changes during our CODE decoding phase in other examples. As a first example, LLaVA-NeXT [41] struggles to distinguish between Haleakala National Park and Diamond Head, both located in Hawaii, and predicts the former during inference with the vanilla decoding method. Using our CODE decoding method, as shown in Fig. 6, the information flow of *"Haleakala"* token is curbed, inducing a token inversion to *"Diamond"*, which matches the ground truth word. This occurs because the logit variation is dynamically adjusted based on both token-level and distributional information.

Question: What is the name of this famous sight in the photo?

GT Answer: The famous sight in the photo is Diamond Head.

Original Tokens	The	famous	sight	in	the	photo	is	Haleakala	…
CODE Tokens	The	famous	sight	in	the	photo	is	Diamond	…
Visual content logit_v	23.42	20.89	24.78	26.45	30.55	28.16	27.05	**18.70**	…
	23.42	20.89	24.78	26.45	30.55	28.16	27.05	17.45	…
Description-only logit_d	18.89	19.06	23.81	24.86	28.64	26.34	24.23	**17.95**	…
	18.89	19.06	23.81	24.86	28.64	26.34	24.23	11.86	…
CODE $\text{logit}_{\text{code}}$	25.57	21.91	25.66	27.89	32.29	29.95	28.58	19.32	…
	25.57	21.91	25.66	27.89	32.29	29.95	28.58	**22.06**	…

Figure 6: Additional token-level case study for analysis.

In another example illustrated in Fig. 7, we show how the adaptive information constraint prevent from rewarding implausible tokens, thus suppress hallucinatory prediction during contrastive decoding. The original tokens from logit_v predict the correct answer at the hallucinatory time step, highlighted in bold. In this case, the original prediction should be preserved, and token inversion should not occur. By our CODE decoding method, β_t dynamically controls the adaptive information constraint, so that those hallucination tokens'(*i.e., "four"* and *"dragon"*) logit values are cut off to $-\infty$ and removed from candidate token pool.

E Further Discussion on Broader Impact

We proposed CODE, which can be seamlessly integrated into LMMs without additional training. There is still a lot of room for progress and mitigation of hallucination issues, as our method cannot assure 100% removal of hallucinations. However, by providing more coherent and contextually accurate responses, our work can potentially be integrated into real-world applications, making user interactions with AI in customer service, education, and personal assistance more effective and satisfying in the near future.

Furthermore, enhanced accuracy and reliability using our method can reduce hallucinations in LMMs, improving the accuracy of AI-generated descriptions in critical fields such as autonomous

https://doi.org/10.52202/079017-4246

Question: How many uncut fruits are in the image?
GT Answer: There are three uncut fruits in the image.

β_t	0.27	0.01	**0.80**	0.11	0.00	**0.80**	0.05	0.00	0.01	0.18
Original Tokens	There	are	three	unc	ut	fruits	in	the	image	.
Description Tokens	There	are	four	unc	ut	dragon	in	the	image	.
Visual content logit_v	22.56	26.53	23.09	22.16	29.47	21.17	25.41	30.00	27.42	25.50
	22.56	26.53	21.66	22.16	29.47	14.63	25.41	30.00	27.42	25.50
Description-only logit_d	20.47	27.09	21.59	23.95	28.20	20.75	25.34	28.36	26.50	21.59
	20.47	27.09	24.28	23.95	28.20	21.80	25.34	28.36	26.50	21.59
CODE Tokens	There	are	three	unc	ut	fruits	in	the	image	.
CODE $\text{logit}_{\text{code}}$	24.10	25.98	23.39	20.56	30.73	21.25	25.47	31.63	28.34	28.69
	24.10	25.98	-inf	20.56	30.73	-inf	25.47	31.63	28.34	28.69

Figure 7: Additional token-level case study for analysis.

driving, robotics, healthcare, and augmented reality. This advancement not only enhances practical applications but also significantly benefits the research community working on hallucination, an area that is not yet fully explored. Our contributions can help pave the way for deeper understanding and new research directions to address these challenges, for more trustworthy AI systems.

F Additional Qualitative Results

Discriminative Capability Case Study: MMVP.

Image:

Question:
Is the following statement correct: In the image, there is one daisy that is under the shadow of a taller daisy.
(a) Correct (b) Incorrect

Ground-truth:
(b)

InternLM-XComposer2 (Opera):
The answer is (a).

InternLM-XComposer2 (VCD):
The answer is (a).

InternLM-XComposer2 (Ours):
The answer is (b).

Image:

Question:
In the image, is the snake's head oriented facing upwards or directed forwards?
(a) Upwards (b) Forward

Ground-truth:
(a)

InternVL 1.5 (Opera):
The snake's head is oriented (b) forward.

InternVL 1.5 (VCD):
The snake's head is directed forwards, so the correct answer is (b) Forward.

InternVL 1.5 (Ours):
The snake's head is oriented facing upwards, so the correct answer is (a) Upwards.

https://doi.org/10.52202/079017-4246

Image:

Question:
Is the person in the picture already standing on the weighing scale?
(a) Yes (b) No

Ground-truth:
(b)

InternVL 1.5 (Opera):
(a) Yes

InternVL 1.5 (VCD):
Yes, the person in the picture is already standing on the weighing scale.

InternVL 1.5 (Ours):
The person in the picture is not yet standing on the weighing scale. The answer is (b) No.

Image:

Question:
Is the needle pointing up or down?
(a) Up (b) Down

Ground-truth:
(a)

LLaVA-NeXT (VCD):
The needle of the syringe is pointing down.

LLaVA-NeXT (Ours):
The needle in the image is pointing up.

Image:

Question:
Based on the image, is the following statement correct: We cannot see the window on the school bus?
(a) Correct (b) Incorrect

Ground-truth:
(a)

LLaVA-NeXT (VCD):
(b) Incorrect. The window of the school bus is visible in the image, located just above the "School Bus" sign.

LLaVA-NeXT (Ours):
The statement "We cannot see the window on the school bus" is (a) Correct. In the image provided, there is no visible window on the school bus. The view is focused on the back of the bus where the stop sign and the words "SCHOOL BUS" are visible, but the window is not shown.

Table 6: More case studies within MMVP, using InternLM-XComposer2, InternVL 1.5 and LLaVA-NeXT.

https://doi.org/10.52202/079017-4246

Discriminative Capability Case Study: RealworldQA.

Image:

Question:
How many circular objects are in the image?
A. There are 2 circular objects.
B. There is 1 circular object.
C. There are 3 circular objects.
Please answer directly with only the letter of the correct option and nothing else.

Ground-truth:
A

InternLM-XComposer2 (Opera):
The answer is C.

InternLM-XComposer2 (VCD):
The answer is C.

InternLM-XComposer2 (Ours):
A.

Image:

Question:
How many animals are in this picture?
A. There are eight animals in this picture.
B. There are nine animals in this picture.
C. There is one animal in this picture.
Please answer directly with only the letter of the correct option and nothing else.

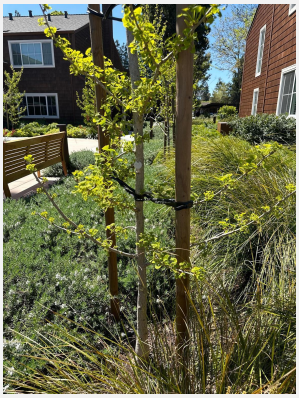
Ground-truth:
B

InternVL 1.5 (Opera):
A

InternVL 1.5 (VCD):
A

InternVL 1.5 (Ours):
B

Image:



Question:
Where is the black material relative to the tree?
A. The black material is at the top of the tree.
B. The black material is at the midpoint or halfway on the tree.
C. There is no black material.
Please answer directly with only the letter of the correct option and nothing else.

Ground-truth:
B

LLaVA-NeXT (VCD):
A

LLaVA-NeXT (Ours):
B

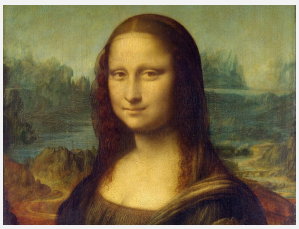
Table 7: More case studies within RealworldQA, using InternLM-XComposer2, InternVL 1.5 and LLaVA-NeXT.

https://doi.org/10.52202/079017-4246

Generative Capability Case Study: LLaVA-Bench (In-the-Wild).

Question: Describe this photo in detail.

Image:



InternLM-XComposer2 (Opera):
The painting you're referring to is the famous "Mona Lisa" by Leonardo da Vinci. It is an oil painting on a poplar wood panel. The artwork depicts a woman, believed to be Lisa Gherardini, the wife of a Florentine merchant named Francesco del Giocondo. The woman is portrayed in a three-quarter view, with her left hand resting on her lap and her right hand holding a veil. She is wearing a dress with a high collar and a pearl necklace. The background features a landscape with mountains and a body of water. The painting is known for its enigmatic smile and the use of sfumato, a painting technique that creates a soft, hazy effect. The "Mona Lisa" is considered a masterpiece of the Italian Renaissance and is housed in the Louvre Museum in Paris, France.

InternLM-XComposer2 (VCD):
The image features a painting of the renowned Mona Lisa, an oil on wood artwork. The central figure is a woman with long red hair and a headband, wearing a blue dress adorned with gold accents. Her face is captured in close-up detail, showcasing her distinct smile and enigmatic gaze. The background is a landscape painted in earth tones, adding depth to the composition. This painting is widely recognized as a masterpiece, capturing the essence of the Renaissance era with its realistic portrayal and intricate details.

InternLM-XComposer2 (Ours):
The painting you're referring to is the famous "Mona Lisa" by Leonardo da Vinci. It is an oil painting on a wooden panel, measuring approximately 77 x 53 centimeters. The painting is a portrait of a woman, who is the focal point of the artwork. She is depicted wearing a dress and a headscarf, with her hair styled in a manner typical of the Italian Renaissance period. The background of the painting features a landscape with mountains and a body of water, adding depth and context to the scene. The painting is known for its realistic portrayal of the woman's features, her enigmatic smile, and the subtle interplay of light and shadow. The "Mona Lisa" is considered a masterpiece of the High Renaissance period and is housed in the Louvre Museum in Paris, France.

https://doi.org/10.52202/079017-4246

Question: Describe this photo in detail.

Image:

InternVL 1.5 (Opera):

The image shows a group of star fruit, also known as carambola or tamarind, arranged and presented. These are a kind a of sweet, fizzi, and tart tasting, and are a popular fruit. There appears a few different ways they're displayed, some whole or unpeel, some are half-cut open. There is a clear visual focus and contrast, as well as a natural and organic feel to the image.

InternVL 1.5 (VCD):

This image depicts several purple-colored fruits, which appear to be durian. Durian is a tropical fruit that is known for its strong, pungent aroma and creamy, sweet flesh. In the center of the image, one of the durians is cut in half, revealing the fruit's inner flesh, which is a creamy white with a star-like pattern of seeds. The cut side of the durian shows the texture of the flesh and the arrangement of the seeds within.

The fruit has a distinctive appearance with a thick, spiky green outer shell that protects the flesh inside. The surface of the fruit is textured and somewhat rough, with the color transitioning from deep purple to green at the points where the spiky projections emerge. The durians are resting on a wooden surface, which provides a rustic backdrop that contrasts with the vibrant colors of the fruit.

In the background, there is a wicker basket containing more of these fruits, suggesting that they have been gathered for consumption or sale. The basket is made of natural materials and complements the overall natural, organic feel of the image. The lighting in the photo is soft and natural, which highlights the textures and colors of the durians and adds to the overall appeal of the image.

Overall, this image presents the durian fruit in a way that showcases both its visual and textural qualities, as well as its role in a natural, possibly agricultural, setting.

InternVL 1.5 (Ours):

The image shows several whole and cut open mangosteens on a wooden surface. Mangosteens are tropical fruits known for their distinctive appearance and sweet, slightly acidic flavor. The fruit has a dark purple, almost black, outer skin that contrasts with the green, bulbous protrusions at the top, which are the remnants of the fruit's calyx. One of the mangosteens is cut in half, revealing the white, segmented flesh inside, which is the edible part of the fruit. The segments are arranged in a flower-like pattern around a central seed pod. The cut-open fruit provides a clear view of the fruit's structure and the contrast between the outer skin and the inner flesh. The wooden surface and the natural lighting give the image a rustic and organic feel, emphasizing the freshness of the fruit.

https://doi.org/10.52202/079017-4246

Question: Describe this photo in detail.

Image:

LLaVA-NeXT (VCD):

In the center of the image, a black bowl with vibrant red and yellow trim commands attention. It cradles a hearty serving of noodles, their creamy yellow hue contrasting with the bowl's dark color. A few pieces of meat, their rich brown tones adding to the variety of colors, are scattered throughout the bowl. Green peas peek out from under the noodles, adding a touch of freshness to the dish. To the left of the bowl, a small white cup filled with a green sauce waits patiently. Its color is a soothing contrast to the bold colors of the bowl and its contents. In the top right corner of the image, a white plate holds a small pile of chopped vegetables. Their bright green color stands out against the plate, drawing the eye to the top right corner of the image.
The arrangement of these objects creates a visually pleasing scene. The bowl of noodles is centrally placed, making it the focal point of the image. The green sauce in the white cup and the plate of chopped vegetables are positioned on either side, framing the bowl and creating a balanced composition. The colors, objects, and their relative positions all contribute to a harmonious and appetizing scene.

LLaVA-NeXT (Ours):

In the heart of the image, a black bowl with a vibrant red and yellow border takes center stage. It cradles a hearty serving of udon noodles, their light brown hue contrasting beautifully with the bowl's dark exterior. The noodles are not alone; they are accompanied by a generous portion of green onions, adding a pop of color and hinting at the freshness of the dish.
To the right of the bowl, a white cup filled with a green sauce waits patiently, ready to enhance the flavors of the udon noodles. The cup's pristine white color provides a stark contrast to the green sauce within, making it an eye-catching element in the composition.
In the background, a blue and white plate holds a serving of meat and green onions. The plate's colors echo those of the bowl, creating a harmonious color palette across the image.
Scattered around these main elements are various condiments and utensils, each playing their part in this culinary tableau. Their presence adds depth to the image, suggesting a meal ready to be enjoyed.
Overall, this image captures a moment of anticipation, just before the enjoyment of a delicious meal begins.

Table 8: More case studies within LLaVA-Bench (In-the-Wild), using InternLM-XComposer2, InternVL 1.5 and LLaVA-NeXT. We marked the expressions where hallucinations occurred in red.

G Failure Cases

As discussed in Appendix. E, even if CODE shows competent performance along various benchmarks and LMM baselines. The hallucination cannot be 100% eliminated. In this section, we attached some of failure cases to shed some light for future work direction. As illustrated in Table. 8, the baseline models fail to correct hallucination even with our CODE method. Upon closer examination of the model responses, it is evident that the hallucinatory responses tend to be biased towards language priors such as *"strawberry-flavored"* or *"holding a glass of beer"*. Our approach mainly uses self-generated description as contrasting reference (*i.e.,* close to the concept of self-correction), thus the strong assumption is on that the amateur model (comprehensive description) should generate not too much deviated responses from true answers.

In the failure examples, we can infer that the reliance on self-generated descriptions may not always suffice, especially when the descriptions themselves are biased or inaccurate. This indicates a need for integrating more robust mechanisms to verify and correct these biases. Additionally, enhancing the model's understanding and processing of visual content could help mitigate such issues. Future work could explore the integration of external knowledge sources and more sophisticated bias detection techniques to further reduce hallucinations and improve the overall accuracy and reliability of LMMs.

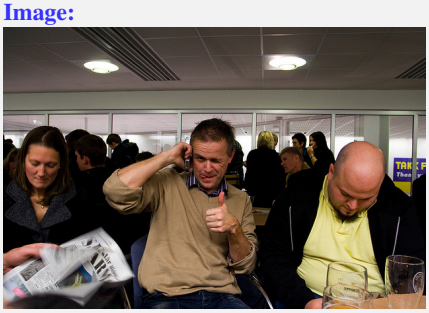
Image:

Question:
Is there any strawberry-flavored yogurt in the fridge?

Generated Text:
The image showcases the inside of a refrigerator, filled with various food items. On the top shelf, there is a carton of milk and a container of yogurt. The middle shelf contains a bag of carrots, a package of strawberries, and a blue plastic container. The bottom shelf holds a carton of blueberries and another carton of yogurt. Additionally, there are two bottles on the door of the refrigerator.

InternLM-XComposer2 (Ours):
Yes, there is a container of strawberry-flavored yogurt in the fridge.

Image:



Question:
What is the man doing in the image?

Generated Text:
The image depicts a group of people gathered in a room, with some sitting and others standing. There are at least 12 people in the scene, engaged in various activities. One man is sitting at a table, holding a cell phone, while another man is sitting nearby, giving a thumbs-up gesture. There are several chairs in the room, with one near the center and another towards the right side. A dining table is also visible in the scene, with a cup placed on it. The people in the room appear to be enjoying their time together, possibly at a social event or gathering.

LLaVA-1.5 (Ours):
The man in the image is sitting at a table, holding a glass of beer, and giving a thumbs-up sign.

Figure 8: Failure case of InterLM-XComposer2 in LLaVA-Bench (In-the-Wild) and LLaVA-1.5 in LLaVA-QA90.

https://doi.org/10.52202/079017-4246

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: They are included in this paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include discussion section for potential limitation of our work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not involve theoretical result to prove.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Implementation section is included, we will open the source to public.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-4246

Answer: [Yes]

Justification: We will open to public in next step.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we have baselines and benchmarks explanations in the manuscript.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Our experiments do not need statistical results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the computational details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We reviewed CoE.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We add it in discussion section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

https://doi.org/10.52202/079017-4246

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This is not relevant with our work.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Our work do not include any license assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowd-sourcing data.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No crowd-sourcing data.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

 https://doi.org/10.52202/079017-4246