ONLY: One-Layer Intervention Sufficiently Mitigates Hallucinations in Large Vision-Language Models

Supplementary Material

This supplementary document is organized as follows:

- The intuitive and theoretical explanation for our motivation is provided in Section A.
- Additional experimental details, including further implementation details, descriptions of other implemented baselines, and license information for the utilized code and datasets, are provided in Section B.
- Additional experimental results on different benchmarks are presented in Section C.
- Additional ablation studies with different parameters are presented in Section D.
- More case studies and GPT-4V-aided evaluations are provided in Section E.
- Potential directions for future work are discussed in Section F.

A. More Explanation on Motivation

A.1. Intuitive Explanation for TVER

Our method is motivated by a principle in information theory: $H(x|y) \leq H(x)$. Let $H(\mathcal{T})$ and $H(\mathcal{V})$ denote the entropy of pure textual and visual attention, respectively. During LVLM decoding, since the model processes both image and text simultaneously, we treat the attention distributions as conditioned on the other modality. This leads to an approximate theoretical form of Eq. (11): $\text{TVER} = \frac{H(\mathcal{T}|\mathcal{V})}{H(\mathcal{V}|\mathcal{T})}$. Since $H(\mathcal{T}|\mathcal{V}) \leq H(\mathcal{T})$ and $H(\mathcal{V}|\mathcal{T}) \leq H(\mathcal{V})$, a higher $H(\mathcal{T}|\mathcal{V})$ indicates behavior closer to purely textual inference, while higher $H(\mathcal{V}|\mathcal{T})$ suggests reliance on visual priors. To approximate the noisy branch used in VCD and M3ID, we aim to enhance textual focus and suppress visual focus, which motivates maximizing TVER for effective textual enhancement.

B. More Experimental Details

B.1. Benchmarks and Metrics

We conduct extensive experiments on the following benchmarks:

- **POPE [19]** is a popular benchmark for assessing object hallucinations in LVLMs. It tests the models with yes-or-no questions regarding the presence of specific objects, such as, "`Is there a {object} in the image?`" The images from the benchmark derive from three existing datasets: MSCOCO [20], A-OKVQA [29], and GQA [13], and comprises three distinct subsets—*random*, *popular*, and *adversarial*—based on how the negative samples are generated. For each dataset setting, the benchmark provides 6 questions per image, resulting in 3,000 test instances. We evaluate the performance of different methods using four metrics: accuracy, precision, recall, and F1 score.
- **CHAIR [28]** evaluates object hallucinations through image captioning, where the LVLMs are prompted to describe 500 randomly selected images from the MSCOCO validation set. The performance is evaluated based on two metrics:

$$\text{CHAIR}_I = \frac{\text{\# hallucinated objects}}{\text{\# all objects mentioned}}, \tag{21}$$

$$\text{CHAIR}_S = \frac{\text{\# sentences with hallucinated object}}{\text{\# all sentences}}. \tag{22}$$

- **MME-Hallucination [11]** is a comprehensive benchmark consisting of four subsets: *existence* and *count* for object-level hallucinations, and *position* and *color* for attribute-level hallucinations. Each subset includes 30 images and 60 questions, with two questions per image. Similar to POPE [19], the benchmark includes yes-or-no questions, and performance is assessed based on binary accuracy. Following the official implementation, the reported score is calculated by combining accuracy and accuracy+, where accuracy is based on individual questions, and accuracy+ is based on images where both questions are answered correctly.
- **MMBench [25]** is a comprehensive benchmark designed to evaluate LVLMs' multimodal understanding and reasoning abilities. It emphasizes tasks that require integrating visual and textual information, assessing a model's performance in diverse, real-world scenarios. MMBench employs a hierarchical ability taxonomy, categorizing Perception and Reasoning as Level-1 (L-1) abilities. This taxonomy is further refined into six Level-2 (L-2) dimensions and twenty Level-3 (L-3) dimensions, providing a detailed framework for assessment.
- **MMVP [32]** is a benchmark designed to assess the fine-grained visual recognition capabilities of LVLMs using CLIP-blind pairs. It comprises 150 image pairs, each paired with a binary-option question. Each image is evaluated separately, and an LVLM's response is deemed correct only if it answers both questions associated with a pair accurately.
- **MM-Vet [39]** is a benchmark for evaluating LVLMs on complex tasks. It defines 6 core vision-language capabil-

ities, including recognition, OCR, knowledge, language generation, spatial awareness, and math. An LLM-based evaluator is used to ensure consistent evaluation across diverse question types. The dataset includes 187 images from various online sources and collects 205 questions, each of which requires one or more capabilities to answer.

- **LLaVA-Bench**[1] includes 24 images depicting complex scenes, memes, paintings, and sketches, accompanied by 60 challenging questions. Selected examples from this dataset are used for qualitative comparisons of responses generated by different decoding methods. Additionally, following Yin et al. [38], we evaluate the accuracy and level of detail in the generated responses using the advanced LVLM, GPT-4V[2].

B.2. More Implementation Details

In our experiments, we adhere to the default query format for the input data used in both LLaVA-1.5 [21], InstructBLIP [9], and Qwen-VL [1]. We set $\alpha_1 = 3$, $\alpha_2 = 1$ by default in our decoding process. Additionally, we set $\gamma = 0.2$ for LLaVA-1.5 and $\gamma = 0.4$ for InstructBLIP/Qwen-VL. We follow VCD [16] to implement adaptive plausibility constraints [18]:

$$p_\theta(y_t) = 0, \quad \text{if } y_t \notin \mathcal{S}(y_{<t}), \tag{23}$$

where $\mathcal{S}(y_{<t}) = \{y_t \in \mathcal{S} : p_\theta(y_t|v, \mathbf{x}, \mathbf{y}_{<t}) \geq \beta \max_w p_\theta(w|v, \mathbf{x}, \mathbf{y}_{<t})\}$. Here, $\mathcal{S}$ is the whole vocabulary of LVLM, and hyperparameter $\beta \in [0, 1]$ controls the truncation of the next token distribution. A larger β indicates more aggressive truncation, keeping only the high-probability tokens. In our implementation, we set the logits for $y_t \notin \mathcal{S}(y_{<t})$ to $-\infty$. By default, we set $\beta = 0.1$ for all tasks. All experiments are conducted on a single 48GB NVIDIA RTX 6000 Ada GPU.

B.3. Pilot Study Details

For Figure 4, we visualize 500 images from the CHAIR [28] benchmark (left) and 3,000 images from POPE [19] (right). For Figure 3, we analyze 3,000 POPE images to examine the relationship between entropy deviation and noise level.

B.4. Devision of Textual and Visual Tokens

In Eq. 8, textual and visual attention are obtained based on the indices corresponding to each modality. The index ranges for both modalities are listed below:

- LLaVA-1.5 [21]:
  Textual indices – [0:35], [611:]; Visual indices – [35:611].
- InstructBLIP [9]:
  Textual indices – [32:]; Visual indices – [0:32].
- Qwen-VL[1]:
  Textual indices – [257:]; Visual indices – [1:257].

B.5. Details of Other Baselines

In this work, we mainly compare the performance of our ONLY with two state-of-the-art contrastive-decoding approaches: VCD [16] and M3ID [10]. The method and implementation details for these approaches are provided below:

- **VCD [16]** contrasts output distributions derived from original and distorted visual inputs. Specifically, given a textual query x and a visual input v, the model generates two distinct output distributions: one conditioned on the original v and the other conditioned on the distorted visual input v', which is obtained by applying pre-defined distortions (e.g., Gaussian noise mask) to v. Then, a new contrastive probability distribution is computed as:

$$p_{vcd}\left(y_t\right) = \text{softmax}[(1+\alpha) f_\theta\left(y|v, \mathbf{x}, \mathbf{y}_{<t}\right) - \alpha f_\theta\left(y|v', \mathbf{x}, \mathbf{y}_{<t}\right)]. \tag{24}$$

  In our implementation, we follow the default setting in VCD [16] and set $\alpha = 1$ for reproduction. To generate v', we use a total of 500 noise steps.
- **M3ID [10]** contrasts output distributions derived from original visual inputs with those from pure text inputs, which lack visual information. The final probability distribution is given by:

$$p_{m3id}\left(y_t\right) = \text{softmax}[f_\theta\left(y|v, \mathbf{x}, \mathbf{y}_{<t}\right) + \frac{1-e^{-\lambda t}}{e^{-\lambda t}}\left(f_\theta\left(y|v, \mathbf{x}, \mathbf{y}_{<t}\right) - f_\theta\left(y|\mathbf{x}, \mathbf{y}_{<t}\right)\right)]. \tag{25}$$

  Following their recommended best practice, we set the hyperparameter λ, which balances the conditioned and unconditioned models, to 0.02.

B.6. Dataset and Code Licensing

Datasets. We list the known license information for the datasets below: POPE [19] and MMVP [32] benchmarks are licensed under MIT License. CHAIR [28] is made available under the BSD 2-Clause License. LLaVA-Bench is available under Apache-2.0 License. MME-Hallucination [11] benchmark dataset is collected by Xiamen University for academic research only. MM-Vet [39] dataset is under the CC BY-NC 4.0 license.

Code. In this work, we also use some code implementations from the existing codebases: LLaVA [21] and VCD [16] are licensed under the Apache-2.0 License. InstructBLIP [9] is under BSD-3-Clause License. Qwen-VL [1] is under the Tongyi Qianwen License.

[1] https://huggingface.co/datasets/liuhaotian/llava-bench-in-the-wild.

[2] https://openai.com/index/gpt-4v-system-card.

Table C1. **Results on MME-Hallucination [11] benchmark.** We report the average MME scores along with the standard deviation across three random seeds for each subset. We also report the total scores achieved by the different methods across all four subsets in the final column. Higher scores (↑) indicate better performance. The best results are **bolded**, and the second-best are underlined.

Model	Method	Object-level		Attribute-level		Total Score ↑
		Existence ↑	Count ↑	Position ↑	Color ↑	
LLaVA-1.5	Regular	173.75 (±4.79)	121.67 (±12.47)	117.92 (±3.69)	149.17 (±7.51)	562.50 (±3.96)
	DoLa	176.67 (±2.89)	113.33 (±10.41)	90.55 (±8.22)	141.67 (±7.64)	522.22 (±16.78)
	OPERA	183.33 (±6.45)	137.22 (±6.31)	122.78 (±2.55)	155.00 (±5.00)	598.33 (±10.41)
	VCD	186.67 (±5.77)	125.56 (±3.47)	128.89 (±6.73)	139.45 (±12.51)	580.56 (±15.13)
	M3ID	186.67 (±5.77)	128.33 (±10.41)	131.67 (±5.00)	151.67 (±20.88)	598.11 (±20.35)
	Woodpecker	187.50 (±2.89)	125.00 (±0.00)	126.66 (±2.89)	149.17 (±17.34)	588.33 (±10.00)
	HALC	183.33 (±0.00)	133.33 (±5.77)	107.92 (±3.69)	155.00 (±5.00)	579.58 (±9.07)
	Ours	**191.67** (±2.89)	**145.55** (±10.72)	**136.66** (±2.89)	**161.66** (±2.89)	**635.55** (±5.85)
InstructBLIP	Regular	160.42 (±5.16)	79.17 (±8.22)	**79.58** (±8.54)	130.42 (±17.34)	449.58 (±24.09)
	DoLa	175.00 (±5.00)	55.00 (±5.00)	48.89 (±3.47)	113.33 (±6.67)	392.22 (±7.88)
	OPERA	175.00 (±3.33)	61.11 (±3.47)	53.89 (±1.92)	120.55 (±2.55)	410.56 (±9.07)
	VCD	158.89 (±5.85)	**91.67** (±18.34)	66.11 (±9.76)	121.67 (±12.58)	438.33 (±16.07)
	M3ID	160.00 (±5.00)	87.22 (±22.63)	69.44 (±9.18)	125.00 (±7.64)	441.67 (±17.32)
	Ours	**180.00** (±5.00)	77.78 (±7.70)	74.44 (±12.05)	**135.55** (±3.85)	**467.77** (±8.55)
Qwen-VL	Regular	155.00 (±3.54)	127.67 (±13.36)	131.67 (±7.73)	173.00 (±9.75)	587.33 (±31.06)
	VCD	156.00 (±6.52)	131.00 (±6.19)	128.00 (±3.61)	**181.67** (±5.14)	596.67 (±11.61)
	M3ID	178.33 (±2.89)	143.33 (±2.89)	150.00 (±2.89)	175.00 (±5.00)	646.66 (±8.50)
	Ours	**180.00** (±5.00)	**146.67** (±5.00)	**156.11** (±6.31)	178.33 (±2.89)	**661.11** (±3.47)

Method	LR	AR	RR	FP-S	FP-C	CP	Overall
Regular	30.51	71.36	52.17	67.58	**58.74**	76.35	64.09
VCD	30.51	**73.37**	53.04	**67.92**	57.34	77.03	64.60
M3ID	30.51	72.36	53.04	67.58	57.34	**77.36**	64.43
Ours	**33.05**	**73.37**	**54.78**	66.55	**58.74**	**77.36**	**64.95**

Table C2. **Detailed results on MMBench benchmark**. Abbreviations adopted: LR for Logical Reasoning; AR for Attribute Reasoning; RR for Relation Reasoning; FP-S for Fine-grained Perception (Single Instance); FP-C for Fine-grained Perception (Cross Instance); CP for Coarse Perception. The best results are **bolded**.

Method	Rec	OCR	Know	Gen	Spat	Math	Total
Regular	30.8	19.0	14.5	17.9	26.9	**11.5**	26.1
VCD	35.6	21.9	18.3	21.9	28.9	3.8	30.9
M3ID	35.0	19.7	18.8	19.0	26.0	7.7	29.9
DoLA	37.2	22.1	17.9	21.0	26.3	7.7	31.7
OPERA	35.4	**25.6**	20.5	**22.9**	30.9	11.5	32.0
HALC	36.2	21.5	17.5	20.1	23.5	**7.7**	30.8
Ours	**37.3**	23.9	**22.9**	22.1	**31.3**	3.8	**32.8**

Table C3. **Detailed results on MM-Vet benchmark**. Abbreviations adopted: Rec for Recognition, OCR for Optical Character Recognition, Know for Knowledge, Gen for Language Generation, Spat for Spatial Awareness, Math for Mathematics. The best results are **bolded**, and the second best are underlined.

C. More Experimental Results and Analysis

C.1. Full Results on MME-Hallucination

In Table C1, we present the full results on the MME-Hallucination benchmark. From the results, our method consistently outperforms others on both object-level and attribute-level data across three LVLM backbones.

C.2. Full Results on MMBench

In Table C2, we present the overall performance on the MMBench benchmark, as well as the detailed performance across six Level-2 abilities: Logical Reasoning (LR), Attribute Reasoning (AR), Relation Reasoning (RR), Fine-grained Perception - Single Instance (FP-S), Fine-grained Perception - Cross Instance (FP-C), and Coarse Perception (CP). We follow VCD [16] to conduct experiments on the MMBench-`dev` set. Our method outperforms other baselines in most abilities and the overall score.

C.3. Results on MM-Vet

In Table C3, we present the overall performance on the MM-Vet [39] benchmark, where we use LLaVA-1.5 as the LVLM backbone. From the results, we observed that our method consistently outperforms others on the MM-Vet benchmark.

C.4. Evaluation on other advanced LVLMs

We further report results of LLaVA-NeXT-7B/13B [23] on POPE (MS-COCO) benchmark in table C4. Our method consistently outperforms existing approaches at both scales while requiring only half the inference time and resources.

Method	LLaVA-NeXT-7B				LLaVA-NeXT-13B			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Regular	85.71	85.27	86.33	85.80	86.74	86.53	87.04	86.78
VCD	87.07	87.40	86.62	87.01	87.09	87.39	86.69	87.04
M3ID	87.48	87.64	**87.27**	87.45	87.84	**87.95**	87.71	87.83
Ours	**87.96**	**88.59**	87.13	**87.86**	**87.94**	87.31	**88.80**	**88.05**

Table C4. **Detailed results with LLaVA-NeXT**. The best results are **bolded**, and the second best are underlined.

D. More Ablation Studies and Analysis

D.1. Effects of α_1 and α_2 in Adaptive Decoding

In Section 3, we introduce collaborative and contrastive decoding, along with hyperparameters α_1 and α_2, which regulate the influence of the textual-enhanced branch. Tables D5 and D6 analyze their impact, showing that the default values $\alpha_1 = 3$ and $\alpha_2 = 1$ yield the best performance across benchmarks. Notably, setting these to 0 reduces our approach to standard decoding, confirming that adaptive decoding significantly enhances hallucination mitigation in LVLMs.

D.2. Effect of β in Adaptive Plausibility Constraint

We perform an ablation study on β, introduced in Eq. 23, by varying its value from 0 to 0.5 while keeping all other hyperparameters fixed. As shown in Table D7, setting $\beta = 0$, which removes the constraint, leads to suboptimal performance across both benchmarks. Our method achieves the best results with $\beta = 0.1$, which we adopt as the default setting.

D.3. Effect of γ in Adaptive Plausibility Constraint

We further studied the influence led by the threshold γ for adaptive decoding. The results in Table D8 show that setting $\gamma = 0.2$ reaches the optimal result for LLaVA-1.5. Besides, we keep $\gamma = 0.4$ for other baseline LVLMs.

D.4. Scaling Up the LVLMs

We extend our evaluation to the 13B variant of the LLaVA-1.5 model to assess the scalability of our approach. Table D9 compares our results with state-of-the-art methods across all three subsets of the POPE benchmark using the 13B model. Our findings show that increasing model size does not mitigate hallucination issues, as the 7B and 13B models exhibit comparable performance. Notably, ONLY consistently outperforms other approaches across all subsets, demonstrating its effectiveness and scalability.

Values	POPE				CHAIR	
	Acc.	Prec.	Rec.	F1	$CHAIR_S$	$CHAIR_I$
$\alpha_1 = 0$	88.13	**94.55**	80.93	87.21	23.5	8.6
$\alpha_1 = 1$	88.27	94.50	81.27	87.38	22.4	7.8
$\alpha_1 = 2$	88.87	89.63	88.10	88.86	21.5	7.2
$\alpha_1 = 3$	**89.70**	89.95	**88.27**	**89.10**	**20.0**	**6.2**
$\alpha_1 = 4$	88.37	88.85	87.94	88.39	22.3	7.6

Table D5. **Sensitivity analysis of hyperparameter** α_1. We present the performance of our approach, based on the LLaVA-1.5 backbone, across two benchmarks for varying values of α_1. Note that we fix $\alpha_2 = 1$ in this experiment.

Values	POPE				CHAIR	
	Acc.	Prec.	Rec.	F1	$CHAIR_S$	$CHAIR_I$
$\alpha_2 = 0$	86.50	86.35	88.13	86.72	24.8	9.3
$\alpha_2 = 1$	**89.70**	89.95	**88.27**	**89.10**	**20.0**	**6.2**
$\alpha_2 = 2$	87.67	96.69	78.00	86.35	22.4	7.6
$\alpha_2 = 3$	87.37	**97.14**	77.00	85.91	23.4	7.3
$\alpha_2 = 4$	87.13	97.12	76.53	85.61	24.2	8.1

Table D6. **Sensitivity analysis of hyperparameter** α_2. We present the performance of our approach, based on the LLaVA-1.5 backbone, across two benchmarks for varying values of α_1. Note that we fix $\alpha_1 = 3$ in this experiment.

Values	POPE				CHAIR	
	Acc.	Prec.	Rec.	F1	$CHAIR_S$	$CHAIR_I$
$\beta = 0$	87.70	93.40	81.13	86.84	24.6	10.1
$\beta = 0.05$	88.17	**94.21**	81.33	87.30	23.7	9.6
$\beta = 0.1$	**89.70**	89.95	**88.27**	**89.10**	**20.0**	**6.2**
$\beta = 0.25$	89.56	89.48	87.63	88.55	21.4	7.6
$\beta = 0.5$	89.47	89.83	86.53	88.15	22.1	7.2

Table D7. **Sensitivity analysis of hyperparameter** β. We present the performance of our approach, based on the LLaVA-1.5 backbone, across two benchmarks for varying values of β.

Values	POPE				CHAIR	
	Acc.	Prec.	Rec.	F1	$CHAIR_S$	$CHAIR_I$
$\gamma = 0.0$	89.13	90.41	86.38	88.35	23.5	8.2
$\gamma = 0.1$	89.20	89.88	86.73	88.28	22.6	8.1
$\gamma = 0.2$	**89.70**	89.95	**88.27**	**89.10**	**20.0**	**6.2**
$\gamma = 0.3$	89.40	93.20	85.00	88.91	21.2	7.1
$\gamma = 0.4$	89.03	93.99	83.40	88.38	21.7	7.0
$\gamma = 0.5$	89.15	92.26	84.29	88.10	22.4	7.6
$\gamma = 0.6$	89.21	91.78	85.39	88.47	23.1	8.1

Table D8. **Sensitivity analysis of hyperparameter** γ. We present the performance of our approach, based on the LLaVA-1.5 backbone, across two benchmarks for varying values of γ.

D.5. Details about Ablation Studies on Layer Selection and Strategies

In Section 4.4, we conduct two ablation studies to validate our proposed method. Detailed results are provided

Table D9. **Results on POPE [19] benchmark using 13B-sized LLaVA-1.5**. Higher (↑) accuracy, precision, recall, and F1 indicate better performance.

	Setup	Method	LLaVA-1.5			
			Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑
MS-COCO	Random	Regular	82.53	78.57	89.47	83.67
		VCD	84.80	80.67	91.53	85.76
		M3ID	85.37	81.30	91.87	86.26
		Ours	**88.63**	**89.66**	87.33	**88.48**
	Popular	Regular	80.53	76.17	88.87	82.03
		VCD	82.23	76.88	92.20	83.84
		M3ID	82.60	77.91	91.00	83.95
		Ours	**85.47**	**83.25**	88.80	**85.94**
	Adversarial	Regular	75.80	70.41	89.00	78.62
		VCD	77.33	71.44	91.07	80.07
		M3ID	77.43	71.65	90.80	80.09
		Ours	**80.63**	**76.33**	88.80	**82.10**

below.

Selection of Layer for Textual Enhancement: In this experiment, we select a single layer from the total of 32 layers for textual enhancement. The F1 scores for our method across the 32 layers are as follows: [85.37, 85.20, 84.74, 84.7, 85.11, 84.68, 85.17, 84.69, 85.18, 84.73, 84.7, 84.74, 84.9, 84.94, 84.88, 85.06, 84.62, 84.67, 84.83, 85.15, 84.72, 84.76, 84.99, 85.03, 84.7, 84.93, 84.76, 84.9, 84.8, 85.12, 84.62, 84.76]. In comparison, the results for regular decoding, VCD [16], and M3ID [10] are 81.27, 83.38, and 84.05, respectively.

Other Strategies for Textual Enhancement: In Table 6, we explore additional strategies for textual enhancement, which include:

- $a^{\mathcal{V}}_{\ell,i} \leftarrow 0$: Setting the visual attention in the attention matrix to zero, inspired by M3ID [10], which uses a visual-free input for contrastive decoding;
- $a^{\mathcal{V}}_{\ell,i} \leftarrow a^{\mathcal{V}}_{\ell,i} + \varepsilon$: Adding noise ε to the visual attention, inspired by VCD [16], which uses a distorted visual input for contrastive decoding;
- $a^{\mathcal{T}}_{\ell,i} \leftarrow a^{\mathcal{T}}_{\ell,i} * 2$: Enhancing textual attention by directly multiplying it by 2;
- $\text{Ratio} \leftarrow \sum a_{\mathcal{T}} / \sum a_{\mathcal{V}}$: Instead of using the text-to-visual entropy ratio as the criterion to select textual-enhanced heads, we use the ratio between the sum of textual attention and visual attention. Heads with a ratio lower than the average across all heads are masked out, as described in Eq. 12.

All of these strategies require minimal additional computation, providing an efficiency advantage over other methods [10, 16]. This demonstrates the effectiveness of using just one layer for mitigating hallucinations in LVLMs, rather than relying on an extra full-process inference.

E. More Case Studies

E.1. Details about GPT-4V-Aided Evaluation

Following VCD [16], we use GPT-4V to evaluate responses in open-ended generation scenarios, scoring them based on accuracy and detailedness. Leveraging GPT-4V's strong human-like capabilities, it can detect incorrect colors, positions, and relationships, allowing for a thorough evaluation of the responses.

Specifically, we apply the prompt in Table E10 to instruct GPT-4V to rate two responses on a scale from 1 to 10 for both accuracy and detailedness:

- **Accuracy** measures the consistency between the responses/descriptions generated by the LVLMs and the given image. A lower score is given if GPT-4V detects any inconsistencies in the content.
- **Detailedness** evaluates the depth and specificity of the responses. A higher score is awarded if the response includes comprehensive descriptions, captures fine-grained details of the image, and provides well-elaborated explanations. Conversely, a lower score is given if the response is vague or lacks sufficient detail.

E.2. More Qualitative Results

In Figure E1, we present additional case studies from LLaVA-Bench to qualitatively demonstrate the effectiveness of our methods in mitigating hallucinations. We also include GPT-4V evaluations, providing accuracy and detailedness scores for each instance.

F. Future Work

In future work, we aim to further improve the speed of our method and develop a more efficient hallucination mitigation approach that surpasses the original LVLM speed, leveraging efficient LVLM techniques like FastV [4] and VScan [42]. Additionally, we plan to explore our method's potential for video hallucination mitigation to demonstrate its adaptability across various tasks.

Description:
AI that scores image description accuracy and detailedness.

Instructions:
You are an AI designed to evaluate and score the performance of two AI assistants in describing a given image. Your primary focus is on the accuracy and detailedness of their descriptions. You will assess the accuracy by checking for hallucinations - any part of the description that is inconsistent with the image content. For detailedness, you will consider how rich the response is in necessary details, excluding any hallucinated parts. You will provide scores on a scale from 1 to 10 for each assistant separately, based on these criteria. After scoring, you will offer an explanation for your evaluation, ensuring it is free from bias and not influenced by the order of presentation of the responses.

Input format:

[Assistant 1]
{Response 1}
[End of Assistant 1]

[Assistant 2]
{Response 2}
[End of Assistant 2]

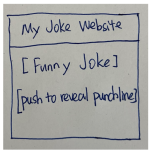
Output format:

Accuracy:
Scores of the two answers:
Reason:

Detailedness:
Scores of the two answers:
Reason:

Table E10. **GPT-4V-aided evaluation setup**. We present the prompt we provided to GPT-4V to evaluate the LVLM responses based on accuracy and detailedness.

Visual Input:

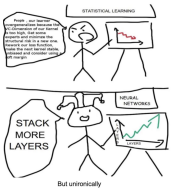


Query:
Describe this photo in detail.

[Regular Decoding] GPT-4V-Aided Evaluation: Accuracy: 2 Detailedness: 3
The image features a myJokeWebsite box placed on a piece of graph paper. This graph paper is filled with multiple sentences, likely specifically written jokes. In addition to the main box, there are smaller boxes placed around the main page, primarily in the lower half of the paper. These smaller boxes contain various jokes, giving the impression of a collection of joke examples to choose from for each joke box.

[Our Approach] GPT-4V-Aided Evaluation: Accuracy: 4 Detailedness: 5
The image is a blue drawing of a joke website sign that says "My Joke Website: Funny Joke." Below the sign, there is a box labeled "Push to Reveal Punchline". The drawing appears to be done in blue ink on a piece of paper, representing the user's intention to access funny jokes on the website.

Visual Input:



Query:
Describe this photo in detail.

[Regular Decoding] GPT-4V-Aided Evaluation: Accuracy: 5 Detailedness: 4
The meme features a cartoon character representing a person, pointing at a graph or animating figures on a large chart, possibly a white board, while surrounded by a dose of humor. The character appears to be reacting to an equation, discussing or teaching the presence. Below the main figure, four additional figures are portrayed, each interacting or holding their hands up in response to the main character's commentary about theory graphs. On the side, there is a description providing additional context.

[Our Approach] GPT-4V-Aided Evaluation: Accuracy: 7 Detailedness: 6
The image consists of two comics, one above the other, each featuring a person in a similar setting. In both comics, the person is standing in front of a chart or graph, which is likely discussing layers. The main focus of the comics is to emphasize the importance of layers and their applications in various fields. In each comic, there is a character standing in front of the chart, while another character is drawing attention to the chart. The drawings are simple, yet they convey the message effectively.

Visual Input:

Query:
Describe this photo in detail.

[Regular Decoding] GPT-4V-Aided Evaluation: Accuracy: 4 Detailedness: 5
The image showcases a large painting of a monkey dog with a sword through its throat, surrounded by a Renaissance-style painting. In the scene, a woman resembling the famous Mona Lisa has a dog wearing a hood, and instead of her looking out of the canvas, it appears as if the dog is staring back. The dog is the focal point of the scene, adding a mysterious twist to a traditional painting.

[Our Approach] GPT-4V-Aided Evaluation: Accuracy: 8 Detailedness: 9
The painting depicts a unique scene where a dog is portrayed as a woman, possibly Mona Lisa, wearing a hood. The dog is adorned with a Renaissance-style dress and a hat, giving it an intriguing resemblance to a human. The background of the painting features mountains and water, further enhancing the artistic portrayal. A rock formation can also be seen in the painting. The overall composition creates a captivating and amusing artistic representation of the dog.

Figure E1. **Case studies on the LLaVA-Bench benchmark**. We compare the responses generated by regular decoding and our method using LLaVA-1.5. GPT-4V-aided evaluation results are also provided alongside the responses. Hallucinated and accurate content is highlighted in red and blue.