

6. Implementation Details

Table 6 summarizes the general and model-specific parameters used in our experiments. We adopt a deterministic decoding setup with fixed temperature and beam size, and vary the maximum number of generated tokens depending on the evaluation benchmark.

To simulate visual hallucinations, we use Stable Diffusion with half of total diffusion steps (T) for forward and reverse steps ($0.5T$). During inference, hidden representations from selected layers are projected onto the orthogonal complement of the hallucination subspace. The rank of this subspace is tuned for each model to optimize hallucination suppression without compromising generation quality.

Parameters	Values
General Settings	
Temperature	0
Number of Beams	3
Maximum New Tokens (CHAIR)	64
Maximum New Tokens (OPOPE)	256
Maximum New Tokens (LLaVA-Bench)	1024
Number of Diffusion Steps	$0.5T$
Editing Layers	16-32
Model-Specific Parameters	
Hallucination Space Rank (LLaVA-1.5)	8
Hallucination Space Rank (MiniGPT-4)	64
Hallucination Space Rank (mPLUG-Owl2)	32

Table 6. General and model-specific parameter settings.

We report the performance of baseline models based on the results presented in the Nullu paper [35], which reproduces several state-of-the-art object hallucination mitigation methods, including HALC [9], VCD [17], DoLa [12], and OPERA [14]. Each method is evaluated using its official implementation and publicly released pre-trained checkpoints. Moreover, we carefully adopted key hyperparameter settings that align with those reported in their original implementations. Specifically, for HALC, we used an amplification factor of 0.05 along with a beam size of 1, reflecting the minimal decoding augmentation strategy employed by the authors. For VCD, we set the diffusion noise step to 500, consistent with the level of perturbation used in their hallucination mitigation framework. In the case of DoLa, we applied a repetition penalty of 1.2 and performed interventions across the early transformer layers, as suggested to promote faithful generation during early stages of decoding. Lastly, for OPERA, we implemented a scaled self-attention mechanism with a scaling factor of 50, in line with the original design to regulate attention intensity. All these hyperparameter choices were selected to faithfully reproduce the respective baselines under settings that remain true to the original authors’ reported configurations.

7. Detailed Results for OPOPE

To evaluate the effectiveness of hallucination mitigation methods in a way that is compatible with post-hoc approaches, we adopt the Offline POPE (OPOPE) protocol. OPOPE is a non-interactive version of the original POPE metric, introduced in HALC [9], and is specifically designed to avoid the limitations of online interaction during inference. While POPE evaluates hallucination through interactive polling prompts such as “Is there a/an [object] in the image?”, this interactive requirement poses challenges for post-hoc mitigation methods and can lead to instability in models with weaker generative backbones.

OPOPE performs hallucination evaluation entirely offline. Following the caption generation pipeline of CHAIR [29], captions are generated for a fixed image set, then tokenized and singularized, with words mapped to MSCOCO object categories using synonym and compound-word mappings. Using POPE’s sampling strategy, three test object lists are created per setting—*Random*, *Popular*, and *Adversarial*—each containing six objects balanced between ground-truth and nonexistent categories. Captions are scanned for these objects, and mentions of nonexistent ones are counted as hallucinations.

To account for the limited recall sensitivity of offline evaluations—where false negatives may occur if hallucinated content is not explicitly verbalized—we adopt the F_β score as our primary evaluation metric, following HALC. The F_β score reduces the penalty on false negatives by weighting precision more heavily. Specifically, we use $\beta = 0.2$ for F_β score:

$$F_\beta = \frac{(1 + \beta^2) \cdot (\text{precision} \cdot \text{recall})}{\beta^2 \cdot \text{precision} + \text{recall}} \quad \text{with } \beta = 0.2.$$

The main paper reports results averaged across the three sampling strategies for brevity, while Table 7 provides a full breakdown by model and sampling configuration. We report Accuracy, Precision, Recall, and F-score for all baselines—Greedy, Beam Search, DoLa [12], OPERA [14], VCD [17], HALC [9], and NullU [35]—along with our method, VISTA. VISTA consistently ranks among the top-performing methods across most settings, demonstrating robustness to varying object distributions and model architectures. These results complement the main findings and provide a detailed view of performance under each hallucination sampling strategy.

We additionally conducted statistical significance tests comparing VISTA to baseline methods for each model. Across all cases, results consistently showed significant improvements. For instance, on LLaVA, VISTA outperforms the baseline with p-values well below 0.01 across all metrics: Accuracy ($p = 0.0033$), Precision ($p < 0.0001$), and $F_{0.2}$ Score ($p < 0.0001$), confirming the robustness of our method.

Setting	Model	Method	Accuracy	Precision	Recall	F Score
Random	LLaVA-1.5	Greedy	81.52	98.41	64.07	96.42
		DoLa	81.38	98.11	64.00	96.14
		OPERA	81.62	98.57	64.17	96.58
		VCD	80.57	98.41	62.13	96.25
		HALC	79.58	98.21	60.27	95.89
		Nullu	81.18	98.05	63.63	96.05
		VISTA	81.84	98.70	64.58	96.68
	MiniGPT-4	Greedy	72.42	98.49	45.53	94.25
		DoLa	72.45	98.58	45.57	94.34
		OPERA	72.57	98.78	45.70	94.52
		VCD	72.35	98.19	45.53	93.97
		HALC	72.08	98.62	44.80	94.25
		Nullu	72.68	99.06	45.80	94.82
		VISTA	72.85	99.37	46.00	95.12
	mPLUG-Owl2	Greedy	79.45	97.74	60.30	95.46
		DoLa	78.33	97.60	58.10	95.09
		OPERA	78.31	97.73	57.96	95.21
		VCD	78.19	98.23	57.42	95.61
		HALC	77.83	97.72	57.00	95.10
		Nullu	79.05	98.40	60.50	95.82
		VISTA	79.56	97.66	60.57	95.41
Popular	LLaVA-1.5	Greedy	78.93	91.17	64.07	89.71
		DoLa	78.72	91.60	64.00	89.26
		OPERA	79.22	91.80	64.17	90.30
		VCD	77.57	89.87	62.13	88.35
		HALC	77.47	91.87	60.27	90.05
		Nullu	79.80	94.06	63.63	92.36
		VISTA	80.02	93.47	64.58	91.89
	MiniGPT-4	Greedy	70.80	92.01	45.53	88.53
		DoLa	70.90	92.33	45.57	88.22
		OPERA	71.10	92.82	45.70	89.27
		VCD	70.33	90.30	45.53	86.98
		HALC	70.92	93.80	44.80	90.00
		Nullu	71.97	96.08	45.80	92.19
		VISTA	72.15	96.43	46.00	92.53
	mPLUG-Owl2	Greedy	76.00	87.90	60.30	86.38
		DoLa	75.20	88.36	58.10	86.00
		OPERA	75.02	88.06	57.96	86.33
		VCD	74.86	88.16	57.42	86.37
		HALC	75.77	91.34	57.00	89.26
		Nullu	77.09	92.34	61.60	90.49
		VISTA	77.95	92.91	60.57	91.03
Adversarial	LLaVA-1.5	Greedy	76.97	86.36	64.07	85.22
		DoLa	76.85	86.18	64.00	85.05
		OPERA	77.03	86.40	64.17	85.26
		VCD	75.88	85.71	62.13	84.48
		HALC	76.57	89.44	60.27	87.80
		Nullu	77.58	88.27	63.63	86.92
		VISTA	78.29	89.03	64.58	87.84
	MiniGPT-4	Greedy	70.43	90.65	45.53	87.32
		DoLa	70.50	88.55	45.57	87.50
		OPERA	70.78	91.63	45.70	88.21
		VCD	69.82	83.41	45.53	85.32
		HALC	70.52	92.22	44.80	88.60
		Nullu	71.10	92.73	45.80	89.21
		VISTA	71.44	93.70	46.00	90.11
	mPLUG-Owl2	Greedy	74.23	83.58	60.30	82.36
		DoLa	73.52	83.98	58.10	82.55
		OPERA	73.17	83.45	57.96	82.06
		VCD	72.85	83.01	57.42	81.61
		HALC	74.02	86.41	57.00	84.72
		Nullu	75.15	87.76	61.60	86.10
		VISTA	76.10	87.90	60.57	86.40

Table 7. Performance comparison of various hallucination mitigation methods across different models and evaluation settings.

Examples of Hallucinated Images



Figure 10. Visual examples from the MSCOCO dataset illustrating hallucinations generated using $0.5T$ Stable Diffusion steps.

This figure shows MSCOCO images before and after applying visual hallucination using Stable Diffusion with $t_h = 0.5T$ denoising steps. While the overall scene struc-

ture remains largely unchanged, the hallucinated images introduce subtle object or texture modifications intended to mislead vision-language models.

8. Extended Ablation Analysis

In this section of the appendix, we provide extended results from our ablation analysis to support the findings presented in the main paper.

8.1. Best Subspace Rank

As shown in the ablation section of the main paper, the best subspace rank for LLaVA-7B was $r = 8$, achieving the lowest CHAIR_S and CHAIR_I scores while preserving or improving BLEU. Here, we extend this analysis to MiniGPT-4 and mPLUG-Owl2.

Figure 11 shows that for MiniGPT-4, the optimal number of SVD components is $r = 64$, where CHAIR scores drop significantly and BLEU remains stable. Similarly, Figure 12 demonstrates that mPLUG-Owl2 achieves the best balance at $r = 32$, where hallucination metrics are minimized and BLEU degradation is minimal. These findings highlight that the optimal subspace rank can vary across models and should be tuned accordingly.

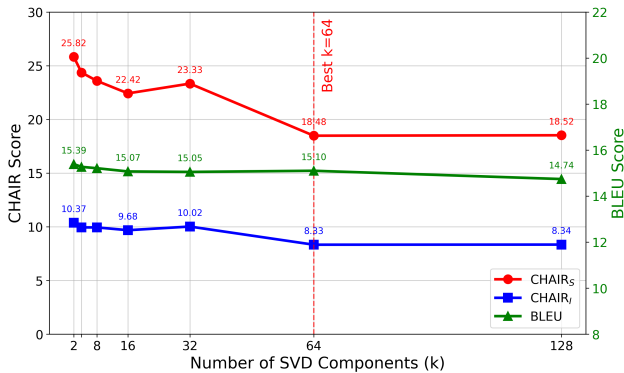


Figure 11. Effect of subspace rank r on CHAIR_S, CHAIR_I, and BLEU for MiniGPT-4 on the CHAIR benchmark. Using $r = 64$ yields the best performance, minimizing hallucinations with minimum BLEU degradation.

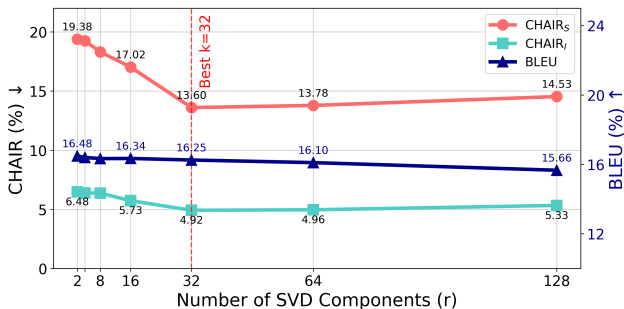


Figure 12. Effect of subspace rank r on CHAIR_S, CHAIR_I, and BLEU for mPLUG-Owl2 on the CHAIR benchmark. Using $r = 32$ yields the best performance, minimizing hallucinations with minimum BLEU degradation.

8.2. Selection of Layers for Projection

VISTA mitigates hallucinations at inference time by projecting the hidden representations of selected LVLM layers onto their corresponding null spaces. To assess the influence of layer choice on suppression effectiveness, we perform an ablation study using various candidate layer ranges. We evaluate each configuration on the MSCOCO validation split using the CHAIR metric, reporting both instance-level (CHAIR_I) and sentence-level (CHAIR_S) hallucination rates, where lower values indicate better suppression.

Layer Range	CHAIR _S ↓	CHAIR _I ↓
8–16	13.52	4.82
8–24	14.10	5.02
16–24	18.10	5.86
16–32	13.05	4.53
20–32	14.10	4.84
24–32	14.00	4.64
28–32	14.48	5.10

Table 8. Ablation study on the transformer layer range used for projection in VISTA.

As shown in Table 8, the choice of layer range significantly impacts hallucination suppression. Projection over earlier layers, such as 8–16 or narrower bands like 28–32, yields suboptimal results, suggesting that local or overly shallow representations lack the semantic expressiveness needed to isolate hallucination directions. Performance improves with broader and deeper projections, with the 16–32 range consistently achieving the lowest CHAIR_S and CHAIR_I scores. This indicates that hallucination-inducing signals are best captured in semantically rich, higher layers spanning a sufficiently deep interval. Consequently, we adopt the 16–32 layer range as the default configuration in our main experiments.

9. Details of GPT-4V Aided Evaluation on LLaVA-Bench

We evaluate our hallucination suppression method using GPT-4V on the LLaVA-Bench benchmark [25], which consists of 24 images. For each example, GPT-4V receives the image, the evaluation prompt, and the responses from both the original and the edited (hallucination-suppressed) models.

GPT-4V compares the two responses in terms of accuracy—how well the answer reflects the image—and detailedness—how specific and complete the response is. It outputs a score for each criterion along with a brief justification. Figure 13 and Figure 14 show the prompt format and a sample GPT-4V evaluation, respectively. This setup enables consistent comparison of response quality grounded in the image content.

Prompt for GPT-4V Aided Evaluation

Descriptions:

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:

[Assistant1]
{Response1}
[End of Assistant1]

[Assistant2]
{Response2}
[End of Assistant2]

Output format:

Accuracy:

- Assistant1: [score]/10
- Assistant2: [score]/10
- Reason: [explanation for accuracy scores]

Detailedness:

- Assistant1: [score]/10
- Assistant2: [score]/10
- Reason: [explanation for detailedness scores]

Figure 13. An illustration of the prompt used to guide GPT-4V for visual question evaluation.

Input:



[Assistant1]

The image features two large subway sandwiches, each cut in half, placed on a black background. The sandwiches are filled with various ingredients, including lettuce, tomatoes, and cheese. The sandwiches are positioned close to each other, creating a visually appealing presentation. The close-up view of the sandwiches highlights their delicious and appetizing appearance.

[End of Assistant1]

[Assistant2]

The image features a close-up view of a Subway Series sandwich, showcasing its delicious ingredients. The sandwich is filled with a generous amount of meat, lettuce, and cheese, making it a mouth-watering meal. The Subway Series logo is also visible in the background, emphasizing the brand's commitment to quality and taste. The sandwich is presented on a black background, allowing the viewer to focus solely on the appetizing details of the meal.

[End of Assistant2]

GPT-4V Output:

Accuracy:

- Assistant1: 8/10
- Assistant2: 9/10

- Reason: Assistant1 accurately describes the presence of sandwiches and their ingredients but incorrectly mentions that the sandwiches are cut in half, which is not visible in the image. Assistant2 accurately describes the sandwich and includes the presence of the Subway Series logo, which is visible in the image, making it slightly more accurate.

Detailedness:

- Assistant1: 7/10
- Assistant2: 8/10

- Reason: Assistant1 provides a good level of detail about the ingredients and presentation but lacks mention of the Subway Series logo, which is a significant detail. Assistant2 includes details about the ingredients and the logo, adding to the richness of the description. However, it could have included more specifics about the ingredients visible, such as the presence of onions.

Figure 14. Example of GPT-4V-based evaluation for image description.

More Examples of VISTA on LLaVA-Bench



Prompt: please describe this image in detail.

Original Model: The image features a small brown dog lying on a wooden floor, possibly a porch or a deck. The dog appears to be sleeping or resting, taking up a significant portion of the floor. The scene is set on a blue surface, which adds a contrasting color to the image. In the background, there is a chair placed near the left edge of the image, and a dining table can be seen in the upper left corner. The overall atmosphere of the scene is relaxed and peaceful, with the dog enjoying its time on the floor.

VISTA: The image features a small, cute French Bulldog laying down on a blue hardwood floor. The dog appears to be sleeping or resting comfortably, occupying a significant portion of the floor's space. In the background, there is a chair positioned on the left side of the scene. The overall atmosphere of the scene is cozy and inviting.



Prompt: please describe this image in detail.

Original Model: The image features a young boy wearing glasses and a bow tie, standing in front of a building. He appears to be smiling and posing for the camera. The boy is wearing a blue suit and a red bow tie, giving him a formal and stylish appearance. The scene is set in a city environment, with a backdrop of buildings and a street visible in the background.

VISTA: The image features a young boy wearing a red bow tie and a blue jacket. He appears to be the main focus of the scene. The boy is also wearing eyeglasses, which adds a distinctive touch to his outfit. In the background, there are a few other people present, but they do not seem to be the main subject of the image. The boy's outfit creates a lively atmosphere.



Prompt: please describe this image in detail.

Original Model: The image features a wooden table with a basket of fruit, including a variety of fruits such as apples, oranges, and bananas. There are three apples on the table, with one placed near the left side, another in the middle, and the third one on the right side. In addition to the apples, there are three oranges on the table. One orange is located on the left side, another one is in the middle, and the third one is on the right side. A banana can also be seen on the table, positioned towards the right side. The table appears to be a dining table, and the fruits are arranged neatly on it.

VISTA: The image features a close-up view of a group of four unripe, dark-colored fruits. The fruits are placed on a dining table, with one of them being sliced open to reveal the inside. The sliced fruit is located in the center of the image, while the other three fruits are positioned around it. In addition to the fruits, there is a basket present in the scene, which is located at the top-left corner of the image. The presence of the basket suggests that the fruits may have been recently harvested or are being displayed for sale or consumption.

Figure 15. More examples of VISTA applied to LLaVA-Bench.