Breaking the Illusion: When Positive Meets Negative in Multimodal Decoding

Supplementary Material

1. More Detailed Experimental Settings

This section provides further details on our experimental setup to ensure full reproducibility. We detail the specific model checkpoints, hyperparameter values, and the computing environment used in our work.

Hyperparameter Details. The final hyperparameter values for our PND framework, used to generate the main results reported in this paper, are listed in Table 1. These values were empirically determined by optimizing the F1-score on a validation split of the POPE dataset. We observed that the framework is robust to minor variations around these optimal values, as discussed in our sensitivity analysis.

Table 1. Hyperparameter settings for the PND framework.

Hyperparameter	Value
Diffusion Timestep	500
α	2.2
γ	0.4
β	0.5
temperature	1.0
top_p	1.0
$chunk\_idx$	0
conv	$llava_{v1}$

2. More Results on the POPE Benchmark

This section provides a detailed breakdown of our PND framework's performance on the POPE benchmark, supplementing the summarized results in the main paper. We demonstrate the consistent and significant effectiveness of PND across various model architectures and scales, confirming its broad applicability.

2.1. Performance Across Diverse MLLM Architectures

To robustly validate the model-agnostic nature of our PND framework, it is essential to evaluate its performance on a set of MLLMs with distinct underlying architectures for vision-language fusion. We selected three leading open-source models for this purpose: **LLaVA-1.5-7B**, which employs a simple yet effective MLP for projection; **InstructBLIP-7B**, which utilizes a more sophisticated Q-Former to bridge modalities; and **Qwen-VL**, which represents another powerful and distinct state-of-the-art architecture. Demonstrating consistent efficacy across these models is key to proving that PND is a truly generalizable decoding strategy, not one tailored to a specific model design.

The detailed results are presented in Table 2 3 4. A clear and consistent trend emerges: PND significantly outperforms the standard decoding baseline across every model. For instance, on the LLaVA-1.5 model, PND yields a dramatic improvement, showcasing its ability to substantially enhance the visual grounding of simpler architectures. Even on powerful models like InstructBLIP and Qwen-VL, which have stronger baselines, our method provides considerable gains. This confirms that even state-of-the-art models are not immune to hallucination and can significantly benefit from our contrastive control mechanism.

Furthermore, the improvements are not confined to a single test condition but are consistently observed across all three POPE subsets: *random*, *popular*, and *adversarial*. The success on the more challenging *popular* and *adversarial* subsets is particularly noteworthy, as it indicates that PND is capable of overriding strong, ingrained linguistic biases within the models, rather than merely correcting simple, random errors. This comprehensive performance across diverse architectures and challenging test conditions strongly validates PND as a versatile and effective solution for object hallucination.

2.2. Impact of Model Scale

A crucial investigation for any enhancement technique is its efficacy and relevance in the era of rapid model scaling. It is often hypothesized that simply increasing a model's parameter count will resolve its inherent flaws. To test this assumption against the persistent problem of hallucination, we conduct a comparative analysis of PND's performance on both 7B and 13B variants of LLaVA-1.5 and InstructBLIP. Our goal is to determine if PND provides meaningful benefits for larger, more capable models.

The results, presented in Table 5 7, offer a nuanced perspective. As expected, the 13B models consistently exhibit stronger baseline performance than their 7B counterparts, confirming the general benefits of scale. However, they are far from immune to hallucination, still demonstrating a significant vulnerability to generating ungrounded content. Crucially, PND delivers substantial and consistent performance gains across *both* 7B and 13B scales for both architectures. In many cases, the PND-enhanced 7B models reach or even exceed the performance of the 13B baseline models, showcasing the profound impact of our targeted decoding strategy.

Furthermore, the application of PND to the 13B mod-

Table 2. Performance comparison of LLAVA1.5-7B across POPE datasets.

Model	Dataset	Category	Methods	Accuracy	Precision	Recall	F1
LLAVA1.5-7B	MSCOCO	adversarial	regular	78.53	82.77	72.06	77.04
			PND	84.03 ↑5.50	89.85 ↑7.08	76.73 ↑4.67	83.48 ↑6.44
		popular	regular	81.56	88.97	72.08	84.87
			PND	86.10 ↑4.54	94.34 ↑5.37	76.80 ↑4.72	88.79 ↑3.92
		random	regular	83.00	92.23	72.06	89.12
			PND	87.33 ↑4.33	97.29 ↑5.06	76.80 ↑4.74	93.41 ↑4.29
	AOKVQA	adversarial	regular	73.36	71.48	77.73	74.48
			PND	78.93 ↑5.57	75.80 ↑4.32	85.00 ↑7.27	80.13 ↑5.65
		popular	regular	82.53	85.82	77.93	81.69
			PND	89.70 ↑7.17	93.49 ↑7.67	85.33 ↑7.40	89.22 ↑7.53
		random	regular	79.30	80.12	77.93	79.01
			PND	86.00 ↑6.70	86.48 ↑6.36	85.33 ↑7.40	85.90 ↑6.89
	GQA	adversarial	regular	75.26	73.17	73.16	76.33
			PND	81.03 ↑5.77	78.68 ↑5.51	85.13 ↑11.97	81.78 ↑5.45
		popular	regular	78.03	77.08	79.80	78.41
			PND	83.96 ↑5.93	83.02 ↑5.94	85.40 ↑5.60	84.19 ↑5.78
		random	regular	84.16	87.43	79.80	83.44
			PND	89.36 ↑5.20	92.75 ↑5.32	85.40 ↑5.60	88.92 ↑5.48

Table 3. Performance comparison of InstructBlip-7B across POPE datasets.

Model	Dataset	Category	Methods	Accuracy	Precision	Recall	F1
InstructBlip-7B	MSCOCO	adversarial	regular	75.66	74.09	78.93	76.43
			PND	82.20 ↑6.54	82.99 ↑8.90	81.00 ↑2.07	81.98 ↑5.55
		popular	regular	78.00	77.30	79.26	78.27
			PND	84.83 ↑6.83	87.83 ↑10.53	80.86 ↑1.60	84.20 ↑5.93
		random	regular	81.10	82.41	79.06	80.70
			PND	87.63 ↑6.53	93.52 ↑11.11	80.86 ↑1.80	86.73 ↑6.03
	AOKVQA	adversarial	regular	70.23	65.58	85.13	74.09
			PND	74.40 ↑4.17	68.98 ↑3.40	88.66 ↑3.53	77.59 ↑3.50
		popular	regular	77.36	73.44	85.73	79.11
			PND	82.40 ↑5.04	78.55 ↑5.11	89.13 ↑3.40	83.51 ↑4.40
		random	regular	80.20	77.15	85.80	81.24
			PND	88.23 ↑8.03	87.55 ↑10.40	89.13 ↑3.33	88.33 ↑7.09
	GQA	adversarial	regular	71.66	67.12	84.93	74.98
			PND	74.86 ↑3.20	69.77 ↑2.65	87.73 ↑2.80	77.73 ↑2.75
		popular	regular	74.33	69.85	85.60	69.32
			PND	78.93 ↑4.60	74.77 ↑4.92	87.33 ↑1.73	80.56 ↑11.24
		random	regular	79.96	76.93	85.60	81.03
			PND	86.26 ↑6.30	85.50 ↑8.57	87.33 ↑1.73	86.41 ↑5.38

els elevates their performance to new state-of-the-art levels. This strongly indicates that PND is not a remedial tool useful only for smaller models, but serves as a vital, complementary approach to scaling. While scaling may enhance a model's general knowledge and fluency, PND specifically reinforces the visual grounding mechanism that remains a structural weakness regardless of model size. Therefore, our method proves to be a valuable and scalable solution for enhancing the fidelity of even the largest and most powerful MLLMs.

Table 4. Performance comparison of QwenVL across POPE datasets.

Model	Dataset	Category	Methods	Accuracy	Precision	Recall	F1
QwenVL	MSCOCO	adversarial	regular	80.96	91.06	68.66	78.29
			PND	82.46 ↑1.50	93.55 ↑2.49	69.73 ↑1.07	79.90 ↑1.61
		popular	regular	82.47	95.00	68.53	79.62
			PND	84.06 ↑1.59	97.49 ↑2.49	69.93 ↑1.40	81.44 ↑1.82
		random	regular	82.86	96.24	68.40	79.96
			PND	84.26 ↑1.40	98.39 ↑2.15	69.66 ↑1.26	81.57 ↑1.61
	AOKVQA	adversarial	regular	79.00	81.25	75.40	78.21
			PND	80.33 ↑1.33	82.60 ↑1.35	76.73 ↑1.33	79.59 ↑1.38
		popular	regular	84.46	92.23	75.26	82.89
			PND	86.36 ↑1.90	94.82 ↑2.59	76.93 ↑1.67	84.94 ↑2.05
		random	regular	85.23	93.71	75.53	83.64
			PND	86.56 ↑1.33	95.21 ↑1.50	77.00 ↑1.47	85.14 ↑1.50
	GQA	adversarial	regular	75.96	79.44	70.06	74.45
			PND	79.50 ↑3.54	83.24 ↑3.80	73.86 ↑3.80	78.27 ↑3.82
		popular	regular	78.26	83.07	71.00	76.56
			PND	81.36 ↑3.10	86.78 ↑3.71	74.00 ↑3.00	79.88 ↑3.32
		random	regular	81.30	90.16	70.26	78.98
			PND	84.76 ↑3.46	94.38 ↑4.22	73.93 ↑3.67	82.92 ↑3.94

Table 5. Performance comparison of LLAVA1.5-13B across POPE datasets.

Model	Dataset	Category	Methods	Accuracy	Precision	Recall	F1
LLAVA1.5-13B	MSCOCO	adversarial	regular	79.80	83.60	74.13	78.58
			PND	84.96 ↑5.16	91.99 ↑8.39	76.60 ↑2.47	83.59 ↑5.01
		popular	regular	82.76	89.60	74.13	81.13
			PND	86.73 ↑3.97	95.84 ↑6.24	76.80 ↑2.67	85.27 ↑4.14
		random	regular	83.40	90.99	74.13	81.70
			PND	87.60 ↑4.20	97.95 ↑6.96	76.80 ↑2.67	86.09 ↑4.39
	AOKVQA	adversarial	regular	76.46	74.84	79.73	77.21
			PND	81.06 ↑4.60	79.27 ↑4.43	84.13 ↑4.40	81.63 ↑4.42
		popular	regular	81.23	82.55	79.20	80.84
			PND	86.96 ↑5.73	88.91 ↑6.36	84.46 ↑5.26	86.63 ↑5.79
		random	regular	83.70	87.03	79.20	82.93
			PND	89.80 ↑6.10	94.55 ↑7.52	84.46 ↑5.26	89.22 ↑6.29
	GQA	adversarial	regular	77.06	75.21	80.73	77.87
			PND	82.80 ↑5.74	81.21 ↑6.00	85.33 ↑4.60	83.22 ↑5.35
		popular	regular	80.73	80.81	80.60	80.71
			PND	86.93 ↑6.20	88.10 ↑7.29	85.40 ↑4.80	86.72 ↑6.01
		random	regular	84.53	87.48	80.60	83.90
			PND	89.83 ↑5.30	93.70 ↑6.22	85.40 ↑4.80	89.36 ↑5.46

3. More Results on the MME Benchmark

To provide a comprehensive validation of PND's impact on general multimodal capabilities, this section presents the full experimental results on all ten perception-oriented tasks of the MME benchmark. As stated in the main paper, a key goal is to demonstrate that our method's benefits extend beyond simple hallucination suppression without causing performance degradation in other areas. The results presented here offer strong evidence that PND, in fact, acts as a holistic performance enhancer.

Table 6. Performance comparison of InstructBlip-13B across POPE datasets.

Model	Dataset	Category	Methods	Accuracy	Precision	Recall	F1
InstructBlip-13B	MSCOCO	adversarial	regular	76.03	76.22	75.66	75.94
			PND	82.83 ↑6.80	84.75 ↑8.53	80.06 ↑4.40	82.34 ↑6.40
		popular	regular	77.60	78.71	75.66	77.15
			PND	85.50 ↑7.90	89.65 ↑10.94	80.26 ↑4.60	84.69 ↑7.54
		random	regular	81.93	86.01	76.26	80.84
			PND	88.46 ↑6.53	96.01 ↑10.00	80.26 ↑4.00	87.43 ↑6.59
	AOKVQA	adversarial	regular	71.13	66.98	83.33	74.27
			PND	76.26 ↑5.13	70.84 ↑3.86	89.26 ↑5.93	78.99 ↑4.72
		popular	regular	75.30	72.02	82.73	77.01
			PND	84.06 ↑8.76	80.96 ↑8.94	89.06 ↑6.33	84.82 ↑7.81
		random	regular	81.83	81.64	82.13	81.88
			PND	89.70 ↑7.87	90.21 ↑8.57	89.06 ↑6.93	89.63 ↑7.75
	GQA	adversarial	regular	69.66	65.94	81.33	72.83
			PND	75.30 ↑5.64	70.72 ↑4.78	86.33 ↑5.00	77.75 ↑4.92
		popular	regular	73.40	70.12	81.53	75.40
			PND	78.96 ↑5.56	75.21 ↑5.09	86.40 ↑4.87	80.42 ↑5.02
		random	regular	81.83	81.98	81.60	81.79
			PND	88.06 ↑6.23	89.37 ↑7.39	86.40 ↑4.80	88.06 ↑6.27

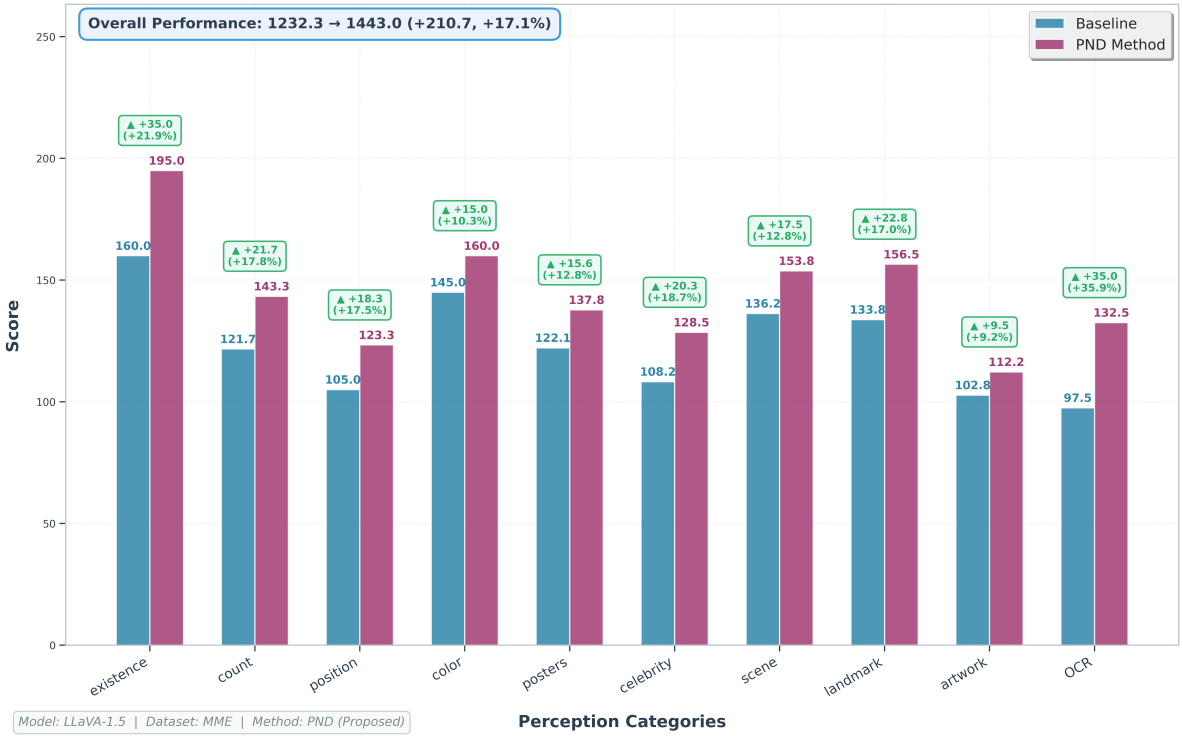

Figure 1. Performance comparison on the 10 MME perception tasks using the LLAVA1.5-7B backbone. Our PND-enhanced model consistently outperforms the greedy decoding baseline across every category, demonstrating its effectiveness as a holistic capability enhancer.

Fig 1 2 3 details the performance of the LLaVA-1.5-7B and InstructBLIP-7B models, with and without our PND framework. A clear and consistent trend of improvement can be observed across every single category for both mod-

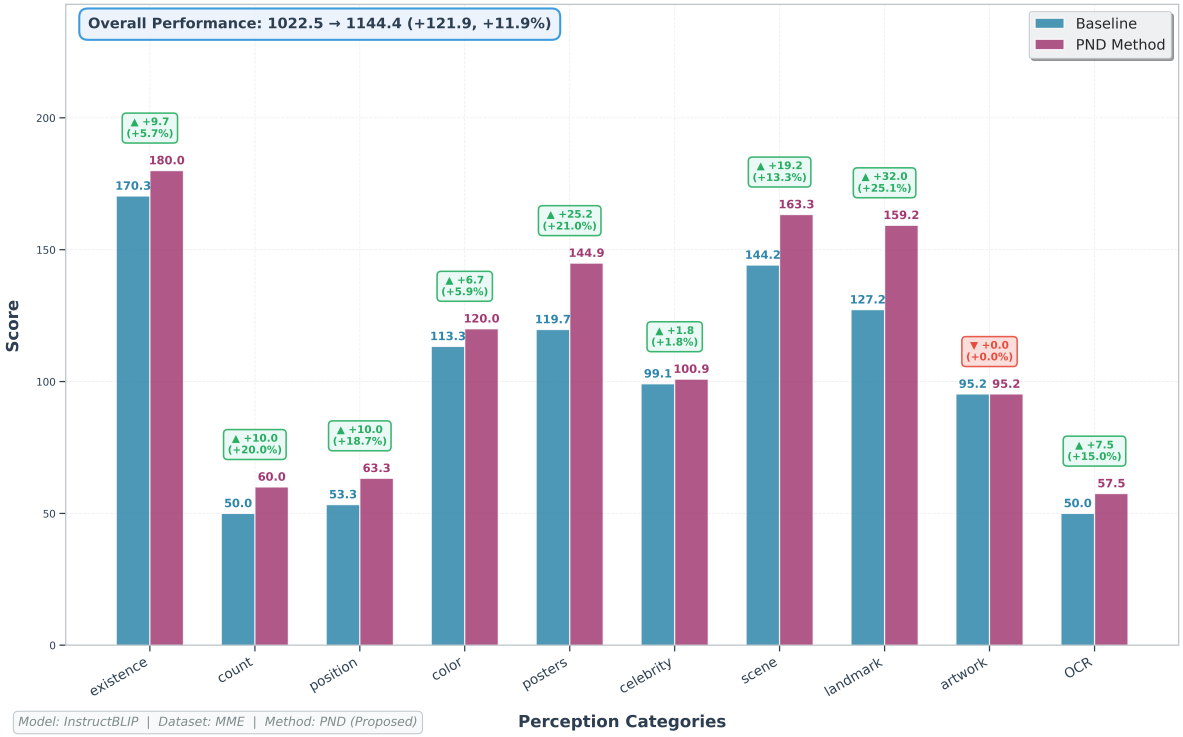

Figure 2. Performance comparison on the 10 MME perception tasks using the InstructBLIP-7B backbone. Our PND-enhanced model consistently outperforms the greedy decoding baseline across every category, demonstrating its effectiveness as a holistic capability enhancer.

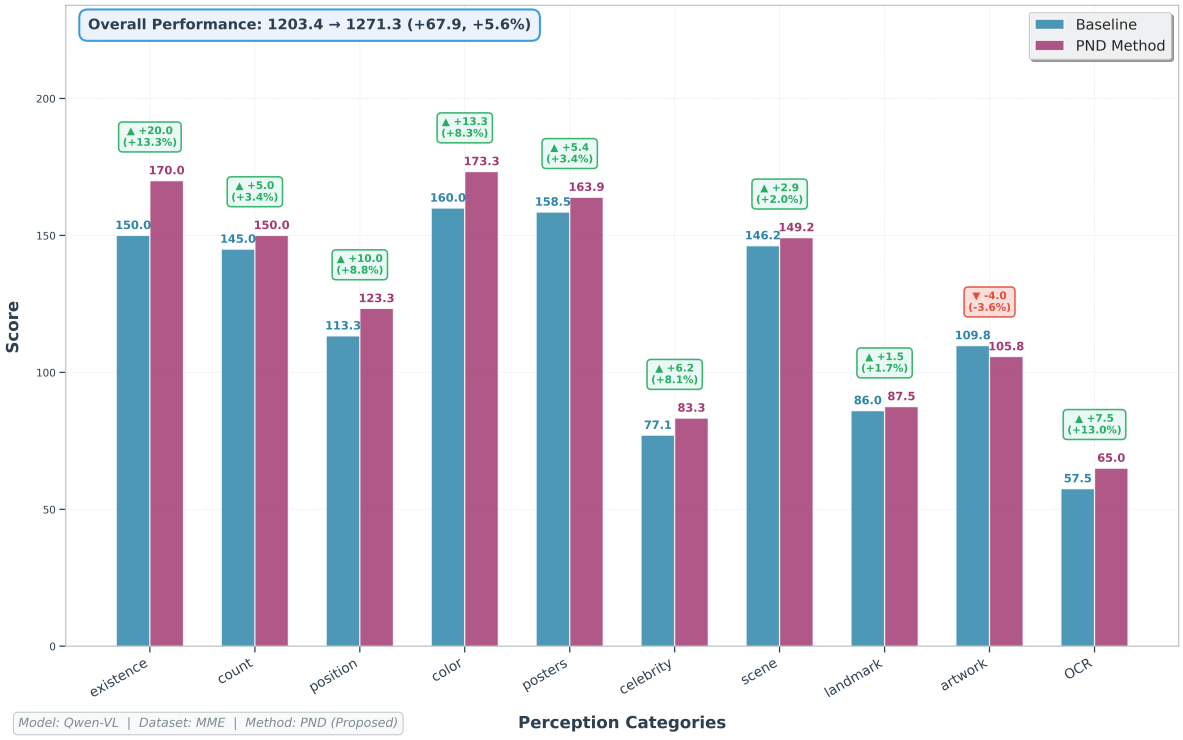

Figure 3. Performance comparison on the 10 MME perception tasks using the Qwen-VL backbone. Our PND-enhanced model consistently outperforms the greedy decoding baseline across every category, demonstrating its effectiveness as a holistic capability enhancer.

Table 7. Ablation Study on Different Layer Configurations

Method	Accuracy	Precision	F1-score
Baseline	78.5	82.7	77.0
Early-layer only	83.2	89.0	82.1
Mid-layer only	83.3	88.5	82.2
Late-layer only	83.0	89.2	81.9
Full Method	**84.0**	**89.8**	**82.7**

els. The gains are not limited to the object- and attribute-level tasks discussed in the main text (*Existence*, *Count*, *Position*, *Color*), but are also evident in more complex and knowledge-intensive domains such as *Celebrity* and *Landmark* recognition, *Artwork* identification, and even challenging fine-grained tasks like Optical Character Recognition (*OCR*)This uniform improvement across such a diverse suite of perceptual tasks is highly significant. It suggests that by forcing a tighter alignment with visual evidence, PND's core mechanism does not merely filter incorrect information but fundamentally enhances the model's ability to "see" and interpret the world more accurately. This provides the strongest evidence for our claim that PND is a general-purpose fidelity enhancement tool, rather than a narrow, specialized fix.

4. More Ablation Experiments

4.1. Ablation on Multi-Layer Fusion Strategy.

To validate the necessity of our multi-layer fusion strategy, we ablate its components by comparing our full PND model against variants that use attention maps from only a single semantic scale: early, middle, or late layers. As shown in Table 7, the full model, which integrates all layers, significantly outperforms every single-scale variant. This confirms our hypothesis that no single layer is sufficient; while early layers excel at localizing fine-grained details and late layers are better at capturing global context, both are prone to errors when used in isolation. Therefore, the integration of these complementary, multi-scale perspectives is a fundamental and necessary component for PND's success in achieving robust visual grounding.

4.2. Ablation on Consensus-Based Destruction Strategy

We conduct an ablation study to validate our specific design for the negative pathway, **Consensus-Based Destruction**, against several alternative negative augmentation strategies. The goal is to show that both the precise targeting via attention consensus and the principled degradation method are crucial for creating effective counterfactuals. As detailed in Table 8, we compare our method against three alternatives:

- *Global Semantic Corruption*: A non-targeted approach that applies global distortions like color shifts and noise, ignoring object salience.
- *Attention-Weighted Noise*: Adds noise proportionally to attention scores without removing the underlying object.
- *Hard Region Removal*: A simpler method that replaces the high-attention region with a blurred or masked-out patch.

As shown in Figure 4, the results confirm our design choices. The *Global Semantic Corruption* strategy performs poorly, as the target object remains largely intact, providing a weak negative signal. *Hard Region Removal* and *Attention-Weighted Noise* are more effective but suboptimal. Hard removal creates strong out-of-distribution artifacts that the model may learn to simply ignore, while weighted noise still leaves the original object features present beneath the noise. Our proposed **Consensus-Based Destruction** yields the best performance. This is because its 'min'-consensus mechanism precisely isolates the core object of interest, and its DDPM-based destruction replaces the object with a semantically empty yet structurally plausible pattern. This creates the most challenging and effective negative counterfactual, forcing the model to reason about the object's absence rather than just identifying an artifact.

4.3. Sensitivity Analysis of Control Gains α, β and γ

We analyze the sensitivity of our framework's key hyperparameters—the positive gain α, the negative gain γ, and the rationality constraint threshold β—on the POPE benchmark. For each parameter, we vary its value while keeping the others fixed at their empirically determined optima.

As shown in Table 9, the analysis reveals distinct optimal settings for each component. Performance peaks at a relatively high positive gain of $\alpha = \mathbf{2.2}$, underscoring the significant benefit of strongly amplifying visually-grounded signals. Conversely, the negative suppression is most effective at a more moderate value of $\gamma = \mathbf{0.4}$, suggesting a measured counterfactual penalty is sufficient to curtail hallucinations without being overly restrictive. Finally, the rationality constraint is optimized at $\beta = \mathbf{0.5}$, indicating a balanced trade-off between adhering to the original distribution and allowing for corrective adjustments. Overall, this analysis confirms that PND's optimal performance stems from a carefully calibrated interplay between a strong positive enhancement, measured negative suppression, and a balanced coherence constraint.

4.4. Ablation on Temperature within PND

We investigate the effect of the temperature sampling parameter on the performance of our PND framework itself. In Table 10, the "Regular (Greedy decoding)" entry serves as the baseline performance without our method. All other

Table 8. Performance Comparison of Different Data Augmentation Methods

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Global Semantic Corruption	82.17	86.21	73.12	80.23
Attention-Weighted Noise	83.78	88.87	75.09	82.67
Hard Region Removal	82.83	87.43	74.34	81.21
DDPM (Ours)	**84.03**	**89.85**	**76.73**	**83.48**

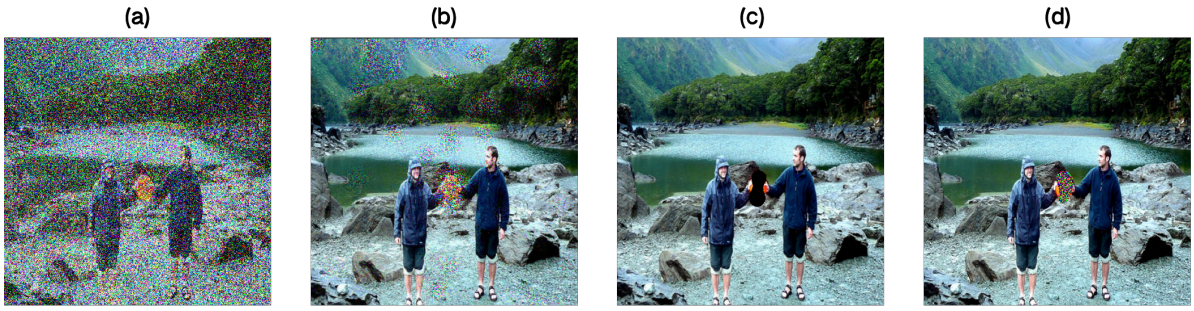

Figure 4. Comparison of different negative augmentation strategies. (a) Global Semantic Corruption: adds global noise but target objects remain visible; (b) Attention-Weighted Noise: adds noise proportionally to attention scores, but original features are partially preserved; (c) Hard Region Removal: directly masks high-attention regions, creating obvious out-of-distribution artifacts; (d) Consensus-Based Destruction (Ours): precisely replaces target objects via DDPM, generating semantically empty yet structurally plausible patterns, creating the most effective negative counterfactuals.

Table 9. Ablation Study on Hyperparameter Settings

Parameter	Value	Accuracy↑	F1-Score↑
Alpha (α)	1.6	0.8310	0.8185
	1.8	0.8383	0.8260
	2.0	0.8400	0.8278
	2.2	**0.8403**	**0.8281**
	2.4	0.8370	0.8246
Gamma (γ)	0.2	0.8397	0.8272
	0.4	**0.8403**	**0.8278**
	0.6	0.8383	0.8256
	0.8	0.8356	0.8244
	1.0	0.8353	0.8224
Beta (β)	0.3	0.8386	0.8257
	0.4	0.8393	0.8263
	0.5	**0.8403**	**0.8278**
	0.6	0.8386	0.8256
	0.7	0.8382	0.8253

results are generated by our full PND framework, evaluated under different temperature settings.

The results in Table 10 first confirm that applying PND with a low temperature (e.g., T=1.0) yields a significant improvement over the baseline. The analysis further reveals that within our framework, performance peaks at T=1.0 and consistently declines as the temperature increases. This indicates that while PND effectively reshapes the output distribution for higher visual fidelity, its precise control is best leveraged in a near-deterministic setting. Higher temperatures introduce excessive randomness that can counteract PND's corrective signals. Therefore, to ensure maximum fidelity, we employ **Temperature**=1 on top of our PND-modified logits in all main experiments.

4.5. Ablation on Sampling Parameters (Top-p and Top-k)

We analyze the impact of stochastic sampling strategies on our PND-enhanced framework by studying two common methods: nucleus sampling (Top-p) and Top-k sampling. This experiment aims to understand the optimal balance between token diversity and the factual fidelity prioritized by our method. The results for both analyses are presented side-by-side in Table 11.

A clear and consistent trend emerges from both sets of experiments. For nucleus sampling, peak performance is achieved with a constrained candidate pool where **top-p is set to 0.2 or 0.4**. Similarly, for Top-k sampling, the optimal result is found at its most deterministic setting, **top-k=1**. In both scenarios, performance consistently degrades as the sampling space expands (by increasing p or k). This strongly indicates that for tasks demanding high visual fidelity, a focused decoding strategy is superior, as introducing stochasticity can re-admit lower-probability, less reliable tokens that PND aims to suppress.

Table 10. Performance Comparison under Different Temperature Parameters

Temperature	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Regular (Greedy decoding)	83.66	89.57	76.20	82.34
1.0	84.03	89.85	76.73	83.48
1.5	83.90	89.57	76.73	82.65
2.0	83.86	89.50	76.73	82.62
2.5	83.56	89.42	76.13	82.24

Table 11. Performance Comparison under Different Top-p and Top-k Parameters

Top-p Parameters					Top-k Parameters				
Top-p	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Top-k	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
0.2	84.66	90.43	77.53	83.48	1	84.70	90.50	77.53	83.51
0.4	84.66	90.43	77.53	83.48	2	83.86	89.50	76.73	82.62
0.6	84.46	90.26	77.26	83.26	3	83.86	89.50	76.73	82.62
0.8	84.03	89.85	76.73	83.48	4	83.86	89.50	76.73	82.62
1.0	84.03	89.85	76.73	83.48	5	83.86	89.50	76.73	82.62

While our analysis identifies that a constrained nucleus ($p = 0.4$) yields the optimal performance, the main experiments in this paper are conducted with $p = 1.0$. This choice is made to ensure a fair and direct comparison against baseline methods, which are typically evaluated without sampling constraints. The results in this ablation therefore show that the performance reported in our main paper is a conservative estimate, with further gains achievable through optimized sampling.

5. More Details for the GCCCE

5.1. Prompt for the GCCCE

Figure 8 illustrates the structure of the prompt used in our GCCCE framework to guide the GPT-4 evaluator. The prompt is designed to be comprehensive and unbiased. It provides the GPT-4 judge with three key pieces of information: 1) the original user instruction given to the model being tested, 2) a ground-truth description of the key visual content in the image, and 3) the MLLM's generated response that is to be evaluated. The core of the prompt is a detailed scoring rubric that explicitly defines each of our four evaluation dimensions: **Relevancy**, **Accuracy**, **Common Sense Plausibility**, and **Fine-grained Precision**. GPT-4 is tasked to provide a score from 1 to 10 for each dimension and to output the results in a structured format for automated parsing.

5.2. Qualitative Case Studies

To provide a more intuitive understanding of our method's effectiveness, this section presents several qualitative case studies. These examples visualize the types of sophisticated failures discussed in the main paper and demonstrate PND's ability to correct them across different MLLMs. The following figures showcase three distinct and challenging hallucination scenarios that PND successfully resolves.

Figure 5 demonstrates PND's ability to suppress a complex, *narrative-level hallucination* on **InstructBLIP**. The baseline model invents a completely fictional story about the banner's design and purpose, whereas PND provides a simple, fact-based description.

Figure 6 showcases its effectiveness against *"runaway" hallucinations* on **LLaVA-1.5**. Here, the baseline model begins with a correct observation (the presence of greenery) but then continues with a plausible yet entirely ungrounded completion (a "roast beef melt sandwich"). PND constrains the entire response to be visually faithful.

Finally, Figure 7 highlights how PND prevents *hallucination-based reasoning errors* on the powerful **Qwen-VL** model. The baseline first hallucinates a nonsensical spatial attribute ("glasses covering his mouth") and then makes faulty deductions based on this false premise. PND breaks this chain of faulty logic by correcting the initial ungrounded assertion.

Collectively, these diverse cases validate PND's robustness in enforcing strict visual fidelity across different models and sophisticated failure modes.

6. Efficiency

As been shown in 12 A practical consideration for our PND framework is its computational overhead. By requiring three parallel forward passes (for the original, positive, and negative inputs) through the MLLM at each decoding step, our method inherently increases inference latency compared to standard greedy decoding. This represents a deliberate trade-off, where we prioritize maximizing visual

[regular]:

The banner attached to the hydrant is in the shape of a giant cannabis leaf, representing the sponsorship of the event by the cannabis industry \u2013 a very unique design choice for a public space, although not something that is encouraged by mainstream organizations.

[PND]:

The banner attached to the hydrant is a green and blue banner with a picture of a fire hydrant on it.

Figure 5. A case study on InstructBlip demonstrating PND's ability to correct severe factual hallucinations. When asked to describe the banner, the baseline model generates a completely ungrounded narrative involving a "giant cannabis leaf". This exemplifies a failure mode where linguistic priors dominate visual evidence. Our PND framework successfully suppresses this complex hallucination, producing a simple, factually accurate description based directly on the visual content.

[regular]:

Yes, there is a roast beef melt sandwich on the plate, and it is garnished with fresh greenery, adding some color and freshness to the meal.

[PND]:

Yes, there is greenery present on the food, as there are green vegetables on the plate.

Figure 6. A case study on LLaVA-1.5 illustrating how PND corrects "runaway" hallucinations that occur after a correct initial statement. The baseline model correctly affirms the presence of greenery but then hallucinates a "roast beef melt sandwich," a plausible but visually ungrounded continuation. PND successfully constrains the generation to be visually faithful, affirming the greenery's presence without fabricating additional objects. This demonstrates PND's ability to maintain visual grounding throughout the entire generation process.

Table 12. Performance vs. inference time (in seconds) on the POPE benchmark for LLaVA-1.5-7B. PND achieves the best overall accuracy and F1-score, while VAF is the most computationally efficient.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	InferTime (s)
Regular	78.53	82.77	72.06	77.04	0.14
VCD	81.13	84.67	76.87	80.43	0.45
VAF	80.88	85.14	74.88	81.39	**0.19**
AGLA	83.13	87.14	**78.12**	82.21	0.63
PND	**84.03**	**89.85**	76.73	**83.48**	0.68

fidelity and reliability over minimizing computational cost. We believe this trade-off is justified for safety-critical applications where accuracy is paramount. A key direction for future work is to improve the efficiency of this contrastive control mechanism. A promising approach is to move beyond requiring an external attention-extraction model (like BLIP). We aim to investigate methods for leveraging the MLLM's **own internal, self-generated attention maps** as the basis for the positive and negative signals. If successful, this would allow for a more tightly integrated and ef-

[regular]:
The picture of the boy with bucked teeth shows him wearing a glasses with rectangular lenses, which are covering his mouth. Therefore, the only facial feature that can be observed is his eyes.

[PND]:
In the picture of the boy with bucked teeth, you can see his bucked teeth and a smile on his face.

Figure 7. A case study on Qwen-VL demonstrating PND's ability to prevent hallucination-based reasoning errors. In response to a query about facial features, the baseline model first hallucinates a nonsensical attribute—that the glasses are "covering his mouth." It then compounds this error by reasoning from this false premise to incorrectly deny the visibility of the boy's mouth and smile. Our PND framework breaks this chain of faulty logic by enforcing strict visual grounding, correctly describing the visible features and preventing the initial attribute hallucination from occurring.

ficient framework that could suppress hallucinations by relying solely on the model's inherent visual grounding capabilities, significantly reducing the additional computational burden.

7. Attention and Saliency Visualization Details

To analyze how multimodal transformers allocate attention across depth, we compute two complementary signals from each cross-modal attention layer: (1) a gradient-based saliency map that captures evidence sensitivity, and (2) the raw attention allocation vectors that describe where the model attends before gradient modulation.

Saliency Map. For each transformer layer i with cross-modal attention matrix $\mathbf{A}^{(i)} \in \mathbb{R}^{H\times T\times S}$, we obtain the gradient of the target loss $\mathcal{L}$ with respect to the attention weights:

$$\nabla \mathbf{A}^{(i)} = \frac{\partial \mathcal{L}}{\partial \mathbf{A}^{(i)}}. \tag{1}$$

Following a Grad-CAM–style formulation, we define the layer saliency score as element-wise magnitude coupling between attention and its gradient:

$$\mathbf{S}^{(i)} = \left| \nabla \mathbf{A}^{(i)} \odot \mathbf{A}^{(i)} \right|, \tag{2}$$

where $\odot$ denotes the Hadamard product. The resulting tensor is averaged over attention heads and token dimensions, yielding a single-layer visual importance map that reflects how strongly the model relies on each image region when producing the final token prediction.

Attention Allocation. To quantify the raw attention distribution across layers, we average $\mathbf{A}^{(i)}$ over heads and spatial positions:

$$\alpha^{(i)} = \mathrm{mean}_{h,s}\left(\mathbf{A}^{(i)}_{h,:,s}\right), \tag{3}$$

producing a token-level attention allocation vector for each layer. This value captures how much each depth stage attends to image features before any gradient-based modulation.

Layer-wise Grouping. To characterize depth-dependent attention behavior, we follow a three-stage partitioning of the L-layer transformer:

$$\text{Early: } 0-10, \quad \text{Middle: } 11-21, \quad \text{Late: } 22-31.$$

For each group $\mathcal{G} \subset \{1,\dots,L\}$, we compute:

$$\bar{\alpha}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \alpha^{(i)}. \tag{4}$$

After normalizing across groups, the triplet $(\bar{\alpha}_{\text{early}}, \bar{\alpha}_{\text{middle}}, \bar{\alpha}_{\text{late}})$ directly reflects the proportion of attention mass assigned to visual tokens at different depths. This forms the basis of Fig. X, where grouped bar charts visualize the evolution of visual attention through the transformer hierarchy.

Aggregation Over Samples. Given N input samples, all saliency maps and attention vectors are averaged:

$$\mathbf{S}^{(i)}_{\text{avg}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{S}^{(i)}_{(n)}, \qquad \alpha^{(i)}_{\text{avg}} = \frac{1}{N}\sum_{n=1}^{N}\alpha^{(i)}_{(n)}.$$

The final visualizations therefore reflect dataset-level attention tendencies rather than single-example noise.

Summary. The visualization pipeline thus combines gradient-based sensitivity (Eq. 2) and raw attention allocation (Eq. 3), aggregated across layers (Eq. 4) and samples, providing a comprehensive picture of how multimodal models distribute attention and how that distribution evolves with depth.

8. Limitations and Future Work

Our work validates PND as an effective method for object hallucination suppression. We identify the following promising directions for future research:

- **Expanding to Broader Hallucination Types:** A primary goal is to extend the PND paradigm beyond object-level errors. Future work will focus on designing more sophisticated augmentation strategies to address subtle hallucinations involving object **attributes** (e.g., incorrect colors) and **relationships** (e.g., faulty spatial relations), which remain significant challenges.
- **Improving Efficiency and Integration:** To address the computational overhead from multiple forward passes, we plan to develop a more integrated framework. A key direction is to leverage the MLLM's **own internal attention mechanisms** as the source for the positive and negative control signals, potentially leading to a highly efficient, single-forward-pass implementation without external modules.
- **Generalizing to Other Modalities:** The core philosophy of PND—enforcing fidelity via counterfactual contrast—is highly generalizable. We aim to adapt this framework for other domains, such as **video-language models** to ensure temporal consistency and prevent action-based hallucinations, and potentially for text-to-audio generation to improve adherence to descriptive prompts.

You are provided with a collection of pre-existing annotations for an image. These annotations are in a textual format and include crucial details, such as object descriptions and relevant question-answer pairs that clarify the image's content. This information should be considered as established facts about the image.

Known Information about the Image:
[annotations]

Here is the instruction for the image:
[questions]

Answer1:
Answer2:

Suppose you are a smart teacher, after looking at the image information above, please score the above answers(0-10) according to the following criteria:
1: Relevancy - whether the response directly follows the instruction.
2: Accuracy - whether the response is accurate concerning the image content.
3: Common Sense Plausibility - whether the response makes logical sense and is plausible in real-world scenarios.
4: Fine-grained Precision - whether the response provides specific, detailed, and precise information rather than vague or general descriptions.

Output format:

Relevancy:
score of the answer1:
score of the answer2:
Accuracy:
score of the answer1:
score of the answer2:
Common Sense Plausibility:
score of the answer1:
score of the answer2:
Fine-grained Precision:
score of the answer1:
score of the answer2:

Figure 8. The prompt structure for our GCCCE framework. It provides the GPT-4 judge with factual context, the model's response, and a clear scoring rubric based on our four evaluation dimensions.