Supplementary Materials of Modeling Visual Flow Leakage to Alleviate Hallucination in Large Vision-Language Models

Anonymous ECCV 2026 Submission

Paper ID #00026

In this supplementary material we provide additional numerical computational cost comparisons between the proposed method and existing methods (in Sec. A). Then, we present more comprehensive results on the POPE dataset (in Sec. B). Next, some more visual examples on handling the additive and subtractive hallucination are presented (in Sec. C). Finally, we provide some further general discussion on hallucination alleviation (in Sec. D).

A Efficiency Comparison

Table 1 compares the computational cost between existing training-free hallucination mitigation methods and the proposed method. The evaluation protocol follows [10], where CHAIR benchmark is used with the LLaVA-1.5 base model under the 128-token setting. The computational cost metrics include both average latency per image (denoted as Avg. Latency, in seconds) and the peak GPU memory usage (denoted as GPU Memory, in MB). The performance metric CHAIR_S is also listed for reference. While achieving the best performance with the lowest CHAIR_S value, the proposed method only has an average latency of 3.42 seconds per image, which is 1% less than the regular decoding baseline. Besides, its GPU memory is nearly the same as the regular decoding model. In summary, FVR achieves the best performance while maintaining efficiency.

Table 1: Efficiency and overall performance comparison using LLaVA-1.5. We report average inference time and memory usage on the CHAIR benchmark (128-token setting), alongside each method's performance. Our method achieves superior results on all benchmarks while incurring minimal overhead. Latency measured on RTX 6000 Ada 48GB; lower latency/memory/CHAIR scores are better.

Method	Avg. Latency (s) ↓	GPU Memory (MB) ↓	CHAIR_S ↓
Regular	3.47 , (×1.00)	14945 , (×1.00)	55.0
VCD [5]	6.97 , (×2.01)	15749 , (×1.05)	54.4
M3ID [3]	7.05 , (×2.03)	15575 , (×1.04)	54.4
OPERA [4]	24.70 , (×7.12)	22706 , (×1.52)	52.6
Woodpecker [11]	10.68 , (×3.08)	22199 , (×1.49)	57.6
HALC [1]	22.61 , (×6.52)	23084 , (×1.54)	51.0
ONLY [10]	3.70 , (×1.07)	14951 , (×1.00)	49.8
Ours (FVR)	3.42 , (×0.99)	14975 (×1.00)	49.6

B More Comprehensive Results on POPE

Due to the limited space, in the main text we only report both the accuracy and F1-score metrics on the POPE dataset. In the supplementary material, we further report the precision and recall metrics as well.

B.1 Results on MS-COCO

LLaVA-1.5 Base Model. Table 2 shows that the proposed FVR attains the strongest performance on the Random and Adversarial settings and remains highly competitive on the Popular setting. In the Random setting, it pushes both accuracy and F1 beyond all the compared methods, while further boosting recall, indicating that it suppresses hallucinations without sacrificing coverage. The advantage becomes most pronounced under the Adversarial setting, where FVR clearly outperforms all the compared methods across accuracy, recall, and F1, highlighting its robustness to the challenging object presence queries.

InstructBLIP Base Model. Table 3 shows that the proposed FVR consistently yields a favorable precision–recall trade-off and is particularly strong on the harder Popular and Adversarial settings. While ONLY sometimes attains the highest precision, FVR substantially improves recall and achieves the best accuracy and F1 in both Popular and Adversarial settings. These outcomes also show that the proposed FVR enhances the robustness to adversarial prompts, rather than merely optimizing for a single metric.

B.2 Results on A-OKVQA

LLaVA-1.5 Base Model. Table 4 shows that the proposed FVR is comparable to the strongest baselines on the Random setting and becomes clearly superior as the prompts grow more adversarial. Under the Popular setting, FVR achieves the best accuracy, recall, and F1, outperforming the compared methods. The advantage is even clearer in the Adversarial setting, where FVR dominates across all four metrics, evidencing that it can retain high coverage of correct objects while being less vulnerable to misleading queries.

InstructBLIP Base Model. Table 5 shows that the proposed FVR remains competitive on the Random setting and becomes the strongest method on both Popular and Adversarial settings. While ONLY often attains the best precision, FVR offers a better global balance, consistently achieving the highest accuracy, recall, and F1 on the more challenging prompt distributions. This pattern indicates that the proposed FVR can stabilize the performance under biased prompts, improving robustness without sacrificing performance on easier cases.

Table 2: Results on MS-COCO [8] (POPE) with LLaVA-1.5 [9] as base model.

Setting	Method	Acc	Prec	Rec	F1
Random	Regular	83.13	81.94	85.00	83.44
	VCD [5]	87.00	86.13	88.20	87.15
	M3ID [3]	87.50	87.38	87.67	87.52
	ONLY [10]	89.70	**89.95**	88.27	89.10
	Ours (FVR)	**89.87**	89.12	**90.80**	**89.90**
Popular	Regular	81.17	78.28	86.27	82.08
	VCD [5]	83.10	79.96	88.33	83.94
	M3ID [3]	84.30	81.58	**88.60**	84.95
	ONLY [10]	**86.00**	**84.44**	88.27	**86.31**
	Ours (FVR)	85.17	83.49	87.67	85.57
Adversarial	Regular	77.43	73.31	86.27	79.26
	VCD [5]	77.17	72.18	88.40	79.47
	M3ID [3]	78.23	73.51	88.27	80.22
	ONLY [10]	79.40	75.00	88.20	81.07
	Ours (FVR)	**82.87**	**78.66**	**90.27**	**83.95**

B.3 Results on GQA

LLaVA-1.5 Base Model. Table 6 shows that the proposed FVR consistently improves the robustness as prompts become more challenging. On the Random setting, FVR is competitive to ONLY's strong performance while further increasing recall. In both the Popular and Adversarial settings, it achieves the best accuracy, recall, and F1 among all the compared methods. These results indicate that FVR not only scales well to the more structured reasoning queries in GQA but also better preserves correctness under the adversarial object-presence tests.

InstructBLIP Base Model. Table 7 shows that the proposed FVR performs the best across all the three settings. It consistently achieves the highest accuracy, recall, and F1 in the Random and Popular settings, and remains the strongest method on the Adversarial setting except for a marginally lower precision than ONLY. To summarize, the proposed FVR yields fewer hallucinations while still answering affirmatively when objects are truly present.

C More Qualitative Examples

In this section, we follow the query format of the POPE dataset, *i.e.*, `Is there a [object] in the image?`, and conduct some yes/no binary object presence experiments. The object presence test is a simple and straight-forward way to inspect the additive and subtractive hallucinations from an LVLM. Some qualitative examples are presented. We use the token activation map [6] as the visualization method. These examples aim to further demonstrate if the proposed method could effectively handle both additive and subtractive hallucinations by presenting visual flow leakage.

Table 3: Results on MS-COCO [8] (POPE) with InstructBLIP [2] as base model.

Setting	Method	Acc	Prec	Rec	F1
Random	Regular	83.07	83.02	83.13	83.08
	VCD [5]	86.23	88.14	83.73	85.88
	M3ID [3]	86.67	88.09	84.80	86.41
	ONLY [10]	**89.23**	**91.83**	86.13	**88.89**
	Ours (FVR)	87.10	86.46	**88.00**	87.24
Popular	Regular	77.00	73.82	83.67	78.44
	VCD [5]	80.07	77.67	84.40	80.89
	M3ID [3]	80.97	77.93	86.40	81.85
	ONLY [10]	83.27	**81.46**	86.13	83.73
	Ours (FVR)	**83.67**	80.98	**88.00**	**84.34**
Adversarial	Regular	74.60	71.26	82.47	76.45
	VCD [5]	77.20	74.29	83.20	78.49
	M3ID [3]	77.47	73.68	85.47	79.14
	ONLY [10]	80.10	**76.89**	86.07	81.22
	Ours (FVR)	**81.07**	76.05	**90.67**	**82.64**

Table 4: Results on A-OKVQA [7] (POPE) with LLaVA-1.5 [9] as base model.

Setting	Method	Acc	Prec	Rec	F1
Random	Regular	81.90	76.63	91.80	83.53
	VCD [5]	83.83	78.05	94.13	85.34
	M3ID [3]	84.67	79.25	93.93	85.97
	ONLY [10]	**86.07**	**80.91**	**94.40**	**87.14**
	Ours (FVR)	85.77	80.53	94.33	86.99
Popular	Regular	75.07	68.58	92.53	78.77
	VCD [5]	76.63	69.59	94.60	80.19
	M3ID [3]	77.80	70.98	94.07	80.91
	ONLY [10]	79.00	**72.17**	94.40	81.80
	Ours (FVR)	**79.10**	72.05	**95.33**	**82.05**
Adversarial	Regular	67.23	61.56	91.80	73.70
	VCD [5]	67.40	61.39	93.80	74.21
	M3ID [3]	68.60	62.22	94.73	75.11
	ONLY [10]	68.70	62.35	94.40	75.70
	Ours (FVR)	**69.70**	**62.67**	**97.53**	**76.27**

C.1 Subtractive Hallucination Alleviation

The results in Fig. 1 and Fig. 2 compare the performance of the baseline and the proposed method in handling subtractive hallucinations. According to the token activation map, in both cases, the baseline can omit the object presence in the image without any proper feature activation. In contrast, the proposed method can precisely discern the object in the image, and assign the object the due feature activation. This observation may be properly explained by the proposed FVR, which mitigates the potential visual flow leakage focus on the key object.

C.2 Additive Hallucination Alleviation

The results in Fig. 3 and Fig. 4 compare the performance of the baseline and the proposed method in handling additive hallucinations. According to the token

Table 5: Results on A-OKVQA [7] (POPE) with InstructBLIP [2] as base model.

Setting	Method	Acc	Prec	Rec	F1
Random	Regular	80.63	76.82	87.73	81.92
	VCD [5]	84.20	80.90	89.53	85.00
	M3ID [3]	85.43	81.77	91.20	86.23
	ONLY [10]	**88.57**	**86.13**	**91.93**	**88.94**
	Ours (FVR)	85.50	82.52	90.00	86.10
Popular	Regular	75.17	70.15	87.60	77.91
	VCD [5]	78.63	73.53	89.47	80.72
	M3ID [3]	78.80	73.38	90.40	81.00
	ONLY [10]	80.83	**75.23**	91.93	82.75
	Ours (FVR)	**81.03**	74.96	**93.33**	**83.15**
Adversarial	Regular	69.87	64.54	88.20	74.54
	VCD [5]	71.00	65.41	89.13	75.45
	M3ID [3]	70.10	64.28	90.47	75.16
	ONLY [10]	72.47	**66.19**	91.87	76.94
	Ours (FVR)	**72.57**	66.04	**93.00**	**77.24**

Table 6: Results on GQA [7] (POPE) with LLaVA-1.5 [9] as base model.

Setting	Method	Acc	Prec	Rec	F1
Random	Regular	82.23	76.32	93.47	84.03
	VCD [5]	83.23	76.73	95.40	85.05
	M3ID [3]	84.20	78.00	95.27	85.77
	ONLY [10]	**86.70**	**80.94**	96.00	**87.83**
	Ours (FVR)	86.10	79.98	**96.20**	87.47
Popular	Regular	73.47	66.83	93.20	77.84
	VCD [5]	72.37	65.27	95.60	77.58
	M3ID [3]	73.87	**66.70**	95.33	78.49
	ONLY [10]	74.03	**66.70**	96.00	78.71
	Ours (FVR)	**74.17**	66.63	**97.00**	**79.06**
Adversarial	Regular	68.60	62.43	93.40	74.84
	VCD [5]	68.83	62.26	95.67	75.43
	M3ID [3]	68.67	62.16	95.40	75.28
	ONLY [10]	69.23	**62.55**	95.87	75.70
	Ours (FVR)	**69.33**	62.50	**96.67**	**75.90**

activation map, in both cases, the baseline can hallucinate the presence of a non-existent object in the image by activating some areas that cause false positives. In contrast, the proposed FVR method reduces the over-activations in the image and by mitigating the potential visual flow leakage to irrelevant image regions, addresses the additive hallucination.

D Further Discussion on Hallucination Alleviation

In this section, we speculate how hallucination alleviation can be improved in the future. Based on the proposed visual flow leakage (VFL) formulation and our observations, we see three directions that can be explored for this important problem.

First, we propose that the community needs **finer-grained hallucination benchmarks** that explicitly separate additive vs. subtractive errors. This is mo-

Table 7: Results on GQA [7] (POPE) with InstructBLIP [2] as base model.

Setting	Method	Acc	Prec	Rec	F1
Random	Regular	79.67	76.05	86.60	80.99
	VCD [5]	82.83	80.16	87.27	83.56
	M3ID [3]	83.07	80.06	88.07	83.87
	ONLY [10]	86.17	**83.84**	89.60	86.63
	Ours (FVR)	**86.23**	82.52	**92.00**	**87.00**
Popular	Regular	73.33	68.72	85.67	76.26
	VCD [5]	76.13	71.10	88.07	78.68
	M3ID [3]	75.17	69.94	88.27	78.04
	ONLY [10]	77.20	**71.79**	89.60	79.72
	Ours (FVR)	**78.27**	71.50	**94.00**	**81.17**
Adversarial	Regular	68.60	63.94	85.33	73.10
	VCD [5]	71.00	65.75	87.67	75.14
	M3ID [3]	71.17	65.79	88.20	75.36
	ONLY [10]	71.93	**65.98**	87.93	75.84
	Ours (FVR)	**72.03**	65.46	**93.33**	**76.90**

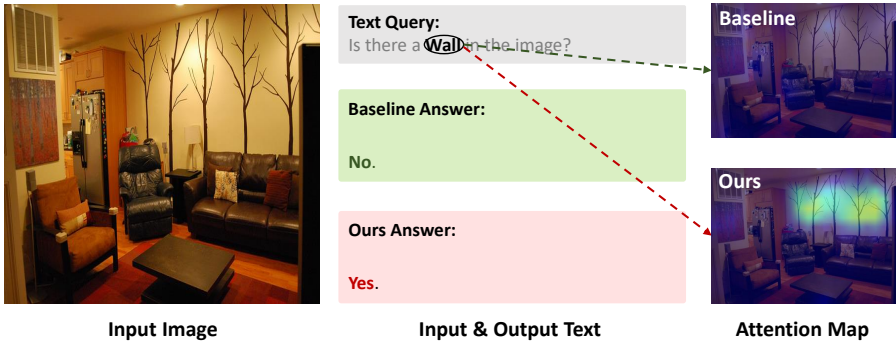

Fig. 1: An example where the proposed method alleviates a subtractive hallucination. Text query: `Is there a wall in the image?`. The token activation map [6] is displayed at the right column.

tivated by our observations on the inherent connection between the visual flow leakage, token activation and commission/omission. Additive (commission) errors refer to fabricating unseen objects, while the subtractive (omission) errors are under-measured in current practice. The existing item-level annotations in hallucination benchmarks can be extended to tag each query with (i) hallucination type (add vs. sub), (ii) severity (minor attribute drift vs. object-level fabrication/omission), and (iii) localization faithfulness against region evidence. These finer-grained hallucination benchmarks will then complement the existing benchmarks for object-presence probing, and further help taking into consideration the missing labels for omission errors and region alignment.

Second, we should further **consider the distribution of positive and negative samples in our benchmarks**, rather than assuming a 50/50 split and adapting evaluation measures under that assumption. By sweeping the class prevalence (e.g., from 20/80 to 80/20 presence rates), benchmarks can overcome

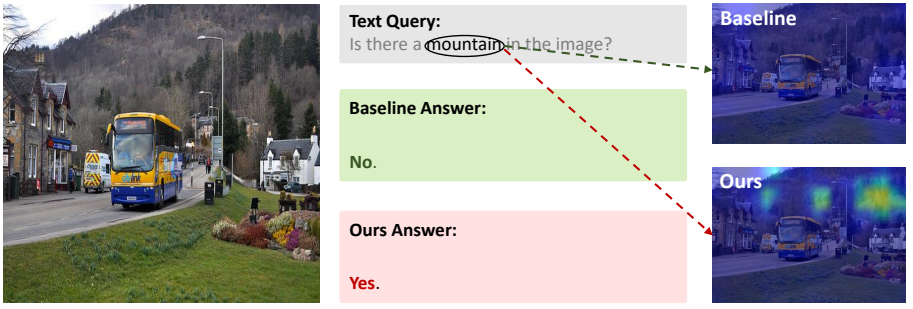

Fig. 2: An example where the proposed method alleviates a subtractive hallucination. Text query: `Is there a mountain in the image?`. The token activation map [6] is displayed at the right column.

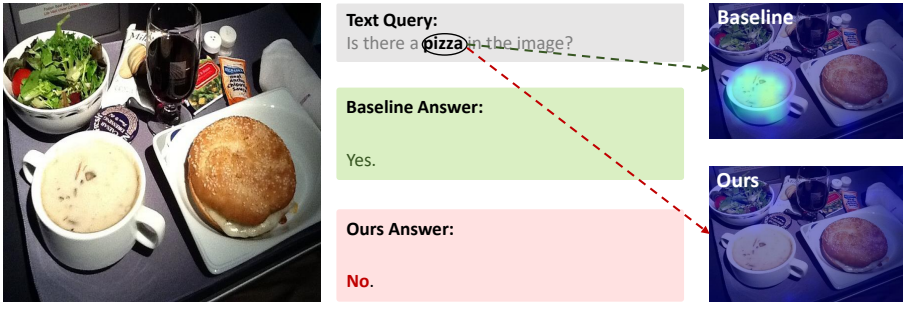

Fig. 3: An example where the proposed method alleviates an additive hallucination. Text query: `Is there a pizza in the image?`. The token activation map [6] is displayed at the right column.

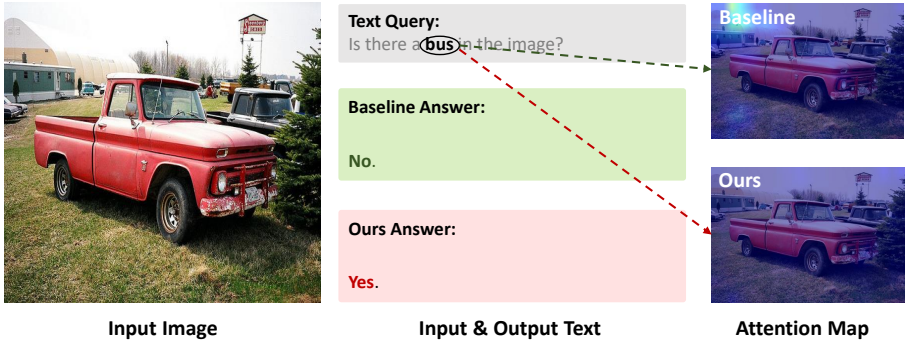

Fig. 4: An example where the proposed method alleviates an additive hallucination. Text query: `Is there a bus in the image?`. The token activation map [6] is displayed at the right column.

the known brittleness of single-point metrics when the object-presence base rate shifts, and make results more comparable across datasets. As an idea, this is similar to looking at a ROC curve, instead of a single accuracy figure, when comparing multiple methods. Benchmarks can allow additive vs. subtractive failure modes to be studied under controlled priors, which can complement existing popular benchmarking evaluations.

Finally, building on the proposed FVR, and our analysis of the results, we believe it will pay off to **consider both head- and layer-conditioned interventions, and joint head-layer interventions, separately**. This type of an approach will enable a more in-depth causal analysis and could further promote new measures that consider belief disagreements with reliability estimations jointly, in the spirit of recent re-evaluations of caption hallucinations.

References

1. Chen, Z., Zhao, Z., Luan, H., Yao, H., Li, B., Zhou, J.: Halc: Object hallucination reduction via adaptive focal-contrast decoding. In: International Conference on Machine Learning. pp. 7824–7846 (2024)
2. Dai, W., Li, Z., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.C.H.: Instructblip: towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems **36**, 49250–49267 (2023)
3. Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., Soatto, S.: Multi-modal hallucination control by visual information grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14303–14312 (2024)
4. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Yu, H., Liu, D., Zhang, W., Yu, A.Z.: Opera: Alleviating hallucination in multimodal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13418–13427 (2024)
5. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13872–13882 (2024)
6. Li, Y., Wang, H., Ding, X., Wang, H., Li, X.: Token activation map to visually explain multimodal LLMs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 48–58 (2025)
7. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023)
8. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision. pp. 740–755 (2014)
9. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
10. Wan, Z., Zhang, C., Yong, S., Ma, M.Q., Stepputtis, S., Morency, L.P., Ramanan, D., Sycara, K., Xie, Y.: Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3225–3234 (2025)

11. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. Science China Information Sciences **67**(12), 220105 (2024)