VASparse: Towards Efficient Visual Hallucination Mitigation via Visual-Aware Token Sparsification

Supplementary Material

1. Experimental Detials

1.1. Experimental Setting

For the settings of the CHAIR and POPE benchmarks, we evaluated the results with the maximum generation token limits of LVLM L_{max} set to 64 and 512, respectively. For the GPT4-assist benchmark [12], we referred to prior work and adopted SHR Evaluation. For the GPT-4 settings, we followed the GPT4-assist configurations and used OpenAI's `gpt-4-0613` version API for evaluation. The parameters for LVLMs were set as follows: `Top-k=False`, `Top-p=1`, and `Temperature=1`. All our experiments, including decoding speed statistics, are conducted on Tesla A100-80G GPUs.

For the proposed token selection strategy, we do not perform sparsification at every decoding step, as this would result in excessive sparsification at each step, leading to overly short generated sequences. In practice, we perform sparsification only after decoding a certain length of new tokens, denoted as L_s. For $L_{\text{max}} = 64$, the beam size is set to 3, and L_s is set to 32. For $L_{\text{max}} = 512$, the beam size is set to 2, and L_s is set to 16. Additionally, in our method, the adaptive plausibility threshold is set to 0.1.

Regarding the comparison of VASparse with SOTAs that are specifically designed for VH mitigation, we adopt the code, hyper-parameters, and pre-trained models of each method outlined in their public repositories and papers respectively. Specifically, for DoLa [3], the parameters are set as follows: the repetition penalty is 1.2, the adaptive plausibility threshold is 0.1, and the pre-mature layers are $[0, 2, \ldots, 32]$. For beam search-based OPERA [4] hyperparameters are set as follows: the self-attention weights scale factor is 50, the attending retrospection threshold is 15, the beam size is 3, and the penalty weights are 1. The VCD [6] hyperparameters are set as follows: the amplification factor is 1, the adaptive plausibility threshold is 0.1, and the diffusion noise step is 500. The HALC [2] hyperparameters are set as follows: the amplification factor is 0.05, the JSD buffer size is 6, the beam size is 1, the FOV sampling uses exponential expansion, the number of sampled FOVs is 4, the exponential growth factor is 0.6, and the adaptive plausibility threshold is 0.1. For post-processing methods, such as LURE and Woodpecker, we follow the settings in HALC [2]. For the SID method [5], we referred to the original configuration in their paper. For all baselines, we follow their implementations and default hyper-parameters as reported in the paper.

1.2. Setting of Empirical Studies

In Section 3, we provide our empirical observations, where all empirical studies are based on LLaVA-1.5 [10]. For the hallucination evaluation results, experiments are conducted on 500 samples randomly selected from the MSCOCO dataset. For decoding speed, we calculate the average number of tokens decoded per second by the model on the 500 samples. Token sparsification methods, such as FastV [1] and SparseVLM [11], directly prune image tokens.

2. Proof of Theorem 1

We aim to prove that in the following optimization problem, our strategy achieves a globally optimal solution:

$$
\begin{aligned}
\min_{M} \quad & \mathcal{E}(M) = \sum_{i=1}^{L} (y_i - M_i y_i)^2 - \lambda P_i M_i \\
\text{s.t.} \quad & M_i \in \{0, 1\}, \quad \forall i = 1, 2, \ldots, L, \\
& \sum_{i=1}^{L} M_i = S,
\end{aligned}
\tag{1}
$$

where:

- $y_i = \langle q, K_i \rangle$ is the inner product of the query vector q and the key matrix vector K_i.
- $P_i \geq 0$ is the selection probability, indicating the priority of selecting a specific element.
- $M_i \in \{0, 1\}$ denotes whether the i-th element is selected.
- The constraint requires exactly S elements in M to be 1.

The goal is to minimize the total error $\mathcal{E}(M)$ when selecting S elements.

Proof First, expand and simplify the objective function $\mathcal{E}(M)$:

$$
\begin{aligned}
\mathcal{E}(M) &= \sum_{i=1}^{L} \left[(y_i - M_i y_i)^2 - \lambda P_i M_i \right] \\
&= \sum_{i=1}^{L} \left[y_i^2 (1 - M_i)^2 - \lambda P_i M_i \right].
\end{aligned}
\tag{2}
$$

Since $M_i \in \{0, 1\}$, it follows that $M_i^2 = M_i$ and $(1 - M_i)^2 = 1 - 2M_i + M_i^2 = 1 - 2M_i + M_i$. Substituting these simplifications, we get:

$$\begin{aligned}\mathcal{E}(M) &= \sum_{i=1}^{L}\left[y_i^2(1-2M_i+M_i)-\lambda P_i M_i\right] \\ &= \sum_{i=1}^{L}\left[y_i^2(1-M_i)-\lambda P_i M_i\right].\end{aligned} \tag{3}$$

Next, remove the constant term $\sum_{i=1}^{L} y_i^2$, as it does not affect the optimization:

$$\begin{aligned}\mathcal{E}(M) &= \sum_{i=1}^{L}\left[y_i^2 - y_i^2 M_i - \lambda P_i M_i\right] \\ &= \sum_{i=1}^{L}\left[y_i^2 - M_i(y_i^2+\lambda P_i)\right].\end{aligned} \tag{4}$$

Thus, the optimization problem can be equivalently transformed into maximizing the following objective function:

$$\begin{aligned}\max_{M} \quad & \sum_{i=1}^{L} M_i(y_i^2+\lambda P_i) \\ \text{s.t.} \quad & M_i \in \{0,1\}, \quad \forall i, \\ & \sum_{i=1}^{L} M_i = S.\end{aligned} \tag{5}$$

Our goal is now to select S elements to maximize the total reward $\sum_{i=1}^{L} M_i\delta_i$, where:

$$\delta_i = y_i^2 + \lambda P_i. \tag{6}$$

Characteristics of the Objective Function

- **Linearity:** The objective function is linear with respect to M_i, with no interaction terms between M_i and M_j.
- **Independence:** The contribution of each M_i to the total reward depends solely on its own δ_i, independent of other variables M_j.

We employ the following selection strategy:

1. Compute the marginal reward δ_i for each element:

$$\delta_i = y_i^2 + \lambda P_i. \tag{7}$$

2. Sort all elements by δ_i in descending order.
3. Select the top S elements, setting their corresponding M_i to 1, and the rest to 0.

Optimality of the Strategy For any feasible solution M, we have:

$$\sum_{i=1}^{L} M_i = S, \quad M_i \in \{0,1\}. \tag{8}$$

Define the total reward for a solution M as:

$$R(M) = \sum_{i=1}^{L} M_i\delta_i. \tag{9}$$

Let the solution chosen by our strategy be M^{ours}, with total reward:

$$R_{\text{ours}} = \sum_{i=1}^{L} M_i^{\text{ours}}\delta_i, \tag{10}$$

where $M_i^{\text{ours}} = 1$ if i belongs to the top S elements with the highest δ_i, and $M_i^{\text{ours}} = 0$ otherwise. Since δ_i is sorted in descending order, the elements chosen by our strategy have the highest individual scores.

For any element i in M such that $M_i = 1$, if its score δ_i is smaller than that of an unselected element j (i.e., $M_j = 0$), swapping these two elements would result in a new total reward:

$$R'(M) = R(M) - \delta_i + \delta_j. \tag{11}$$

Since $\delta_j > \delta_i$, this increases the total reward. Thus, any feasible solution M with lower-scoring elements can always be improved by following our selection strategy.

Finally, for any feasible solution M, we have:

$$R_{\text{ours}} = \sum_{i=1}^{L} M_i^{\text{ours}}\delta_i \geq \sum_{i=1}^{L} M_i\delta_i = R(M). \tag{12}$$

Conclusion The total reward achieved by our algorithm is no less than that of any other feasible solution. Therefore, the solution provided by our strategy is globally optimal.

3. More evidence of empirical observations

We present additional evidence on the attention sparsity and attention sinking of LLaVA-1.5 in Figure 1. Our research findings confirm that the self-attention in most layers of the LVLM decoder is sparse. Furthermore, we observe a significant attention ”sinking” effect on certain text tokens within the LVLM’s attention mechanisms. These results further confirm the characteristics of attention sparsity and attention sinking in LVLMs.

4. More results on CHAIR benchmark

We set the maximum generation length to 512 and evaluated our method using the CHAIR benchmark, as shown in Table 1. We can observe that when setting the maximum generation length to 512, our method still outperforms the baseline method in most metrics, while achieving competitive decoding speed. For all results, we set different random seeds and run them five times, and record the average of the results from the five runs.

5. More results on POPE benchmark

Following HALC [2], we utilize offline POPE (OPOPE) benchmark with both accuracy and F1-score as evaluation metrics to assess VH. We conduct experiments under two different maximum text length settings: 64 and 512 tokens. As

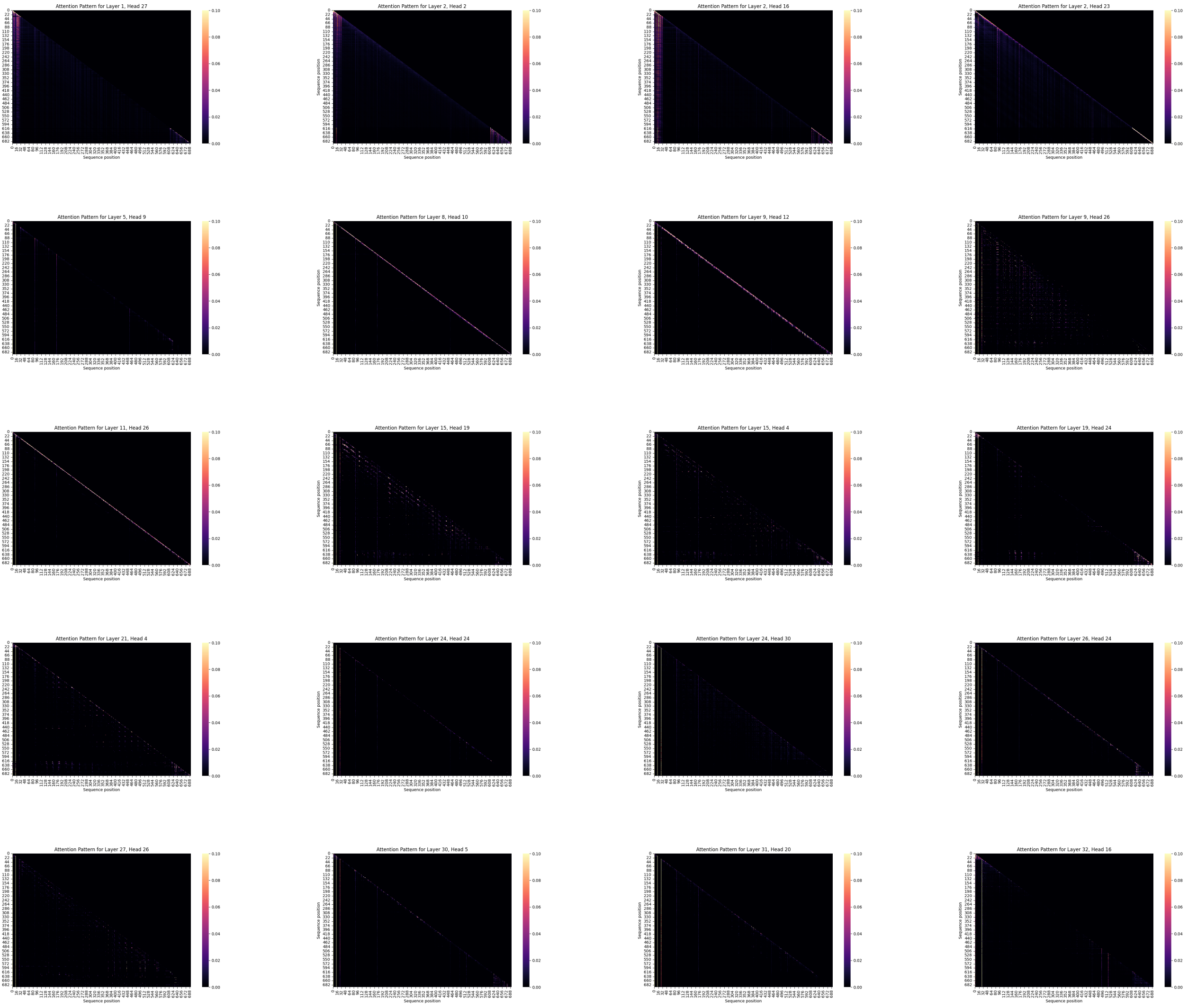


Figure 1. More visualization and evidence of sparsity of attention and sinking attention on the LLaVA-1.5.

shown in Tables 2 and 3, we observe several key findings: (1) VASparse consistently achieves optimal performance across most experimental settings, surpassing both state-of-the-art decoding methods and post-processing approaches under both the 64 and 512-token settings. (2) The effectiveness of VASparse remains robust across different text length configurations. The performance improvements persist when extending the maximum text length from 64 to 512 tokens, indicating the method's scalability; (3) VASparse demonstrates consistent VH mitigation capabilities across three distinct LVLM architectures, highlighting its versatility and plug-and-play nature. This architectural agnosticism suggests broad applicability across different model frameworks.

6. Qualitative Study

To visually demonstrate the effectiveness of our approach, we present generated captions from our method and baseline approaches in Figures 2 and 3on the MSCOCO dataset. We consistently used *Please describe this image in detail.* as the input prompt across all experiments. The results indicate that captions generated by our VASparse method exhibit notably fewer hallucinated descriptions. To further evaluate our method's effectiveness in mitigating VH, we conducted experiments on LLaVA-Bench [8], which consists of 24 distinct images with expert-annotated descriptions and corresponding evaluation questions. In alignment with previous studies [2, 6, 9], we employed this benchmark for qualitative assessment of VH reduction. The visual results are presented in Figure 4, 5 and 6, where we consistently

Methods	LLaVA-1.5			MiniGPT-4			mPLUG-Owl2		
	CHAIR$_i$ ↓	CHAIR$_s$ ↓	TPS↑	CHAIR$_i$ ↓	CHAIR$_s$ ↓	TPS↑	CHAIR$_i$ ↓	CHAIR$_s$ ↓	TPS↑
FastV$*$	17.62	51.94	33.18	19.67	54.59	38.05	22.65	67.68	24.37
SparseVLM$*$	18.09	52.40	32.06	19.85	55.27	37.49	23.04	68.42	23.19
Woodpecker†	13.27	49.72	-	13.76	44.07	-	18.39	59.58	-
LURE†	13.08	47.95	-	13.49	43.92	-	17.85	57.73	-
Greedy	14.63	49.66	31.17	14.06	43.65	36.28	19.07	61.28	19.96
Beam Search	13.62	48.89	29.89	13.90	44.45	32.10	17.12	54.66	19.58
OPERA	12.98	47.60	4.07	15.42	42.42	5.27	17.86	56.29	3.49
VCD	14.82	49.76	17.55	17.09	43.80	17.68	19.46	62.44	9.77
DoLa	13.75	50.03	23.40	13.85	44.20	24.75	18.43	60.18	14.23
SID	13.29	47.09	19.57	13.68	43.65	22.67	18.47	60.82	12.85
HALC	12.93	46.35	2.04	13.73	43.68	3.68	17.63	56.12	1.50
Ours	**12.46**	**46.21**	27.53	**13.29**	**43.02**	29.74	**17.02**	**53.70**	17.86

Table 1. Comparison of the average CHAIR evaluation results (instance levels CHAIR$_i$ and sentence levels CHAIR$_s$)and token per second (TPS) during decoding with different baselines on MSCOCO datasets of five random runs, with whole statistical results in Appendix. * represents the image token sparsity method and † is the post-hoc methods.

	Methods	Max New Token 512					
		Random		Popular		Adversarial	
		Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
LLaVA-1.5	Greedy	77.19	71.74	72.74	67.99	71.18	66.76
	Beam Search	78.38	**73.60**	75.06	70.74	72.87	**68.96**
	OPERA	78.01	72.98	74.31	69.81	73.25	68.95
	VCD	77.82	72.98	74.56	70.19	72.62	68.63
	DoLa	76.69	71.07	72.12	67.26	70.80	66.23
	SID	77.93	72.84	74.89	69.34	72.77	68.30
	HALC	77.08	72.16	74.15	69.09	72.46	68.04
	Ours	**78.57**	72.33	**75.16**	**70.51**	**73.37**	68.88
MiniGPT4	Greedy	69.14	56.55	65.84	54.04	65.67	53.91
	Beam Search	68.90	55.78	65.67	53.32	65.61	53.28
	OPERA	69.77	57.24	**66.90**	55.04	65.38	53.85
	VCD	69.32	57.05	65.14	53.89	65.25	53.98
	DoLa	69.02	56.31	66.08	54.07	65.84	53.90
	SID	69.05	56.53	65.58	53.53	65.45	53.52
	HALC	69.13	56.86	65.62	53.63	65.73	53.69
	Ours	**69.84**	**57.36**	66.31	**55.68**	**66.02**	**54.10**
mPLUG-Owl2	Greedy	76.21	70.16	71.61	81.48	69.38	64.63
	Beam Search	75.83	69.87	71.83	81.75	69.02	64.29
	OPERA	73.56	65.33	70.32	**84.43**	67.90	60.82
	VCD	75.74	69.16	70.67	80.63	69.08	63.77
	DoLa	76.33	70.22	71.67	81.72	69.55	64.71
	SID	75.72	69.31	71.79	81.90	69.12	64.10
	HALC	75.62	69.04	70.24	82.40	68.35	63.51
	Ours	**76.51**	**70.45**	**72.19**	82.44	**69.72**	**64.98**

Table 2. Comparison of the average Accuracy and F1-score evaluation results under different settings (i.e., *Random, Popular, Adversarial*) with different baselines and our VASparse on offline POPE benchmark [2, 7] of five random runs, with whole statistical results in Appendix. Higher F1-score indicate better performance and bold indicates the best results. We set the maximum generated length to 512.

	Methods	Max New Token 64					
		Random		Popular		Adversarial	
		Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
LLaVA-1.5	Woodpecker†	70.82	59.73	68.62	58.53	68.49	58.07
	LURE†	71.10	60.08	69.17	58.63	69.16	58.34
	Greedy	70.55	58.75	68.93	57.42	67.91	56.64
	Beam Search	71.32	60.38	69.31	58.98	69.02	58.43
	OPERA	71.02	59.80	69.31	58.42	68.79	58.00
	VCD	71.08	60.05	68.96	58.34	68.55	58.02
	DoLa	70.73	59.36	69.14	58.08	68.32	57.44
	SID	71.47	61.63	69.42	59.62	69.36	58.83
	HALC	70.76	60.46	69.17	59.33	69.25	58.50
	Ours	**72.03**	**62.13**	**70.18**	**60.93**	**70.31**	**59.20**
MiniGPT4	Woodpecker†	68.05	53.84	65.49	51.70	65.06	51.27
	LURE†	68.12	53.91	65.96	52.37	65.17	51.38
	Greedy	68.02	53.71	65.31	51.68	65.41	51.92
	Beam Search	68.26	53.97	66.02	52.27	65.55	51.93
	OPERA	67.73	53.08	65.37	51.32	65.19	51.20
	VCD	67.96	53.26	65.61	51.50	65.02	51.07
	DoLa	68.08	53.83	65.55	51.93	65.25	51.72
	SID	68.09	53.86	65.69	51.98	65.28	51.77
	HALC	68.18	53.93	65.83	52.06	65.31	51.80
	Ours	**68.55**	**54.87**	**66.23**	**52.93**	**65.91**	**52.70**
mPLUG-Owl2	Woodpecker†	68.61	58.10	67.28	53.07	66.58	55.42
	LURE†	68.78	58.28	67.35	53.15	66.89	55.65
	Greedy	69.67	57.40	68.02	53.43	67.14	55.43
	Beam Search	68.79	55.31	66.92	52.89	65.90	53.12
	OPERA	69.08	55.70	67.37	53.41	66.43	53.66
	VCD	**70.49**	**58.63**	68.55	54.87	67.31	56.13
	DoLa	69.61	57.21	67.90	53.38	67.08	55.24
	SID	69.34	55.82	67.80	53.46	67.01	56.07
	HALC	69.66	56.29	67.67	53.38	66.95	55.84
	Ours	70.38	58.27	**68.70**	**55.28**	**67.86**	**56.77**

Table 3. Comparison of the average Accuracy and F1-score evaluation results under different settings (i.e., *Random, Popular, Adversarial*) with different baselines and our VASparse on offline POPE benchmark [2, 7] of five random runs, with whole statistical results in Appendix. Higher F1-score indicate better performance and bold indicates the best results. † denotes the post-hoc method. We set the maximum generated length to 64.

used the prompt *Please describe this image in detail.* to generate image captions.

References

[1] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models, 2024. 1

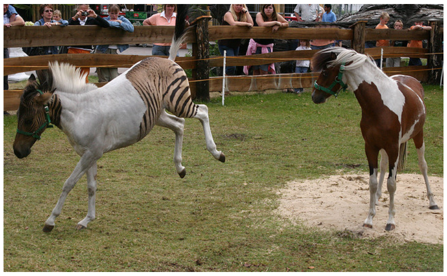
[2] Zhaorun Chen, Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *ArXiv*, abs/2403.00425, 2024. 1, 2, 3, 4, 5

[3] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *ArXiv*, abs/2309.03883, 2023. 1

[4] Qidong Huang, Xiao wen Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Neng H. Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *ArXiv*, abs/2311.17911, 2023. 1

[5] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models, 2024. 1

[6] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shi-



VASparse (Ours): The image features two zebras standing in a grassy field, with one of them jumping in the air. The zebra in the air appears to be performing a trick, while the other zebra stands nearby, possibly observing the action. There are several people in the scene, some of whom are watching the zebras from behind a fence. The spectators are scattered around the field, with some standing closer to the zebras and others further away.

Greedy: The image features a zebra and a horse standing in a grassy field, both appearing to be enjoying their time together. The zebra is positioned on the left side of the field, while the horse is on the right. There are several people in the scene, with some standing near the fence and others scattered around the field. A few individuals are closer to the zebra, while others are near the horse. A car can be seen in the background, parked on the left side of the field.

HALC: The image features a zebra and a horse in a fenced-in area, likely at a zoo or a similar facility. The zebra is in the middle of a jump, while the horse stands nearby, watching the zebra's actions. In addition to the main zebra and horse, there are a few other people scattered around the scene, some closer to the fence and others further away. A car can be seen in the background, likely belonging to one of the visitors.

VCD: In this image, there is a zebra standing on a sandy area, possibly in a zoo enclosure. The zebra appears to be in motion, possibly running or galloping. Additionally, there is a horse present in the scene, standing on the other side of a fence. There are several individuals in various positions around the fence, watching the zebra and the horse. A car can be seen in the background, stopping by the roadside.

DoLa: The image features two zebras in a grassy field, with one of them jumping in the air. The zebra in the air is positioned towards the left side of the field, while the other zebra is standing on the right side. There are several people watching the zebras from behind a wooden fence. Some of them are standing closer to the zebras, while others are positioned further away. The onlookers appear to be enjoying the sight of the zebras in action. There are several cars parked by the fence in the distance.

Figure 2. Qualitative results comparing our VASparse and other methods with LLaVA-1.5 backbone.

jian Lu, Chunyan Miao, and Li Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *ArXiv*, abs/2311.16922, 2023. 1, 3

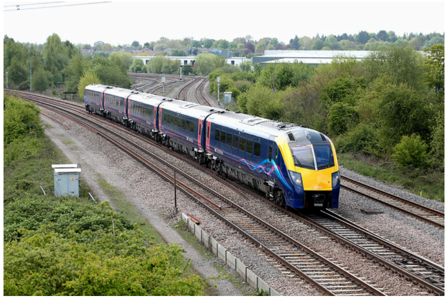
[7] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023. 4, 5

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *ArXiv*, abs/2310.03744, 2023. 3

[9] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xingguo Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *ArXiv*, abs/2310.16045, 2023. 3

[10] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv*, abs/2306.02858, 2023. 1

[11] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng,



VASparse (Ours): The image features a blue and yellow passenger train traveling down the tracks. The train is quite long, occupying a significant portion of the scene. The train is moving along the tracks, which are surrounded by trees and bushes, creating a serene and natural environment. There are multiple train tracks visible in the image, with the main train traveling on one of them. Additionally, there are two smaller trains in the scene.

Greedy: The image depicts a blue and yellow train traveling down the tracks, surrounded by trees and bushes. There are multiple train tracks visible in the scene, with the main train occupying one of them. In addition to the train, there are several traffic lights scattered throughout the image, likely indicating the presence of intersections or crossings. These traffic lights are positioned at various heights and distances from the train.

SID: The image features a blue and yellow train traveling down the tracks, surrounded by trees. The train is moving along the tracks, and there are several other tracks visible in the scene. The train appears to be a passenger train, as it is designed for transporting people rather than cargo. The train's colorful design and the presence of multiple tracks suggest that this is a busy transportation hub or a well-connected railway network.

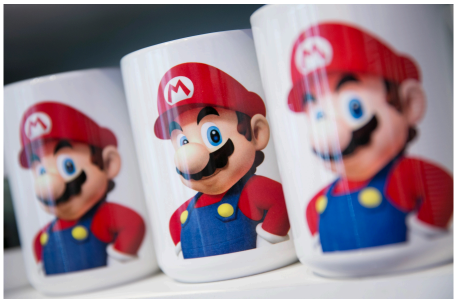
VCD: The image features a blue and yellow passenger train traveling down the tracks. The train is quite long, stretching across the entire width of the scene. It appears to be moving at a moderate speed, as it is not too close to the viewer. There are several traffic lights along the train tracks, with some located closer to the train and others further away. These traffic lights are likely used to control the train's movement and ensure safe passage through the area.

OPERA: The image features a blue and yellow train traveling along the railroad tracks. It appears to be a commuter train with multiple cars, as it passes through the countryside. There are several trees visible in the background, providing a scenic atmosphere for the train journey. The train is situated between two sets of train tracks, which indicates a well-organized transportation system.

Figure 3. Qualitative results comparing our VASparse and other methods with LLaVA-1.5 backbone.

Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024. 1

[12] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023. 1



VASparse (Ours): The image showcases a group of three coffee cups on a counter, each featuring a large, colorful mug with a Super Mario Bros. design. The cups are stacked in a row, with one placed in the middle and the other two on either side, creating a visually appealing display. The coffee cups are all positioned side by side, making it easy for viewers to appreciate the design and details of each cup.

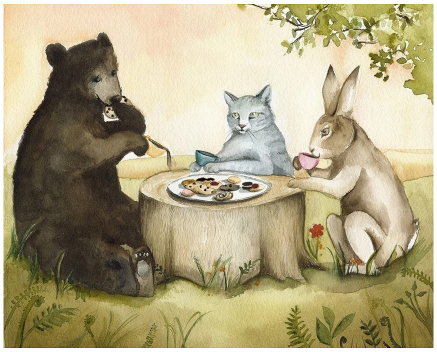
Greedy: The image features a collection of four coffee mugs, each with a unique design. The mugs are arranged in a row, with one mug on the left, another in the middle, and two more on the right side. The mugs are all white and have a Mario theme, with each mug featuring a different Mario character. The characters are displayed in various positions, adding visual interest to the arrangement.

OPERA: The image showcases a collection of four coffee mugs, which featuring a different Mario character. All mugs are placed side by side, showcasing their unique designs. The Mario characters are displayed on the front of each mug, making them an appealing choice for fans of the iconic video game character.

VCD: The image showcases a collection of four coffee mugs, each featuring a different Mario character. The mugs are placed side by side, showcasing their unique designs. The Mario characters are prominently displayed on the front of each mug, making them an appealing choice for fans of the iconic video game character.

SID: The image showcases a collection of four coffee mugs, each featuring a different Mario character. The mugs are arranged in a row, with each character occupying a separate mug. The first mug on the left features a close-up of Mario's face, while the second mug on the right displays a full-body image of the iconic character. The third mug in the middle showcases a side view of Mario, and the fourth mug on the far right showcases a back view of the character.

Figure 4. LLaVA-Bench results comparing our VASparse and other methods with LLaVA-1.5 backbone.



VASparse (Ours): The image features a group of three animals sitting around a dining table, enjoying a meal together. The animals include a large brown bear, a cat, and a rabbit. They are all seated on the grass, with the bear on the left side, the cat in the middle, and the rabbit on the right side of the table. A plate of food is placed on the table, and there are several donuts scattered around the table. The animals are eating and drinking from cups.

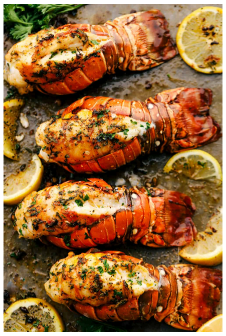
Greedy: The image features a group of three animals sitting around a dining table, enjoying a meal together. The animals include a large brown bear, a cat, and a rabbit. They are all seated on the grass, with the bear on the left side, the cat in the middle, and the rabbit on the right side of the table. The table is set with various food items, including a cake, a bowl, and a cup. The cake is placed in the center of the table, while the bowl and cup are positioned closer to the right side.

OPERA: The image depicts a whimsical scene of three animals sitting around a dining table, enjoying a meal together. The table is filled with a variety of food items, including donuts, cakes, and sandwiches. The animals are positioned around the table, with a bear on the left side, a cat in the middle, and a rabbit on the right side. Each animal has a cup and a spoon placed in front of them, indicating that they are eating from bowls.

VCD: The image portrays a charming scene of three animals sitting at a wooden dining table. A black bear is on one side, a white cat on another side, and a rabbit on the opposite side. All three animals are eating food, with a cup placed in front of each animal. There is a plate of food on the table, and several carrots are spread out around the table. The animals appear to be enjoying their meal together, creating a whimsical and heartwarming atmosphere.

DoLa: The image depicts a whimsical scene of a bear, a cat, and a rabbit sitting around a dining table, enjoying a meal together. The bear is on the left side of the table, the cat is in the middle, and the rabbit is on the right side. The table is set with a variety of food items, including a cake, a bowl, and a cup. The cake is placed in the center of the table, while the bowl and cup are positioned closer to the right side. The animals seem to be enjoying their time together.

Figure 5. LLaVA-Bench results comparing our VASparse and other methods with LLaVA-1.5 backbone.



VASparse (Ours): The image features a close-up view of a delicious meal consisting of several cooked lobster tails. The lobster tails are accompanied by a few lemon wedges, which add a refreshing touch to the dish. The combination of the cooked lobster tails and the lemon wedges creates an appetizing and visually appealing presentation.

Greedy: The image showcases a close-up view of several cooked lobster tails with lemon wedges placed on top of them. The lobster tails are arranged in a visually appealing manner, with some tails overlapping each other. The lemon wedges are spread across the tails, adding a burst of citrus flavor to the dish. The combination of the cooked lobster tails and the lemon wedges creates a mouth-watering and appetizing presentation.

HALC: The image showcases a close-up view of several cooked lobster tails on a dining table. The lobster tails are arranged in a visually appealing manner, with some tails overlapping each other. There are a total of five lobster tails in the scene, each with varying sizes and positions. In addition to the lobster tails, there are three lemon wedges placed around the tails, adding a touch of citrus flavor to the dish. The combination of the cooked lobster tails.

VCD: The image features a close-up view of a delicious meal consisting of several cooked lobster tails. The tails are placed on a dining table, and they are generously seasoned with herbs and spices. The lobster tails are arranged in a visually appealing manner, with some tails overlapping each other. In addition to the lobster tails, there are a couple of lemon wedges placed on the table, adding a touch of citrus flavor to the dish.

DoLa: The image showcases a delectable arrangement of cooked lobster tails and lemon wedges on a dining table. There are six large lobster tails laid out, some placed next to each other, and others arranged in a visually appealing manner. The lemon wedges are scattered around the lobster tails, with a total of seven lemon wedges visible in the scene.

Figure 6. LLaVA-Bench results comparing our VASparse and other methods with LLaVA-1.5 backbone.