Steering LVLMs via Sparse Autoencoder for Hallucination Mitigation

Zhenglin Hua[1,2†], Jinghan He[3,4†], Zijun Yao[5], Tianxu Han[6], Haiyun Guo[3,4*], Yuheng Jia[1,2*], Junfeng Fang[7*]

[1]School of Computer Science and Engineering, Southeast University

[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications

[3]Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences

[4]School of Artificial Intelligence, University of Chinese Academy of Sciences

[5]Department of Computer Science and Technology, Tsinghua University

[6]Wuhan University of Technology [7]National University of Singapore

huazhenglin2003@gmail.com, hejinghan2022@ia.ac.cn

Abstract

Large vision-language models (LVLMs) have achieved remarkable performance on multimodal tasks. However, they still suffer from hallucinations, generating text inconsistent with visual input, posing significant risks in real-world applications. Existing approaches to address this issue focus on incorporating external knowledge bases, alignment training, or decoding strategies, all of which require substantial computational cost and time. Recent works try to explore more efficient alternatives by adjusting LVLMs' internal representations. Although promising, these methods may cause hallucinations to be insufficiently suppressed or lead to excessive interventions that negatively affect normal semantics. In this work, we leverage sparse autoencoders (SAEs) to identify semantic directions closely associated with faithfulness or hallucination, extracting more precise and disentangled hallucination-related representations. Our analysis demonstrates that interventions along the identified faithful direction can mitigate hallucinations, while those along the hallucinatory direction can exacerbate them. Building on these insights, we propose **S**teering LVLMs via **S**AE **L**atent Directions (SSL), a plug-and-play method based on SAE-derived latent directions to mitigate hallucinations in LVLMs. Extensive experiments demonstrate that SSL significantly outperforms existing decoding approaches in mitigating hallucinations, while maintaining transferability across different model architectures with negligible additional time overhead. The code is available at https://github.com/huazhenglin2003/SSL.

1 Introduction

LVLMs (Liu et al., 2023; Dai et al., 2023b; Liu et al., 2024b) have demonstrated impressive capabilities in jointly processing visual and textual modalities, achieving strong performance on tasks such as VQA (Antol et al., 2015) and image captioning (Li et al., 2022). However, LVLMs still suffer from hallucination (Yin et al., 2011), where the generated text does not align with the visual content. This limitation poses significant challenges to their deployment in critical applications, including medical diagnosis(Gu et al., 2024) and autonomous driving(You et al., 2024), where factual consistency and reliability are essential.

To mitigate hallucination in LVLMs, researchers explore various strategies, including refining decoding algorithms (Leng et al., 2024; Huang et al., 2024; Kim et al., 2024), incorporating external

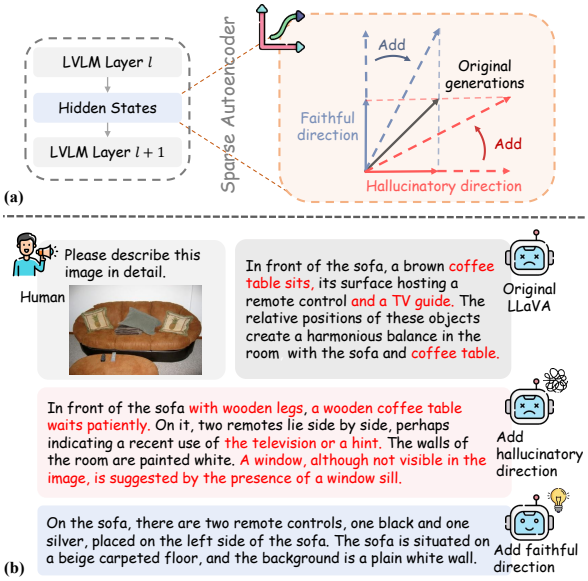


Figure 1: The figure shows, from top to bottom, the original response from the LLaVA-Next-8b, the response after intervention along the hallucinatory direction, and the response after intervention along the faithful direction. These results demonstrate that targeted interventions along faithful directions suppress hallucinatory generations, while perturbations along hallucinatory directions tend to elicit more factually incorrect content.

† Equal contribution.
* Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 10808–10828
November 4-9, 2025

knowledge bases (Qu et al., 2024), and leveraging additional annotated data for model fine-tuning (Park et al., 2024). While promising, these approaches often incur substantial computational cost and time. Recent works (Liu et al., 2025; Jiang et al., 2024; Li et al., 2025) try to explore more efficient alternatives by adjusting LVLMs' internal representations. However, these methods may cause hallucinations to be insufficiently suppressed or lead to excessive interventions that negatively affect normal semantics. Therefore, extracting fine-grained and reliable representations related to hallucinations remains a key challenge in advancing the reliability of LVLMs.

Recently, SAEs have shown success in extracting fine-grained semantic representations—specifically capturing whether the model knows certain entities—of abstract concepts in the field of large language models (LLMs) (Ferrando et al., 2025). SAEs builds on the Linear Representation Hypothesis (Park et al., 2023), which posits that internal model representations can be expressed as sparse combinations of interpretable semantic directions (Tigges et al., 2024; Li et al., 2023a). Building on this foundation, Zhang et al. (2024a) utilize SAEs to interpret LVLM representations and analyze open-set semantics. However, their work does not address hallucination modeling or intervention. Inspired by these works, we extend the application of SAE-based analysis from LLMs to LVLMs, with a particular focus on mitigating hallucinations. Specifically, we leverage the SAE provided by Zhang et al. (2024a) to identify latent directions correlated with hallucinatory semantics as well as those aligned with faithful content, and steer the internal semantic representations of LVLMs, aiming to understand and mitigate hallucinations more precisely and directly. As illustrated in Figure 1, targeted interventions along faithful directions suppress hallucinatory generations, while perturbations along hallucinatory directions tend to elicit more factually incorrect content.

Building on this insight, we propose **S**teering LVLMs via **S**AE **L**atent Directions (SSL), a plug-and-play approach based on SAE-derived latent directions to mitigate hallucinations in LVLMs. During the visual feature merging stage, we inject faithful semantic directions to amplify grounded semantic features and improve image–text consistency. In the subsequent language generation stage, we reduce projection onto hallucinatory semantic directions, thereby reducing the risk of generating factually incorrect content. Remarkably, although the SAE was trained on the LLaVA-Next-8b model, the extracted hallucination and factuality directions generalize seamlessly to other architectures (*e.g.*, LLaVA1.5-7b model (Liu et al., 2024c) and InstructBLIP-7b model (Dai et al., 2023a)). Experimental evaluation on established LVLM hallucination benchmarks shows that SSL outperforms existing decoding approaches, confirming its effectiveness and efficiency in hallucination reduction.

Our main contributions are as follows:

- We leverage SAEs to identify semantic directions that are highly correlated with hallucinatory and faithful content in the representation space of LVLMs.
- We propose SSL, a plug-and-play method that injects factuality semantic directions during visual feature fusion to reinforce grounded content and suppresses hallucination directions during language generation to proactively mitigate hallucinatory outputs.
- Extensive experiments demonstrate that SSL outperforms existing decoding approaches on widely used hallucination benchmarks with negligible time overhead, exhibiting transferability across different architectures.

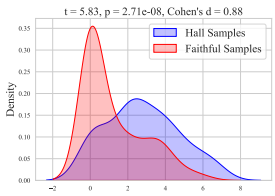
2 Preliminary

LVLM generation. LVLMs take both image and text as input and encode them into a sequence of tokens. During autoregressive generation, the model first concatenates the system tokens X_s, prompt tokens X_t, and visual tokens X_v in a predefined order to form the initial input. At the first generating step $t = 1$, the model predicts the output token based on this initial context. At each subsequent step $t > 1$, the previously generated tokens $X_o^{<t}$ are appended to the end of the initial input, resulting in the current sequence $\left[X_s, X_t, X_v, X_o^{<t}\right]$. The model then generates the next token autoregressively according to the conditional probability distribution, continuing until an end-of-sequence token is produced or a maximum sequence length is reached:

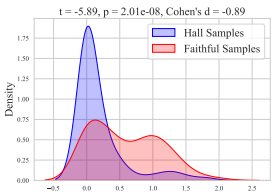
$$y_t = \arg\max p_\theta(y_t \mid X_s, X_t, X_v, X_o^{<t}), \quad (1)$$

where y_t is the token generated at time step t.

Sparse autoencoders. SAEs have been proven to be effective for separating overlapping features (Bricken et al., 2023; Ferrando et al., 2025). In this



(a) Hallucinatory latent KDE distribution



(b) Faithful latent KDE distribution

Figure 2: KDE plots of the selected latent activations for test samples labeled as hallucination and faithfulness. The overlaid plots visualize the distributional differences, with annotated t-statistic, p-value, and Cohen's d effect size indicating the statistical separation between the two groups.

work, we use the SAE provided by Zhang et al. (2024a), which operates on the residual stream $h_l \in \mathbb{R}^d$ from the l-th layer of LVLMs. The SAE projects these representations into a higher-dimensional latent space $z(x) \in \mathbb{R}^{d_{\text{SAE}}}$ and applies a ReLU activation:

$$z(x) = \text{ReLU}(W_{\text{enc}}x + b_{\text{enc}}), \tag{2}$$

where W_{enc} and b_{enc} denote the encoder's weight matrix and bias, respectively. To enforce sparsity, a top-k operation retains only the k largest activations in $z(x)$, zeroing out the rest to obtain the sparse latent vector $z_k(x) = \text{TopK}(z, k)$. The decoder then reconstructs the original representation via a linear combination of the active components:

$$\text{SAE}(x) = W_{\text{dec}}^T z_k(x) + b_{\text{dec}}, \tag{3}$$

where W_{dec} and b_{dec} denote the decoder's weight matrix and bias. During training, the loss function combines the reconstruction error with an auxiliary loss proposed by Gao et al. (2025), aiming to encourage the utilization of meaningful features in the latent representation $z_k(x)$ and to prevent feature inactivity, thereby enhancing the overall expressiveness of the sparse encoding. We refer to each component of $z_k(x)$ as a latent activation, and each row vector of W_{dec} as a latent direction.

Steering with SAE Latents. The SAE reconstructs model representations as a linear combination of latent directions and a bias, effectively approximating the original input. Each latent activation $z_j(x)$ corresponds to a specific decoder direction $d_j = W_{\text{dec}}[j, :]$, enabling targeted adjustment of the representation through activation steering (Turner et al., 2023). This technique allows us to steer the residual stream by modifying the representation as follows:

$$x_{\text{steer}} \leftarrow x + \alpha d_j, \tag{4}$$

where α is a tunable parameter that determines the strength of the intervention.

3 Method

In this work, we introduce SSL, a plug-and-play method for steering LVLMs. Our method consists of two principal components: semantic directions identification and steering LVLMs via SAE latents.

3.1 Semantic Directions Identification

Residual Stream Dataset Construction for Hallucinations and Faithfulness. To investigate whether there exist directions in LVLMs that are highly correlated with hallucinatory and faithful semantics, we randomly sampled $4,000$ image-text pairs from the MSCOCO dataset (Lin et al., 2014). Using the LLaVA-Next-8b model (Liu et al., 2024d) for the image captioning task, we extract the residual stream representations from the 25th layer when the model generated object tokens classified as either hallucinatory or faithful.

Given that a small proportion of object words are tokenized into multiple subword units, we exclude these cases to simplify the analysis. Furthermore, because each sample exhibited an imbalance between the counts of hallucinatory and faithful object terms, we enforce class balance by sampling an equal number of residual vectors from each category per image–text pair. See Appendix A for a description of the process. Finally, we construct a balanced dataset containing $1,784$ samples and divide it into a training set and a test set in a $9:1$ ratio, for direction mining and direction effectiveness validation, respectively.

Semantically Hallucinatory and Faithful Direction Identification via SAE. Inspired by Meng et al. (2022); Ferrando et al. (2025), we leverage SAE to identify latent directions aligned with hallucinatory and faithful semantics. Specifically, each residual stream sample from the training set is passed through the SAE, and we record the activation frequency of each latent activation across hallucinatory samples $\mathcal{X}_{\text{hall}}$ and faithful samples $\mathcal{X}_{\text{faithful}}$. For a given latent activation j, its activation frequencies on hallucinatory samples f_j^{hall} and

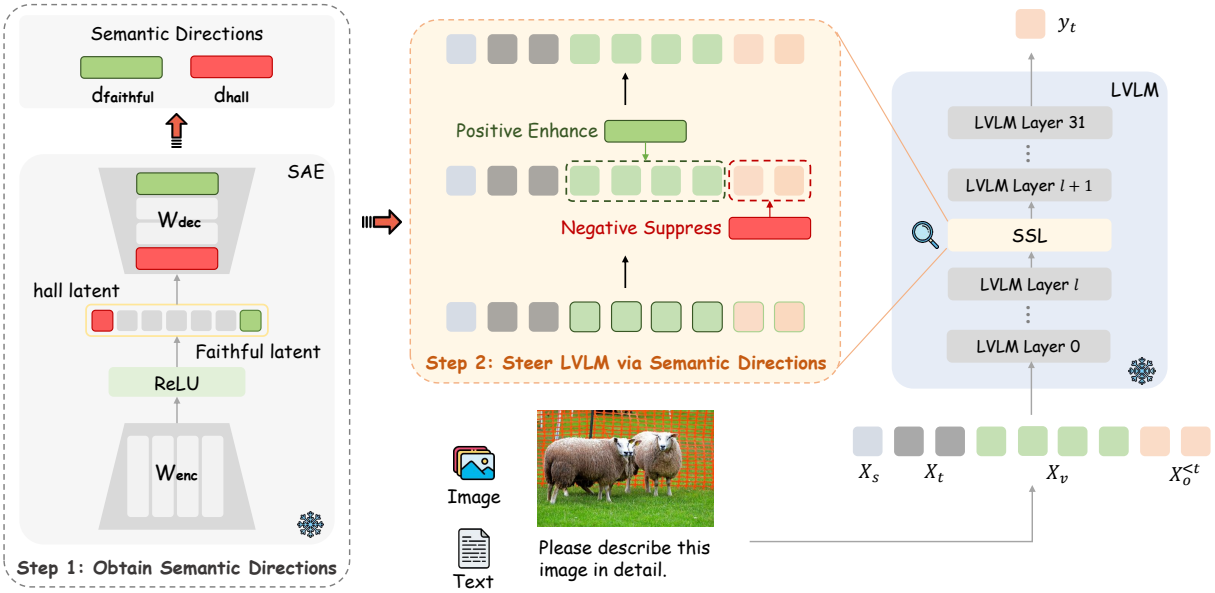


Figure 3: Overview of the proposed SSL approach leveraging SAE to identify semantically aligned directions and mitigate hallucination in LVLMs. We use SAE to identify latent directions within the internal representation space of LVLMs that are associated with hallucinatory and faithful semantics, denoted as d_{hall} and d_{faithful}, respectively. These semantic directions are then used to modulate the residual stream at the l-th layer, steering the generation process toward greater factual consistency.

faithful samples f_j^{faithful} are computed as:

$$f_j^{\text{hall}} = \frac{1}{N_{\text{hall}}} \sum_{x \in \mathcal{X}_{\text{hall}}} \mathbb{I}\left(z_j(x) > 0\right), \\ f_j^{\text{faithful}} = \frac{1}{N_{\text{faithful}}} \sum_{x \in \mathcal{X}_{\text{faithful}}} \mathbb{I}\left(z_j(x) > 0\right), \tag{5}$$

where N_{hall} and N_{faithful} represent the number of hallucinatory and faithful samples, respectively. To quantify the semantic relevance of each latent activation, we compute the difference in activation frequencies as follows:

$$s_j^{\text{hall}} = f_j^{\text{hall}} - f_j^{\text{faithful}}, \\ s_j^{\text{faithful}} = f_j^{\text{faithful}} - f_j^{\text{hall}}. \tag{6}$$

These values reflect the relevance of latent dimension j to hallucinatory and faithful semantics, respectively. Finally, we identify the latent activation with the highest s_j^{hall} as the hallucination semantic direction (hereafter referred to as the *hall* latent), and the one with the highest s_j^{faithful} as the faithful semantic direction (hereafter referred to as the *faithful* latent).

Validation of the Effectiveness of Semantic Directions. We begin by analyzing the distributional differences of the hallucinatory latent and faithful latent activations across both sample types in the test set. These distributions are visualized using kernel density estimation (KDE) plots as shown in Figure 2. We further quantify the separation using independent two-sample t-tests and compute Cohen's d to assess effect sizes. Both latent activations exhibit statistically significant distributional shifts, with substantial effect sizes, confirming their discriminative power.

To probe the semantic alignment of the activations, we further conduct Spearman rank correlation analysis between the activation values of hallucinatory samples and their associated hallucinatory object terms. The hallucinatory latent correlates positively with hallucinatory objects (Spearman's $\rho = 0.42$, $p = 9.95 \times 10^{-9}$), whereas the faithful latent correlates negatively ($\rho = -0.44$, $p = 9.85 \times 10^{-10}$). Given the binary nature of the labels and the symmetry of rank correlation, we infer a positive correlation between the faithful latent and samples.

Finally, to quantitatively evaluate the predictive power of SAE-derived directions in distinguishing hallucinatory from faithful samples, we design a set of classification experiments based on logistic regression. The model takes the latent activation values—either individually or in combination—as input features for a binary hallucination classification task. As illustrated in Figure 4, the results demonstrate that the latent semantic directions extracted

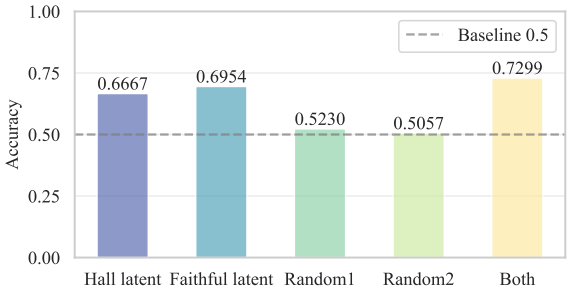


Figure 4: Comparison of classification accuracy using different latent activations. Hall latent and faithful latent correspond to the identified hallucinatory and faithful latent activations, respectively. Random1 denotes a single latent activation randomly selected from the SAE latent space, while Random2 represents a feature combination of two randomly selected latent activations. The dashed line indicates the baseline accuracy of 0.5.

by the SAE are discriminative (see Appendix B for more experimental details). Furthermore, combining the *hall* latent and *faithful* latent as input features yields further performance improvements. A more detailed discussion on clarifying the boundary between hallucinatory and faithful semantic directions can be found in Appendix C.

3.2 Steering LVLMs via SAE Latents

Steering Strategy. A core architectural mechanism involves multi-layer residual connections that progressively model input information. At each layer, semantic representations are passed through residual flows, which can be divided according to the input sequence into four main components: system token, prompt tokens, visual tokens, and output tokens. Among these, visual tokens interact with language tokens to guide the model's understanding of image content. Injecting faithful direction at this position enables the model to increase visual faithfulness. Output tokens represent the model's autoregressive language generations, influenced by both prompts and visual inputs, thereby reflecting the model's semantic behavior. Suppressing hallucinatory directions during this stage helps reduce hallucination tendencies and enhances factual consistency in language generation.

Following the method described in Section 2, we identify two semantic direction vectors: the hallucinatory direction d_{hall} and the faithful direction d_{faithful}. During the visual feature fusion stage, we incorporate d_{faithful} to improve the faithfulness of visual understanding. In the subsequent language generation stage, we suppress activations along d_{hall} to reduce the risk of hallucinatory outputs. Semantic steering at layer l is defined as follows:

$$\begin{aligned} X_{l,\text{v}} &\leftarrow X_{l,\text{v}} + \alpha \cdot d_{\text{faithful}}, \\ X_{l,\text{o}}^{<t} &\leftarrow X_{l,\text{o}}^{<t} - \alpha \cdot d_{\text{hall}}, \end{aligned} \tag{7}$$

where α is a tunable hyperparameter controlling the strength of semantic steering.

Adaptive Steering Parameters (ASP). The setting of the steering strength α plays a crucial role in determining the effectiveness of semantic intervention. Traditional steering approaches often rely on a fixed hyperparameter α to linearly combine the steering vector with the residual representations. However, this fixed strategy can result in unstable or suboptimal performance, as the magnitude of residual vectors can vary across model layers and token positions. In such cases, a change that is too small may fail to induce meaningful guidance, while an excessively large change may cause semantic distortion or instability.

To address this limitation, we propose an adaptive feature steering mechanism, which dynamically adjusts the steering strength based on the norm of the residual vector at each token at a given layer. This approach ensures more stable and context-aware intervention across varying model states. Specifically, the adaptive steering strength α is computed as:

$$\alpha = \gamma \cdot \frac{\|x_{\text{residual}}\|}{\|d_{\text{steer}}\| + \epsilon}, \tag{8}$$

where γ is a scaling factor, x_{residual} denotes the residual vector, d_{steer} is the steering direction, and ϵ is a small constant to avoid numerical instability. An overview of the proposed SSL method is presented in Figure 3. The complete procedure of SSL is provided in Algorithm 1.

4 Experiments

4.1 LVLMs

We conduct experiments on three representative LVLMs: LLaVA-NeXT-8b (Liu et al., 2024d), LLaVA-1.5-7b (Liu et al., 2024c) and InstructBLIP-7b (Dai et al., 2023a). These models share a modular structure comprising an image encoder, a projection module, and a language model. LLaVA-1.5 and LLaVA-NeXT use an MLP to project all image tokens into the LLM's input space, while InstructBLIP employs a Q-Former to select a compact set of informative visual tokens, reducing redundancy.

Algorithm 1: SSL

Input: Scaling factor γ; steering layer l_s; semantic directions d_{hall}, d_{faithful}; residual stream at layer l_s: $[X_{l_s,s}, X_{l_s,t}, X_{l_s,v}, X_{l_s,o}^{<t}]$

if $L = l_s$ **then**
  for *token* x *in residual stream* **do**
    if $x \in X_{l_s,v}$ **then**
      $x \leftarrow x + \gamma \cdot \frac{\|x\|}{\|d_{\text{faithful}}\|+\epsilon} \cdot d_{\text{faithful}}$
    else if $x \in X_{l_s,o}^{<t}$ **then**
      $x \leftarrow x - \gamma \cdot \frac{\|x\|}{\|d_{\text{hall}}\|+\epsilon} \cdot d_{\text{hall}}$
    else
      // System and prompt tokens remain unchanged

Compared to LLaVA-1.5, LLaVA-NeXT upgrades the LLM from 7b to 8b parameters and supports higher-resolution inputs for visual understanding.

4.2 Benchmarks

CHAIR. We evaluate object hallucination using the Caption Hallucination Assessment with Image Relevance (CHAIR) metric (Rohrbach et al., 2018), which compares generated image captions against ground-truth annotations to detect hallucinatory objects mentioned in the captions but absent from the image. CHAIR includes two metrics at both captions level (CHAIR_S) and object level (CHAIR_I):

$$\text{CHAIR}_S = \frac{|\{\text{captions w/ hallucinatory objects}\}|}{|\{\text{total captions}\}|},$$
$$\text{CHAIR}_I = \frac{|\{\text{hallucinatory objects}\}|}{|\{\text{total mentioned objects}\}|}. \quad (9)$$

We randomly sample 500 images from the COCO 2014 validation set (Lin et al., 2014) and conduct five runs with different random seeds. For all LVLMs, captions are generated using the prompt "Please describe this image in detail." We report the mean and standard deviation for each metric.

POPE. We further evaluate object hallucination using the POPE benchmark (Li et al., 2023b), a question-answering dataset designed to assess the factual consistency of generated image descriptions. POPE contains 500 images from the MSCOCO dataset (Lin et al., 2014), each paired with binary questions of the form: "Is there a <object> in the image?" The dataset comprises three subsets—random, popular, and adversarial—which differ in their object sampling strategies. Model performance is measured using standard classification metrics: Accuracy, Precision, Recall, and F1 score. To provide an overall assessment, we report the average results across all three subsets.

LLaVA-Bench. We evaluate LVLM performance using the LLaVA-Bench (In-the-Wild) benchmark (Liu et al., 2024c), a comprehensive set designed to assess models across diverse and challenging visual scenarios. The benchmark includes 24 images from varied real-world contexts, such as indoor scenes, outdoor environments, and internet memes, paired with 60 carefully curated questions spanning open-ended QA, fine-grained descriptions, and complex reasoning. We prompt the GPT-4o model to evaluate the LVLMs' outputs along two dimensions: factual accuracy and response detail.

4.3 Baselines

We compare the performance of base LVLMs using greedy decoding and beam search decoding. Additionally, we also conduct a comparison between SSL and the popular training-free approaches that require neither external data nor auxiliary models. Specifically, DoLa (Chuang et al., 2024) derives the next-token distribution by contrasting logits from later and earlier layers; VCD (Leng et al., 2024) employs contrastive learning by comparing the output distributions generated from original and perturbed images; OPERA (Huang et al., 2024) enhances generation quality by alleviating excessive reliance on previously generated tokens during beam search; and CODE (Kim et al., 2024) enhances vision-language alignment by using self-generated captions as internal references.

4.4 Implementation Details

We set γ to 0.6, 0.8, and 0.1 for LLaVA-NeXT, LLaVA-1.5, and InstructBLIP, respectively, to balance effective mitigation of hallucination with minimizing the invasiveness of state interventions. SSL is applied at 16th layer for LLaVA-NeXT, 31th layer for LLaVA-1.5 and 8th layer for InstructBLIP. We faithfully replicate all baseline methods, implementing them based on their open-source codebases and configuring them according to the hyperparameters reported in the original papers. All experimental results are obtained under consistent base model, prompt, and generation parameter settings to ensure a fair comparison. For all methods involving beam search, we set the max_new_token

	LLaVA-NeXT-8b			LLaVA-1.5-7b			InstructBLIP-7b		
	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	*Avg.Len*	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	*Avg.Len*	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	*Avg.Len*
Greedy	$29.60_{\pm 0.89}$	$8.03_{\pm 0.41}$	165.61	$49.44_{\pm 1.57}$	$14.19_{\pm 0.76}$	82.97	$\underline{45.44}_{\pm 2.43}$	$\underline{13.07}_{\pm 0.71}$	92.11
Beam	$\underline{27.20}_{\pm 1.19}$	$\underline{7.20}_{\pm 0.39}$	174.17	$53.60_{\pm 2.39}$	$15.47_{\pm 0.45}$	87.38	$48.68_{\pm 1.65}$	$13.59_{\pm 0.43}$	95.92
DoLa	$29.04_{\pm 1.08}$	$7.86_{\pm 0.24}$	166.14	$50.64_{\pm 2.33}$	$14.51_{\pm 0.88}$	82.32	$46.12_{\pm 1.85}$	$13.09_{\pm 0.90}$	91.80
VCD	$31.36_{\pm 1.99}$	$8.40_{\pm 0.79}$	165.43	$51.68_{\pm 1.85}$	$15.29_{\pm 0.83}$	83.03	$50.84_{\pm 2.41}$	$14.51_{\pm 0.97}$	91.44
OPERA	-	-	-	$\underline{44.04}_{\pm 0.94}$	$\underline{13.23}_{\pm 0.46}$	75.79	$45.88_{\pm 2.31}$	$13.15_{\pm 0.87}$	93.51
CODE	$30.76_{\pm 0.92}$	$8.09_{\pm 0.42}$	158.07	$47.72_{\pm 0.79}$	$14.13_{\pm 0.56}$	78.43	$50.88_{\pm 2.05}$	$14.21_{\pm 0.92}$	89.62
SSL	$\mathbf{26.36}_{\pm 1.94}$	$\mathbf{6.32}_{\pm 0.63}$	163.86	$\mathbf{40.88}_{\pm 2.11}$	$\mathbf{12.30}_{\pm 1.18}$	82.37	$\mathbf{44.04}_{\pm 3.91}$	$\mathbf{12.64}_{\pm 1.38}$	100.26

Table 1: CHAIR results on MSCOCO dataset averaged over 5 random seeds. The best and second-best results are indicated in **bold** and underlined, respectively. *Avg.Len* represents the average length of the generated descriptions.

	LLaVA-NeXT-8b F1 score ↑	LLaVA-1.5-7b F1 score ↑	InstructBLIP-7b F1 score ↑
Greedy	89.10	84.99	<u>85.37</u>
Beam	89.30	85.31	84.41
DoLa	<u>89.49</u>	85.08	85.22
VCD	88.91	84.42	84.68
OPERA	-	<u>85.46</u>	84.42
CODE	88.93	84.64	84.81
SSL	**89.66**	**85.47**	**85.56**

Table 2: POPE results averaged over popular, adversarial, and random splits. The best and second-best results are indicated in **bold** and underlined, respectively.

	Accuracy ↑	Detailedness ↑
LLaVA-NeXT	6.2891	**6.0278**
LLaVA-NeXT w/ SSL	**6.3671**	5.2667
LLaVA-1.5	**5.3333**	**4.7000**
LLaVA-1.5 w/ SSL	4.7167	4.5667
InstructBLIP	5.5056	4.1111
InstructBLIP w/ SSL	**5.5823**	**4.2711**

Table 3: Evaluation results on the LLaVA-Bench (In-the-Wild) benchmark based on prompting GPT-4o.

to 512 and the beam size to 5.

5 Results

CHAIR. Table 1 reports the performance of SSL on the CHAIR benchmark compared to all baseline approaches. Due to excessive memory requirements, OPERA fails to produce results on LLaVA-NeXT-8b. Notably, although the SAE from Zhang et al. (2024a) were trained on LLaVA-NeXT-8b, the identified semantic directions generalize well across different model architectures. SSL consistently outperforms all baselines across all three LVLMs, while only incurring a marginal trade-off in caption length or descriptive richness.

POPE. As shown in Table 2, applying SSL to LVLMs with different architectures consistently improves performance on the POPE benchmark. This demonstrates the robustness of SSL in enhancing models across a spectrum of capabilities and further validates the generalizability of the semantic directions captured by our approach.

LLaVA-Bench. Table 3 presents the evaluation results on LLaVA-Bench using GPT-4o. The results highlight the effectiveness of SSL in enhancing model accuracy across highly diverse and challenging tasks.

5.1 Ablation Study

Effectiveness of ASP. To validate the necessity of adaptively adjusting the steering strength, we conduct an ablation study by replacing ASP with a fixed α steering parameter as shown in Equation 4 equal to the γ value. As shown in Table 4, removing the adaptive strategy ASP results in consistent performance drops across all three model architectures, highlighting the importance of ASP in effectively mitigating hallucinations.

Layer Selection Ablation. We conduct an ablation study on LLaVA-NeXT-8b to examine the impact of applying guidance at different layers. Figure 5 shows that the choice of guidance layer significantly affects model performance, setting γ to 0.8. For LLaVA-NeXT-8b, applying SSL at middle layers yields more effective mitigation of hallucinations, with layer 15 achieving the best performance. For further analysis on steering layer selection and scaling factor γ settings across different models, please refer to Appendix D and E.

5.2 Further Analysis

Analysis of Reverse-SSL for Inducing Hallucinations in LVLMs. To further validate the effectiveness of the semantic directions we identified, we compare the CHAIR benchmark across

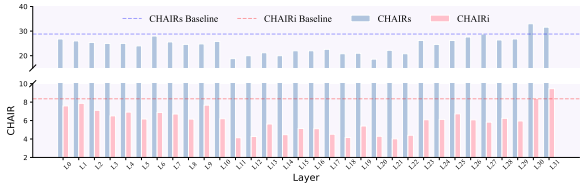


Figure 5: Results of SSL applied across different layers.

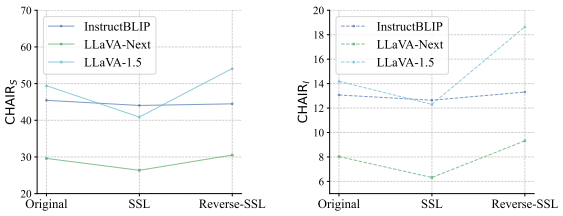
	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$
LLaVA-NeXT w/ ASP	**26.36**	**6.32**
LLaVA-NeXT w/ fixed α	28.40	7.79
LLaVA-1.5 w/ ASP	**40.88**	**12.30**
LLaVA-1.5 w/ fixed α	47.21	13.12
InstructBLIP w/ ASP	**44.04**	**12.64**
InstructBLIP w/ fixed α	45.76	13.12

Table 4: Ablation study on the necessity of adaptively adjusting the steering parameters.

three model architectures under three settings: original model state, SSL, and Reverse-SSL (see Appendix F for details on Reverse-SSL). As shown in Figure 6, across all three model architectures, applying Reverse-SSL significantly increases hallucinations, while standard SSL guidance consistently reduces hallucinations. To further clearly illustrate the effectiveness of SSL, we present several qualitative examples in Appendix G.

Additional Time Analysis. During each generation step, SSL dynamically adjusts the steering strength through a single scaling and weighting operation, introducing negligible computational overhead. Compared to the overall generation process, the additional latency introduced by SSL is minimal. A comparison of inference time between SSL and other baselines is shown in Figure 7.

Effect of Model Size on Method Performance. To further analyze the impact of model size on our method, we conducted additional experiments on the 7B, 8B, and 11B models, evaluating SSL performance on both the CHAIR and POPE benchmarks, with results reported in Table 5 and Table 6. Specifically, for the 11B model, we use Llama-3.2-11b-Vision-Instruct released by Meta Llama, where SSL is applied with $\gamma = 0.4$ at the 32th layer. The results demonstrate that SSL effectively mitigates hallucinations across models of different sizes and consistently outperforms other baseline methods.



(a) $CHAIR_S$ across models (b) $CHAIR_I$ across models

Figure 6: CHAIR evaluation across three multi-modal models—InstructBLIP-7b, LLaVA-NeXT-8b, and LLaVA-1.5-7b—under three generation settings: original, SSL, and Reverse-SSL.

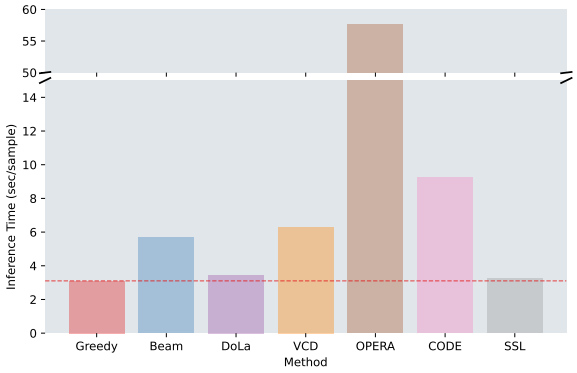


Figure 7: Comparison of inference time for different methods measured using identical hardware.

6 Related Works

LVLMs refer to the phenomenon where the generated textual content is inconsistent with the visual input. This issue arises from various factors, such as dataset bias, insufficient visual perception by the encoder, and misalignment across modalities (Liu et al., 2024a). While prior studies have proposed diverse strategies to mitigate hallucination, the internal mechanisms within LVLMs that give rise to such inconsistencies remain largely underexplored.

Liu et al. (2025) enhance the stability of visual representations by steering latent features during generation, preventing premature degradation. Jiang et al. (2024) remove hallucination-related feature components through linear orthogonalization by projecting the internal image representations of vision models into the language space, thereby purifying the input and reducing hallucinations. Li et al. (2025) uncover phenomena such as early activation and progressive loss of visual information in LVLMs, and propose injecting continuous visual streams during inference to compensate for these effects, significantly reducing hallucinations.

	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$
LLaVA-1.5-7b	49.44	14.19
LLaVA-1.5-7b w/ SSL	**40.88**	**12.30**
LLaVA-Next-8b	29.60	8.03
LLaVA-Next-8b w/ SSL	**26.36**	**6.32**
Llama-3.2-11b-Vision-Instruct	31.40	7.40
Llama-3.2-11b-Vision-Instruct w/ SSL	**28.92**	**6.69**

Table 5: CHAIR results on MSCOCO dataset averaged over 5 random seeds. The best are indicated in **bold**.

	Accuracy ↑	F1 score ↑
LLaVA-1.5-7b	83.86	84.99
LLaVA-1.5-7b w/ SSL	**84.55**	**85.47**
LLaVA-Next-8b	89.10	89.10
LLaVA-Next-8b w/ SSL	**89.19**	**89.66**
Llama-3.2-11b-Vision-Instruct	87.26	87.80
Llama-3.2-11b-Vision-Instruct w/ SSL	**87.48**	**88.07**

Table 6: POPE results averaged over popular, adversarial, and random splits. The best are indicated in **bold**.

Unlike previous methods, our work directly identifies hallucinatory and faithful semantic directions using SAEs. We then dynamically adjust these directions during visual-linguistic fusion and generation to proactively reduce hallucination outputs.

Furthermore, our approach contributes to the practical interpretability of SAEs in LVLMs, demonstrating their potential for understanding and controlling internal semantic representations.

7 Conclusion

This work explores the relationship between the hallucination in LVLMs and their internal latent representations. We construct a residual stream dataset for hallucinatory and faithful object tokens, and use SAE to extract the semantic directions corresponding to hallucination and Factuality. Based on this insight, we propose SSL, a plug-and-play method that amplifies true semantics while suppressing potential hallucinations. Extensive experiments demonstrate that SSL outperforms existing methods. Furthermore, although the SAE was trained on LLaVA-Next, the semantic directions it extracted generalize well across different model architectures, further showcasing the potential of SAE in understanding and controlling the internal semantic representations of models.

Limitations

Currently, the only fully open-source multi-modal SAE is provided by LLM-Labs, trained on the 25th layer of the LLaVA-Next 8b model. As a result, our study does not include a comparison of SAEs trained on other model architectures across different multimodal models. Future work can focus on training multi-modal SAEs on various architectures to investigate whether the findings from this study generalize across different models.

Acknowledgements

This work was supported by National Key R&D Program of China under Grant No.2024YFC3015501, also supported by the National Natural Science Foundation of China under Grant 62576094, and Grant U24A20322.

Ethical Consideration

In our proposed SSL method, positive steering significantly reduces hallucinations in LVLMs, while negative steering increases them. Positive steering not only improves model performance but also aligns with ethical principles such as safety and reliability. In contrast, negative guidance may lead to more hallucinations and generate false information. Therefore, such mechanisms should be applied with caution, supported by thorough validation and human oversight.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. In *Transformer Circuits Thread*.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2025. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision – ECCV 2024*, pages 19–35, Cham. Springer Nature Switzerland.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023a. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023b. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv preprint*, abs/2305.06500.

Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2025. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. 2024. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. *ArXiv preprint*, abs/2410.04514.

Zishan Gu, Changchang Yin, Fenglin Liu, and Ping Zhang. 2024. Medvh: Towards systematic evaluation of hallucination for large vision language models in the medical context. *ArXiv preprint*, abs/2407.02730.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. 2024. Cracking the code of hallucination in lvlms with vision-aware head divergence. *ArXiv preprint*, abs/2412.13949.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427, Los Alamitos, CA, USA. IEEE Computer Society.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. 2024. Interpreting and editing vision-language representations to mitigate hallucinations. *ArXiv preprint*, abs/2410.02762.

Junho Kim, Hyunjun Kim, KIM YEONJU, and Yong Man Ro. 2024. CODE: Contrasting self-generated description to combat hallucination in large multi-modal models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882, Los Alamitos, CA, USA. IEEE Computer Society.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. 2025. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *ArXiv preprint*, abs/2502.03628.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *ArXiv preprint*, abs/2402.00253.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024d. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485.

Sheng Liu, Haotian Ye, and James Zou. 2025. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Dongmin Park, Zhaofang Qian, Guangxing Han, and Ser-Nam Lim. 2024. Mitigating dialogue hallucination for large vision language models via adversarial instruction tuning. *ArXiv preprint*, abs/2403.10492.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*.

Xiaoye Qu, Qiyuan Chen, Wei Wei, Jishuo Sun, and Jianfeng Dong. 2024. Alleviating hallucination in large vision-language models with active retrieval augmentation. *ArXiv preprint*, abs/2408.00555.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Curt Tigges, Oskar J. Hollinsworth, Atticus Geiger, and Neel Nanda. 2024. Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 58–87, Miami, Florida, US. Association for Computational Linguistics.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *ArXiv preprint*, abs/2308.10248.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2011. A survey on multimodal large language models. *ArXiv preprint*, abs/11/12.

Junwei You, Haotian Shi, Zhuoyu Jiang, Zilin Huang, Rui Gan, Keshu Wu, Xi Cheng, Xiaopeng Li, and Bin Ran. 2024. V2x-vlm: End-to-end v2x cooperative autonomous driving through large vision-language models. *ArXiv preprint*, abs/2408.09251.

Kaichen Zhang, Yifei Shen, Bo Li, and Ziwei Liu. 2024a. Large multi-modal models can interpret features in large multi-modal models. *ArXiv preprint*, abs/2411.14982.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2024b. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *ArXiv preprint*, abs/2411.09968.

A Construction of Residual Stream Dataset for Hallucinatory and Faithful Object Words

Figure 8 illustrates the construction process of the residual stream dataset. We begin by randomly sampling 4,000 image-text pairs from the MSCOCO dataset and extracting residual stream vectors from the 25th layer of the LLaVA-Next-8b model corresponding to object words identified as either hallucinatory or faithful during inference. It is worth noting that some object words are tokenized into multiple subword tokens by the model's tokenizer. Given the relatively low frequency of such cases and to facilitate consistent statistical analysis, we exclude these incomplete subword instances from our dataset. For example, the word "backpack" may be split into two tokens—"back" and "pack"—by the tokenizer. Since these subtokens do not independently convey the complete semantic meaning of the original word, they are omitted from further analysis.

To ensure class balance, we extract an equal number of residual stream vectors for hallucinatory and faithful object words from each sample. This process results in a balanced dataset comprising 1,784 instances, with a 1:1 ratio of positive (hallucinatory) and negative (faithful) samples. We further divide the dataset into training and test sets using a 9:1 split while maintaining the class distribution in both subsets. The training set is used to identify semantic directions closely associated with hallucination, and the test set is employed to evaluate the generalizability and discriminative power of the extracted semantic features. Using sparse autoencoder analysis, we find that the direction corresponding to latent activation index 36992 is highly correlated with hallucination, whereas index 47230 aligns closely with faithful outputs.

B A Set of Classification Experiments Based on Logistic Regression

To quantitatively evaluate the discriminative power of the latent activation directions extracted by the SAE in distinguishing hallucinatory from faithful samples, we conduct a series of classification experiments based on logistic regression. Specifically, for the i-th sample, let $z(x_i) \in \mathbb{R}^{d_{\text{SAE}}}$ denote

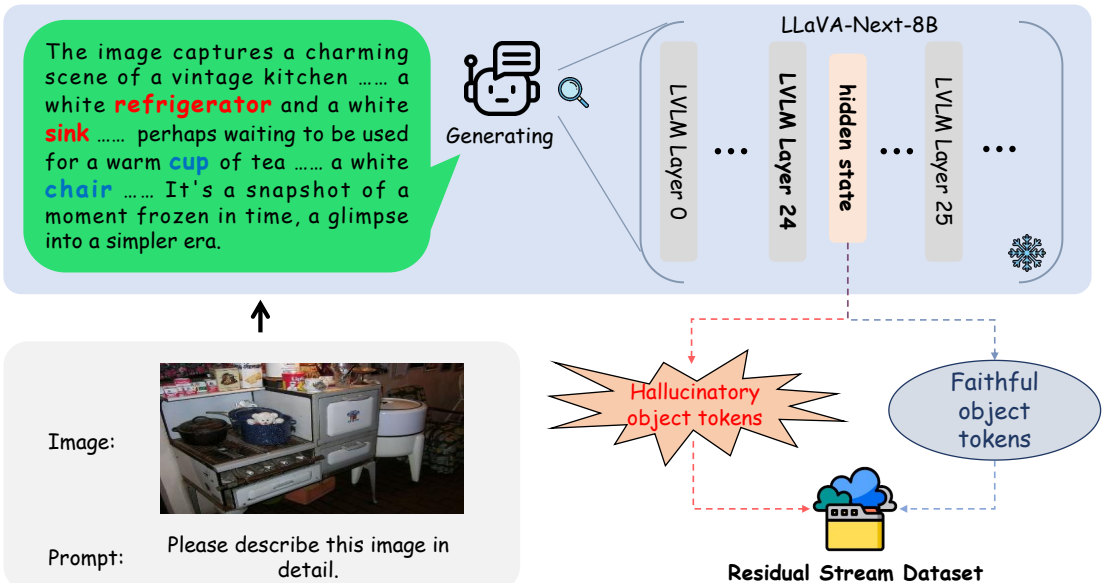


Figure 8: The process of construction of residual stream dataset for hallucinatory and faithful object words.

the SAE latent representation. From a total of N samples, we extract individual latent dimensions indexed by j, denoted as $z_j(x_i)$, and construct five types of input features:

Hall latent: The dimension $hall$ with the highest correlation to hallucinated object words is selected, forming a one-dimensional feature:

$$X_{\text{hall}} = [z_{\text{hall}}(x_1), \ldots, z_{\text{hall}}(x_N)]^T \in \mathbb{R}^{N\times 1}. \tag{10}$$

Faithful latent: The dimension $faithful$ most correlated with faithful object words is selected, forming:

$$X_{\text{faithful}} = [z_{\text{faithful}}(x_1), \ldots, z_{\text{faithful}}(x_N)]^T \in \mathbb{R}^{N\times 1}. \tag{11}$$

Random 1: A single latent dimension $r_1 \sim \mathcal{U}\{0, \ldots, d_{\text{SAE}} - 1\}$ is randomly sampled to form a one-dimensional baseline feature $X_{r_1} = \mathbb{R}^{N\times 1}$.

Random 2: Two latent dimensions (r_1, r_2) are randomly sampled to construct a two-dimensional feature $X_{r_2} \in \mathbb{R}^{N\times 2}$.

The corresponding label vector is $y = [y^{(1)}, y^{(2)}, \ldots, y^{(N)}]^T$, where $y^{(i)} \in \{0, 1\}$, with 1 indicating a hallucinatory sample and 0 a faithful sample. All input features are standardized before being fed into the logistic regression model. The model is trained on the training set and evaluated on the held-out test set using classification accuracy and confusion matrices as evaluation metrics.

As shown in Figure 9, the confusion matrices for the three main feature groups (*Hall* latent, *Faithful* latent, Both). The results show that the model achieves balanced performance across both positive and negative classes, with no noticeable prediction bias. Figure 4, the performance of Random1 and Random2 approximates the random baseline ($\approx$ 0.5), while using Hall latent and faithful latent individually yields classification accuracies of 66.67% and 69.54%, respectively. Combining the two features (Both) further improves performance to 72.99%.

	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	*Avg.Len*
LLaVA-NeXT	29.60	8.03	165.61
w/ SSL $\gamma = 0.2$	28.36	7.38	163.20
w/ SSL $\gamma = 0.4$	27.92	6.69	158.01
w/ SSL $\gamma = 0.6$	26.36	6.32	163.86
w/ SSL $\gamma = 0.8$	22.28	5.14	185.80
w/ SSL $\gamma = 1.0$	13.28	4.06	163.62
w/ SSL $\gamma = 1.5$	2.56	0.59	53.80

Table 7: Ablation study on the scaling factor of LLaVA-Next.

These findings suggest that the latent representations extracted by the SAE encode semantically discriminative signals for hallucination detection. Furthermore, combining hallucination and factuality-related latent directions provides complementary information that enhances classification performance.

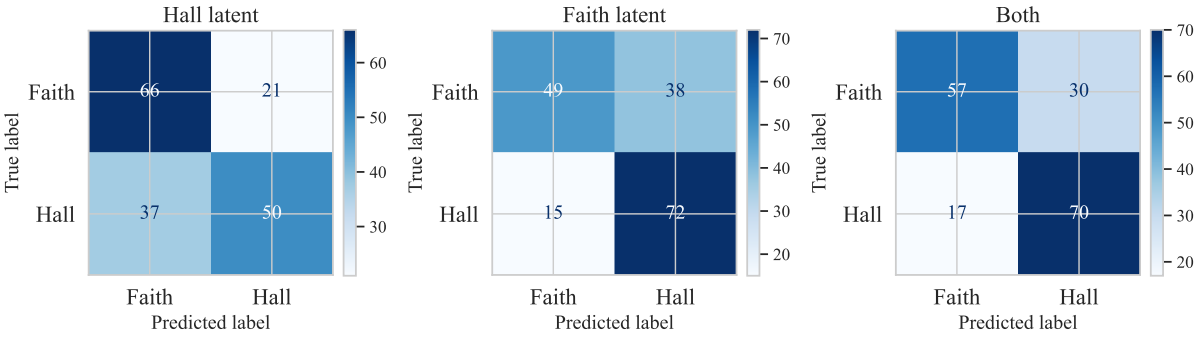


Figure 9: The confusion matrices for three main feature groups.

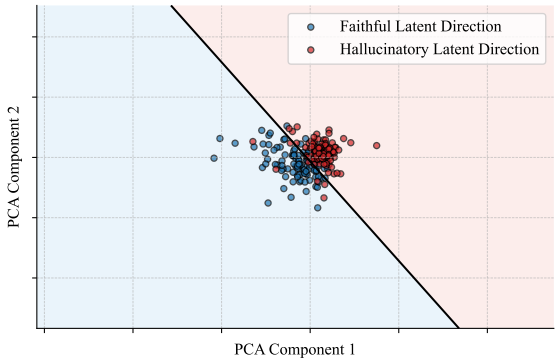


Figure 10: Visualization of the linear SVM decision boundary in the 2D PCA-reduced latent space. Blue points represent faithful latent directions, while red points represent hallucinatory latent directions. The shaded regions indicate the SVM classification regions, and the black line shows the decision boundary.

C Boundary Between Hallucinatory and Faithful Semantic Directions

First, we computed the activation frequency of each latent direction across hallucinatory and faithful samples. If a direction was significantly more active in hallucinatory samples, it was labeled as hallucinatory; conversely, it was labeled as faithful. Based on this criterion, we selected the top 128 directions with the largest activation differences between the two categories (excluding dead latents) for further analysis.

We then used a Support Vector Machine (SVM) classifier in the high-dimensional space to identify a separating decision boundary between the two categories, and applied Principal Component Analysis (PCA) to project the features into two dimensions for intuitive visualization. As shown in Figure 10, although hallucinatory and faithful semantic directions appear as relatively tight clusters in the 2D projection space, the linear decision boundary learned by the SVM still effectively separates them. This apparent overlap is mainly due to PCA prioritizing the preservation of global variance rather than discriminative features, with much of the key class-distinguishing information compressed into higher-dimensional subspaces. Nevertheless, the presence of a clear linear boundary even after dimensionality reduction indicates that hallucinatory and faithful semantic directions are inherently separable, thereby validating the existence of a semantic boundary between the two groups.

D Choice of Steering Layer

Figures 11, 12 and 13 present the results of ablation studies investigating the effect of introducing SSL at individual layers of the LLaVA-Next ($\gamma = 0.6$), LLaVA-1.5 ($\gamma = 0.8$) and InstructBLIP ($\gamma = 0.1$) models, respectively. For LLaVA-Next, we observe that applying SSL at the middle layers more effectively mitigates hallucinations, consistent with the results shown in Figure 5. For LLaVA-1.5, we observe that applying SSL at either the layer1 or deeper layers consistently mitigates hallucination. This observation aligns closely with findings reported by Zhang et al. (2024b), He et al. (2024) and Chen et al. (2025), who also found that layer1 or deeper layers interventions can significantly reduce hallucination in LLaVA-1.5. In contrast, for InstructBLIP, introducing SSL at shallow layers yields more substantial improvements, while deeper layer interventions contribute less noticeably to performance. We hypothesize that this is attributable to architectural and training differences in InstructBLIP, specifically, its shallow layers may already perform substantial cross-modal alignment early in the pipeline, making early-stage semantic guidance more impactful on overall generation quality. A deeper analysis of the layer-specific mechanisms in different multimodal architectures

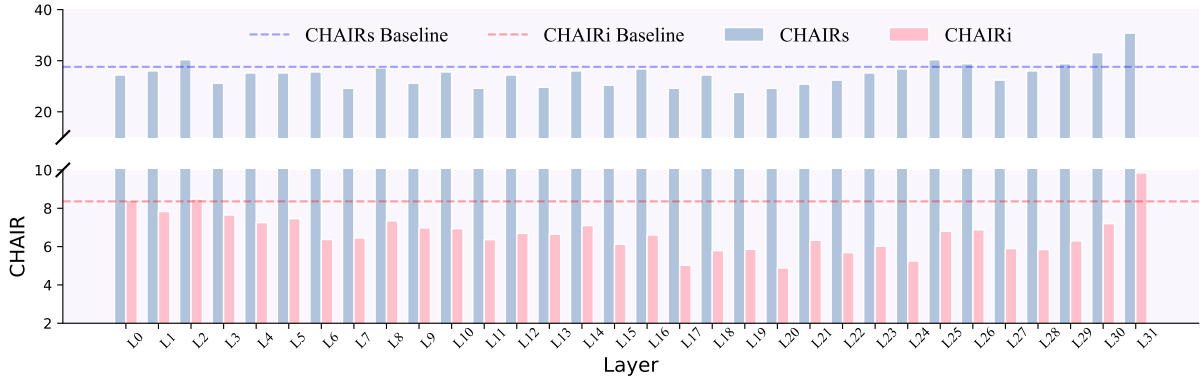


Figure 11: Ablation study on steering a specific layer of LLaVA-Next.

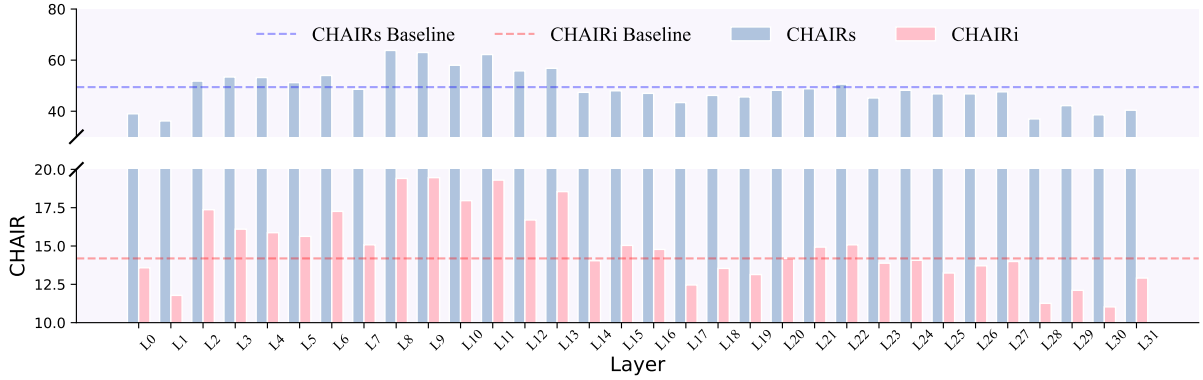


Figure 12: Ablation study on steering a specific layer of LLaVA-1.5.

Algorithm 2: Reverse-SSL

Input: Scaling factor γ; steering layer l_s; semantic directions d_{hall}, d_{faithful}; residual stream at layer l_s: $[X_{l_s,s}, X_{l_s,t}, X_{l_s,v}, X_{l_s,o}^{<t}]$

if $L = l_s$ **then**
  for *token* x *in residual stream* **do**
    if $x \in X_{l_s,v}$ **then**
      $x \leftarrow x - \gamma \cdot \frac{\|x\|}{\|d_{\text{faithful}}\|+\epsilon} \cdot d_{\text{faithful}}$
    else if $x \in X_{l_s,o}^{<t}$ **then**
      $x \leftarrow x + \gamma \cdot \frac{\|x\|}{\|d_{\text{hall}}\|+\epsilon} \cdot d_{\text{hall}}$
    else
      `// System and prompt tokens remain unchanged`

is left for future work.

E Choice of Scaling Factor

Tables 7, 8, and 9 report the ablation results on the effect of the scaling factor γ in the SSL. For the LLaVA series of models, setting γ to 0.6 or 0.8 effectively reduces hallucinations, indicating that moderate levels of semantic intervention are beneficial. However, when $\gamma > 1.0$, the supervision becomes overly aggressive, disrupting the model's behavior and leading to abnormal hallucination metrics. In contrast, for InstructBLIP, larger values similarly result in performance degradation. These findings highlight the importance of carefully calibrating the intensity of semantic guidance to balance model control and generation quality across different architectures.

	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	*Avg.Len*
LLaVA-1.5	49.44	14.19	82.97
w/ SSL $\gamma = 0.2$	48.16	14.10	83.45
w/ SSL $\gamma = 0.4$	47.16	14.04	83.24
w/ SSL $\gamma = 0.6$	45.96	13.36	83.63
w/ SSL $\gamma = 0.8$	40.88	12.30	82.37
w/ SSL $\gamma = 1.0$	33.80	10.18	86.02
w/ SSL $\gamma = 1.5$	17.16	7.66	275.16

Table 8: Ablation study on the scaling factor of LLaVA-1.5.

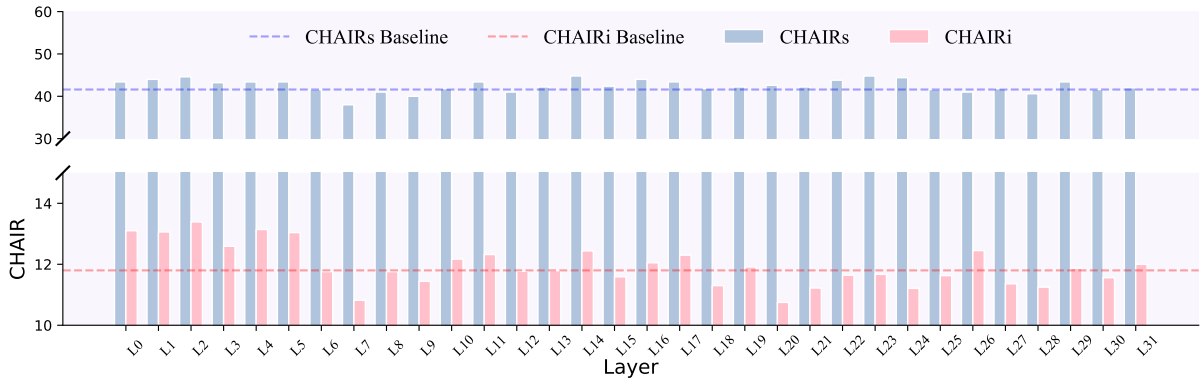


Figure 13: Ablation study on steering a specific layer of InstructBLIP.

	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	*Avg.Len*
InstructBLIP	45.44	13.07	92.11
w/ SSL $\gamma = 0.1$	44.04	12.64	102.86
w/ SSL $\gamma = 0.2$	41.92	12.10	100.26
w/ SSL $\gamma = 0.3$	38.40	11.67	105.38
w/ SSL $\gamma = 0.4$	37.21	11.45	119.39
w/ SSL $\gamma = 0.5$	8.61	10.55	44.49
w/ SSL $\gamma = 0.6$	3.42	14.02	52.14

Table 9: Ablation study on the scaling factor of InstructBLIP.

F Reverse-SSL Approach

To further validate the efficacy of the identified semantic directions, we extend the standard SSL by introducing Reverse Steering LVLMs via SAEs Latents (Reverse-SSL), an approach that deliberately induces the model to generate more hallucinations. Specifically, at each residual flow layer, we divide the input sequence into four contiguous segments: system tokens, prompt tokens, visual tokens, and output tokens. During the visual-token stage, we inject a specific reverse-direction vector that deliberately shifts the visual features away from the true image semantics. At the onset of autoregressive language generation, we inject the specific reverse-direction to amplify the previously distorted visual signal, thereby biasing subsequent text outputs toward content that is either factually incorrect or substantially divergent from the original prompt. The complete algorithmic procedure for Reverse-SSL is presented in Algorithm 2.

G More Qualitative Results

Figures 14, 15, and 16 present additional qualitative examples on the LLaVA-Next, LLaVA-1.5, and InstructBLIP models, respectively, to demonstrate the effectiveness of our proposed SSL approach in mitigating hallucinated objects. With the integration of SSL, the generated descriptions by LVLMs exhibit improved fidelity to the visual content while maintaining the richness and informativeness of the language output.

H Details on the GPT-4o Evaluation

To evaluate the performance of LVLMs on the LLaVA-Bench benchmark, we adopt GPT-4o as the reference evaluator. Following the template provided in Table 10 of Gong et al. (2024), each evaluation instance includes the original image, the base output of the LVLM, and its SSL-enhanced counterpart. The evaluation focuses on both the accuracy and fineness of the generated responses. To mitigate potential biases caused by output order, we randomly swap the positions of the two outputs with a probability of 0.5 before each evaluation. Each sample is evaluated four times to compute an average score. Figures 17, 18, and 19 illustrate representative evaluation examples based on three model architectures: LLaVA-Next, LLaVA-1.5, and InstructBLIP, respectively.

I License

The open-source data utilized in this work was employed exclusively for academic research, consistent with the original intended usage. All the used intellectual artifacts' license allows for academic usage.

GPT-4o Prompt

You are required to score the performance of two AI assistants in describing a given image. You should pay extra attention to the hallucination, which refers to the part of descriptions that are inconsistent with the image content, such as claiming the existence of something not present in the image or describing incorrectly in terms of the counts, positions, or colors of objects in the image. Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria:
1. Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinations should be given higher scores.
2. Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not count as necessary details.

Please output the scores for each criterion, containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.
[Assistant 1]
{}
[End of Assistant 1]

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

Output format:
Accuracy:
Reason:
Detailedness:
Reason:

Table 10: The prompt used for GPT-4o evaluation.

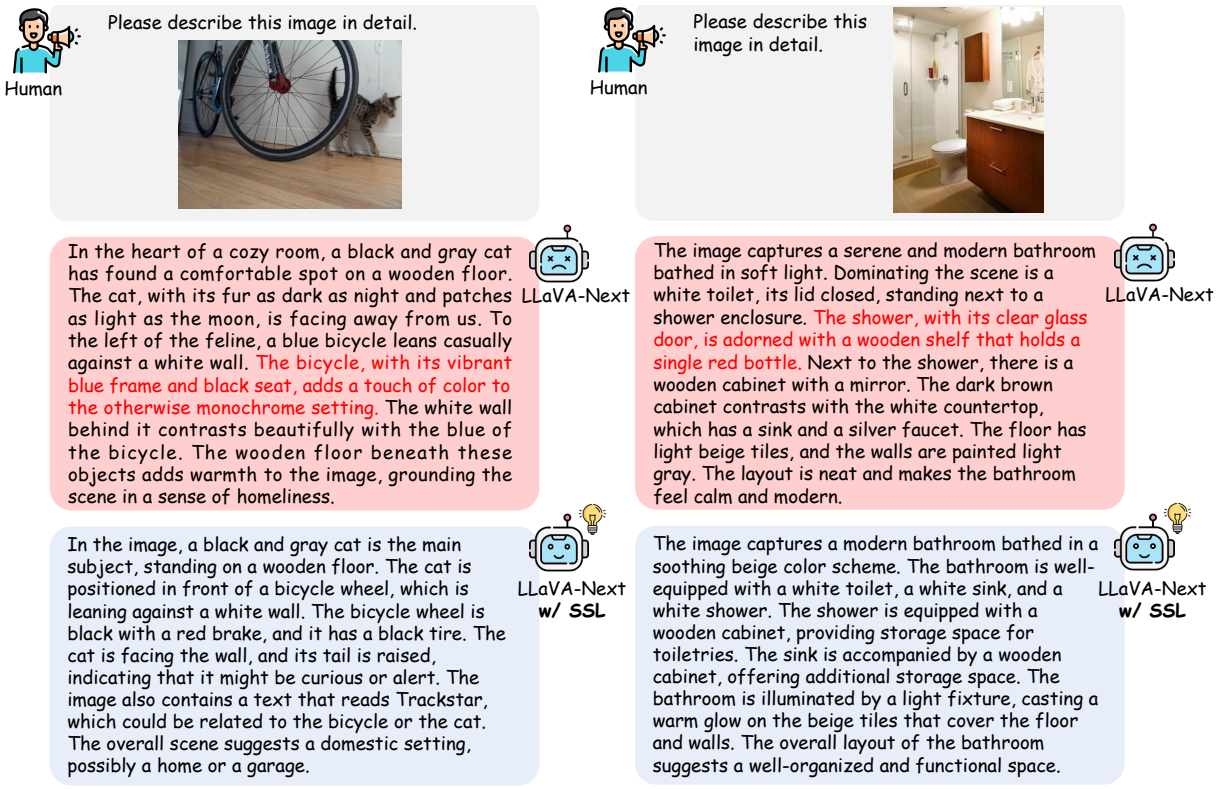


Figure 14: More examples of SSL in reducing hallucinated content in LLaVA-Next.

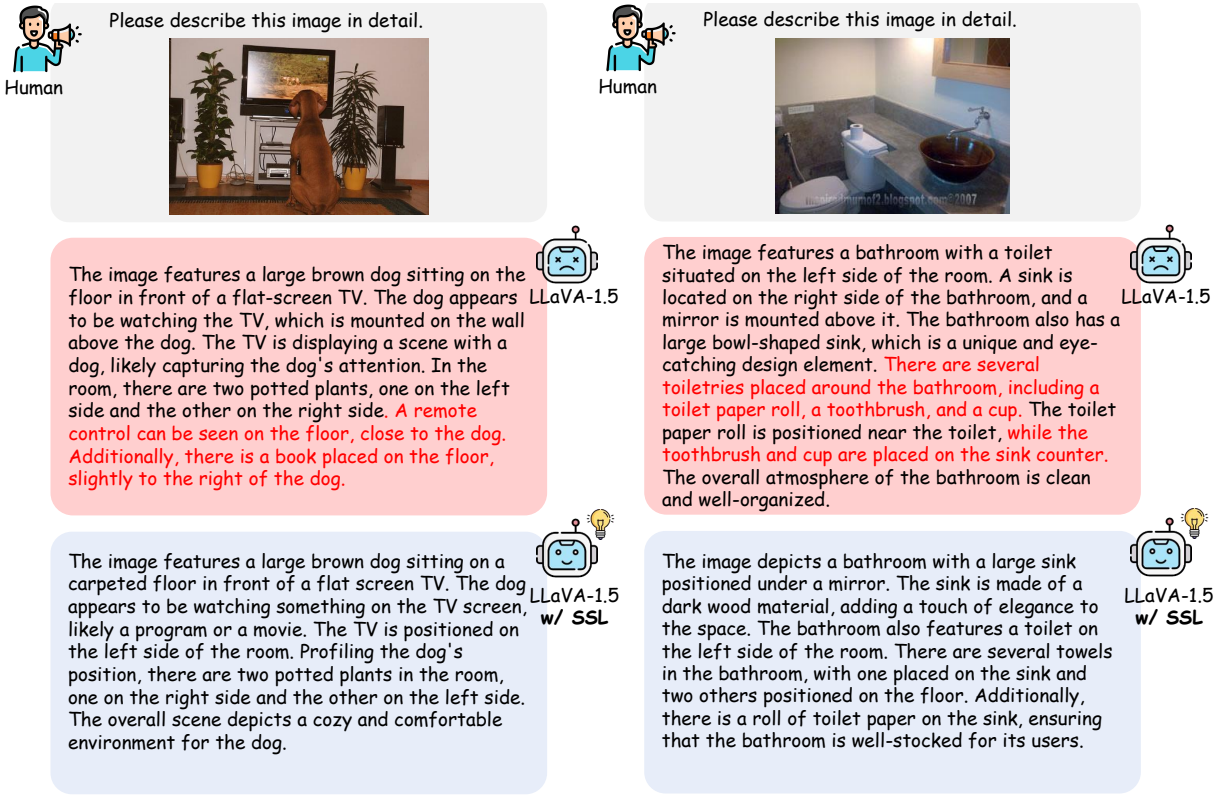


Figure 15: More examples of SSL in reducing hallucinated content in LLaVA-1.5.

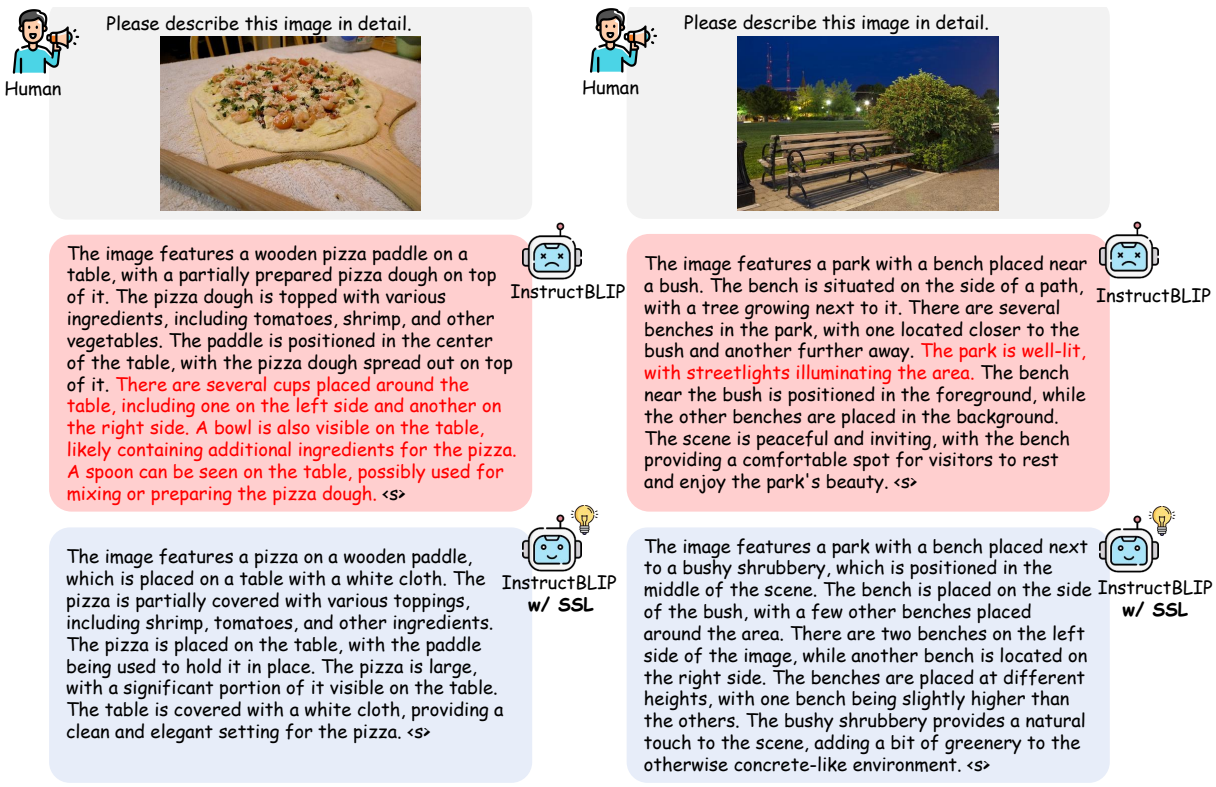


Figure 16: More examples of SSL in reducing hallucinated content in InstructBLIP.

[Input Image]

[Input Instruction] Describe this photo in detail.

[Model Outputs]

LLaVA-Next

The image is a digital artwork that pays homage to the famous painting \"Mona Lisa\" by Leonardo da Vinci. The central figure in the artwork is a dog, specifically a Chihuahua, dressed in a black dress with a white collar. The dog is wearing a brown headscarf and is seated against a backdrop of a landscape. The landscape features a river, mountains, and a castle. The dog's expression is neutral, and it is looking directly at the viewer. The artwork is a playful and creative take on the original painting, replacing the human subject with a dog.

LLaVA-Next w/ SSL

The image is a digital artwork that features a dog dressed in a style reminiscent of the Mona Lisa. The dog is depicted with a brown coat and a black face, with a large, expressive nose and eyes. It is wearing a brown headband and a dark brown shawl with a pattern that resembles the folds of the Mona Lisa's dress. The dog's arms are crossed, and it is looking directly at the viewer with a calm and serene expression. The background of the image is a landscape with a mountainous terrain, featuring a river winding through the mountains. The sky is filled with clouds, and the overall color palette is dominated by earth tones, with a focus on browns and grays. The painting style is reminiscent of the impressionist style, with loose brushstrokes and a focus on the interplay of light and shadow. The image is a playful and creative blend of the famous painting with the subject of a dog, combining elements of art history with a modern twist.

GPT-4o Evaluation

Accuracy: **4.0 9.0**
Reason: Assistant 1 inaccurately identifies the dog as a Chihuahua and mentions a black dress with a white collar, as well as a castle, none of which are present in the image. Assistant 2 accurately describes the dog in the style of the Mona Lisa, mentioning the correct attire and general description, with only minor inaccuracies like referring to the face as black instead of brown.
Detailedness: **5.0 9.0**
Reason: Assistant 1 gives less accurate details and misses several aspects of the image, resulting in less informative content. Assistant 2 provides a detailed description of the image, including the style, attire, and background elements. It accurately conveys the artistic style and elements borrowed from the Mona Lisa.

Figure 17: The performance of SSL on reducing hallucinations on LLaVA-Next-8b.

[Input Image]

[Input Instruction] What material appears to make up the creature?

[Model Outputs]

LLaVA-1.5

The creature appears to be made of a combination of plastic and glass.

LLaVA-1.5 w/ SSL

The creature appears to be made of a combination of materials, including a red or orange substance that resembles a horse or a camel, and a pair of eyeglasses. The overall appearance suggests that the substance is likely a combination of clay or a similar material, along with the eyeglasses, which add a creative and playful touch to the overall design.

GPT-4o Evaluation

Accuracy: **3.0 5.0**
Reason: Assistant 1 accurately identifies the materials as plastic and glass, which may relate to the eyeglasses, but it doesn't address the fiery aspects or colors. Assistant 2 accurately describes the creature as having a red or orange appearance and mentions eyeglasses. However, the description of the material as clay is speculative and doesn't fully align with the molten, fiery appearance.
Detailedness: **2.0 6.0**
Reason: Assistant 1 gives minimal description and lacks detail, primarily naming materials without describing the creature's form or distinctive features. Assistant 2 provides more details about the creature's appearance, color, and shape, mentioning both a horse or camel likeness and speculating on the material, despite some inaccuracies.

Figure 18: The performance of SSL on reducing hallucinations on LLaVA-1.5-7b.

[Input Image]

[Input Instruction] Describe this photo in detail.

[Model Outputs]

InstructBLIP

In the image, a man is standing on the back of a yellow car, ironing clothes on the roof. He is positioned on the right side of the car, with the laundry basket placed on the roof rack. The car is parked on a busy street, with several other vehicles and a truck in the background. The scene is set in a bustling city, with people walking around and cars passing by. The man's actions are likely to draw attention from the pedestrians, as he is ironing clothes on the roof of the car. <s>

InstructBLIP w/ SSL

The image shows a man in a yellow shirt standing on the back of a yellow SUV, ironing clothes on the roof of the vehicle. He is surrounded by various items, including a suitcase, a backpack, and a purse. The scene takes place on a busy city street, with multiple cars and a traffic light visible in the background. The man's presence on the roof of the SUV is a unique and creative way to iron clothes, making it a memorable and attention-grabbing image. <s>

GPT-4o Evaluation

Accuracy: **6.0 7.0**
Reason: Both assistants correctly describe a man ironing on the back of a yellow vehicle. However, Assistant 1 inaccurately mentions the man standing on the roof and the presence of additional items such as a suitcase, a backpack, and a purse, which are not visible. Assistant 2 incorrectly mentions ironing on the roof, but doesn't add non-existent items, therefore has a slightly more accurate description..
Detailedness: **5.0 6.0**
Reason: Assistant 1 provides some details, despite the inaccuracies, such as the mention of the city street and multiple cars, but includes incorrect additional items. Assistant 2 offers a more accurate context of the scene, noting the bustling city and people, though details about specific objects are still inaccurate. Both descriptions lack accuracy in detailedness, but Assistant 2 is slightly more coherent.

Figure 19: The performance of SSL on reducing hallucinations on InstructBLIP-7b.