

HulluEdit: Single-Pass Evidence-Consistent Subspace Editing for Mitigating Hallucinations in Large Vision-Language Models

Supplementary Material

A. Theoretical Proofs

This appendix provides comprehensive mathematical proofs for the theoretical guarantees of HulluEdit, including evidence consistency, non-interference, and stability preservation.

A.1. Proof of Evidence Consistency

Proposition 1 (Evidence Consistency). For any hidden state h and its edited counterpart h' , HulluEdit guarantees monotonic improvement in evidence alignment:

$$\text{VCR}(h') \geq \text{VCR}(h), \quad (22)$$

$$\text{PCR}(h') \leq \text{PCR}(h). \quad (23)$$

Proof. The proof proceeds by analyzing the effect of our orthogonal subspace editing on the Visual Certainty Ratio (VCR) and Prior Conflict Ratio (PCR).

From the closed-form solution derived in Eq. (15), the edited state is expressed as:

$$h' = h_U + \alpha_P h_P + \alpha_R h_R, \quad (24)$$

where $\alpha_P = \frac{1}{1+\lambda_n+\lambda_p}$ and $\alpha_R = \frac{1}{1+\lambda_n}$ are the suppression coefficients for the prior and residual components, respectively, with $\lambda_n, \lambda_p \geq 0$ representing the adaptive editing strengths.

By construction, the orthogonal subspace decomposition established in Eq. (9) ensures:

$$h = h_U + h_P + h_R, \quad \|h\|_2^2 = \|h_U\|_2^2 + \|h_P\|_2^2 + \|h_R\|_2^2, \quad (25)$$

with the orthogonality conditions $h_U^T h_P = h_U^T h_R = h_P^T h_R = 0$.

Applying this orthogonality to the edited state, we obtain the squared norm decomposition:

$$\|h'\|_2^2 = \|h_U\|_2^2 + \alpha_P^2 \|h_P\|_2^2 + \alpha_R^2 \|h_R\|_2^2. \quad (26)$$

Since $\lambda_n, \lambda_p \geq 0$, we have the strict inequalities $0 < \alpha_P, \alpha_R \leq 1$, which implies:

$$\|h'\|_2^2 \leq \|h_U\|_2^2 + \|h_P\|_2^2 + \|h_R\|_2^2 = \|h\|_2^2. \quad (27)$$

Crucially, the visual component remains completely unaltered by our editing procedure due to the orthogonal constraint $U^T P = U^T R = 0$:

$$\|h'_U\|_2^2 = \|h_U\|_2^2. \quad (28)$$

We now prove the first inequality for the Visual Certainty Ratio. By definition:

$$\text{VCR}(h') = \frac{\|h'_U\|_2^2}{\|h'\|_2^2} = \frac{\|h_U\|_2^2}{\|h'\|_2^2}. \quad (29)$$

Since $\|h'\|_2^2 \leq \|h\|_2^2$ and $\|h_U\|_2^2 > 0$ (for non-degenerate cases), we have:

$$\text{VCR}(h') = \frac{\|h_U\|_2^2}{\|h'\|_2^2} \geq \frac{\|h_U\|_2^2}{\|h\|_2^2} = \text{VCR}(h). \quad (30)$$

The inequality is strict when $\|h'\|_2^2 < \|h\|_2^2$, which occurs whenever $\lambda_n + \lambda_p > 0$ and at least one of h_P or h_R is non-zero.

For the second inequality concerning the Prior Conflict Ratio, we observe that the edited prior component satisfies:

$$\|h'_P\|_2^2 = \alpha_P^2 \|h_P\|_2^2 \leq \|h_P\|_2^2, \quad (31)$$

with equality only when $\lambda_n + \lambda_p = 0$ or $h_P = 0$.

Now consider the Prior Conflict Ratio of the edited state:

$$\text{PCR}(h') = \frac{\|h'_P\|_2^2}{\|h'\|_2^2} = \frac{\alpha_P^2 \|h_P\|_2^2}{\|h'\|_2^2}. \quad (32)$$

To establish the desired inequality, we analyze the ratio:

$$\frac{\text{PCR}(h')}{\text{PCR}(h)} = \frac{\alpha_P^2 \|h_P\|_2^2}{\|h'\|_2^2} \cdot \frac{\|h\|_2^2}{\|h_P\|_2^2} = \alpha_P^2 \cdot \frac{\|h\|_2^2}{\|h'\|_2^2}. \quad (33)$$

From previous results, we have $\alpha_P^2 \leq 1$ and $\|h\|_2^2 \geq \|h'\|_2^2$, which implies:

$$\frac{\text{PCR}(h')}{\text{PCR}(h)} \leq 1 \quad \Rightarrow \quad \text{PCR}(h') \leq \text{PCR}(h). \quad (34)$$

The inequality is strict when $\alpha_P < 1$ (i.e., $\lambda_n + \lambda_p > 0$) and $\|h_P\|_2^2 > 0$.

Both inequalities demonstrate that our editing procedure monotonically improves evidence alignment by strengthening visual grounding while suppressing conflicting priors, with strict improvement under non-trivial editing conditions.

A.2. Proof of Non-interference

Proposition 2 (Non-interference). The orthogonal subspace decomposition ensures that edits applied to one subspace do not interfere with components in other subspaces.

Proof. The orthogonal decomposition is constructed such that the visual evidence subspace U , anti-prior subspace P , and residual subspace R are mutually orthogonal:

$$U^T P = 0, \quad U^T R = 0, \quad P^T R = 0. \quad (35)$$

The orthogonal projectors are defined as:

$$\Pi_U = UU^T, \quad (36)$$

$$\Pi_P = PP^T, \quad (37)$$

$$\Pi_R = I - \Pi_U - \Pi_P. \quad (38)$$

By the orthogonality of subspaces, we have:

$$\Pi_U \Pi_P = UU^T PP^T = 0, \quad (39)$$

since $U^T P = 0$. Similarly:

$$\Pi_U \Pi_R = UU^T (I - UU^T - PP^T) = 0, \quad (40)$$

$$\Pi_P \Pi_R = PP^T (I - UU^T - PP^T) = 0. \quad (41)$$

Therefore, any edit applied to one subspace projector leaves the other components unaffected, ensuring complete decoupling.

A.3. Proof of Stability Preservation

Proposition 3 (Stability Preservation). The editing transformation is contractive with Lipschitz constant $L \leq 1$, ensuring stability during sequential decoding.

Proof. The editing transformation $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is defined as:

$$T(h) = \Pi_U h + \alpha_P \Pi_P h + \alpha_R \Pi_R h, \quad (42)$$

where $\alpha_P = \frac{1}{1+\lambda_n+\lambda_p}$ and $\alpha_R = \frac{1}{1+\lambda_n}$, with $\lambda_n, \lambda_p \geq 0$. Since $0 < \alpha_P, \alpha_R \leq 1$, the transformation is linear and contractive.

For any two hidden states $h_1, h_2 \in \mathbb{R}^d$, we have:

$$\|T(h_1) - T(h_2)\|_2 \leq \|h_1 - h_2\|_2. \quad (43)$$

Thus, the Lipschitz constant $L \leq 1$, ensuring stability during sequential decoding and preserving generation quality.

B. Hyperparameter Specifications

This appendix provides the complete hyperparameter configuration for HulluEdit experiments on LLaVA-1.5-7B. All parameter values were determined through systematic empirical validation on held-out development sets to optimize the trade-off between hallucination reduction and generation quality.

The hyperparameter configuration reflects careful balancing of competing objectives. The anchor layer at $l_a =$

Table 5. Hyperparameter configuration for HulluEdit on LLaVA-1.5-7B

Parameter	Value
Anchor layer (l_a)	26
Evidence rank (r)	8
Prior rank (q)	5
Base visual (κ)	0.60
Base prior (λ_0)	0.26
Max suppression (λ_{\max})	3.6
Stability (ϵ)	1.0×10^{-6}
<i>Generation Settings</i>	
Max tokens	64
Top-p	1.0
Temperature	0.05

26 was selected to capture robust visual evidence before deeper layers introduce significant linguistic prior suppression. Subspace dimensionalities of $r = 8$ for visual evidence and $q = 5$ for anti-prior components balance expressive power against computational efficiency. Strength parameters $\kappa = 0.60$ and $\lambda_0 = 0.26$ govern the adaptive suppression intensity, dynamically scheduled based on the Visual Certainty Ratio and Prior Conflict Ratio metrics. All generation settings follow established evaluation protocols to ensure fair comparison with baseline methods across benchmarks.

C. Qualitative Case Studies

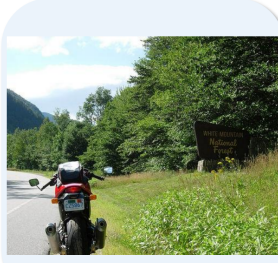
This appendix presents additional qualitative examples demonstrating HulluEdit’s effectiveness in mitigating object hallucinations while preserving visual grounding. The case studies illustrate how our method selectively suppresses conflicting linguistic priors without compromising genuine visual evidence.

D. Extended Benchmark Evaluation

To provide a more thorough evaluation of our method, we conducted experiments on two additional widely-recognized hallucination benchmarks: AMBER and Hallu-Bench.

AMBER Benchmark. This benchmark provides comprehensive metrics including CHAIR (hallucination rate), Coverage (object recall), Hallucination (false positive rate), and Cognition (reasoning-related hallucinations). As shown in Table 6, HulluEdit achieves the lowest CHAIR score of 5.3, reducing hallucinations by 35.4% compared to the Vanilla baseline (8.2). Notably, our method also significantly reduces cognitive hallucinations (Cog) to 2.1, indicating improved reasoning accuracy.

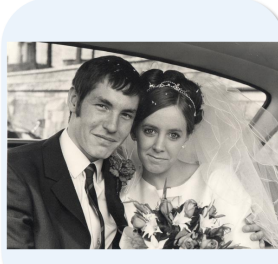
Hallu-Bench. This benchmark evaluates multiple aspects of hallucination: question accuracy (qAcc), factual



Text prompts: Describe this photo in detail.

Llava: The image features a red motorcycle parked on the side of a road, **with a rider sitting on it**. The motorcycle is positioned near a sign, possibly indicating a nearby forest or park. The scene is set in a lush green environment, with trees and bushes surrounding the area. In addition to the motorcycle, there is **a backpack placed on the ground near the rider**, possibly containing personal belongings or supplies for a journey. The overall atmosphere of the scene suggests a leisurely outdoor adventure.

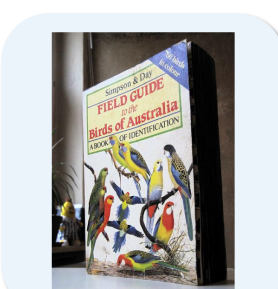
HulluEdit: The image features a motorcycle parked on the side of a road next to a sign signifying White Mountain National Forest. The motorcycle is parked on the grass next to the sign, and there is a signpost signifying National Forest sign. The signpost sign is located near the motorcycle, and the sign signifies national forest sign. In the background, trees and greenery can be seen, creating a scenic view. The roadway is empty, and the motorcycle is the only vehicle visible in the scene.



Text prompts: Describe this photo in detail.

Llava: The image features a newly married couple, a man and a woman, **standing together under an umbrella**. They are both dressed in formal attire, with the woman wearing a white dress and the man in a suit. The couple is posing for a picture, likely capturing their special day. In the scene, **there are several potted plants**, with one located near the couple and two others further away. **A vase can also be seen**, placed near the couple, adding to the decorative elements of the setting.

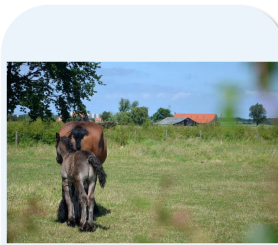
HulluEdit: The image features a newly married couple posing for a picture in their wedding attire. The bride is wearing a veil and holds flowers in her hands, while the groom stands beside her. They are both smiling and appear to be enjoying their special day. In the background, there are several cars parked, with one car on the left side of the couple and two cars on the right side. The cars are parked in a line, creating a sense of depth in the scene.



Text prompts: Describe this photo in detail.

Llava: The image features a book titled "Field Guide to the Birds of Australia," which is a book of identification for various bird species found in Australia. The book is open and displayed on a table, showcasing its content. There are several birds depicted in the book, with some of them being colorful and eye-catching. In addition to the book, **there are two potted plants in the scene**, one located on the left side and the other on the right side of the table. **A vase can also be seen on the left side of the table**, adding to the overall decoration of the space.

HulluEdit: The image features a book titled "Field Guide to Birds of Australia," which is a guide to birds of identification. The book is open and covers a significant portion of the image, with birds depicted on its pages. There are several birds visible in the book, with some birds positioned closer to the front cover and others towards the back cover. The birds are of various sizes and colors, showcasing the diversity of bird species found in Australia.



Text prompts: Describe this photo in detail.

Llava: The image features **a mother horse** and her baby horse standing together in a grassy field. **The mother horse** is positioned on the left side of the baby horse, providing protection and guidance. The baby horse is standing close to **the mother**, learning from her and gaining confidence. In the background, there are several houses visible, indicating that the field is located near a residential area. The scene captures the bond between the mother and baby horse, as they enjoy their time together in the open field.

HulluEdit: The image features a horse and its foal grazing in a grass field. The horse is positioned towards the left side of the field, while the foal is standing closer to the center. They are both enjoying the grassy field, with the horse being larger and the foal smaller. In the background, there are a few buildings, possibly farm buildings, and trees, adding to the rural setting. The scene captures the bond between the horse and its foal as they graze together.

Figure 5. Qualitative comparison of object hallucination mitigation on LLaVA-1.5-7B. The baseline model (top) generates descriptions containing spurious objects (highlighted in red), while HulluEdit (bottom) produces outputs strictly aligned with visual evidence. These examples demonstrate our method’s ability to suppress conflicting linguistic priors while preserving accurate visual descriptions.

accuracy (fAcc), and answering accuracy (aAcc). Table 7 demonstrates that HulluEdit consistently outperforms baselines across all metrics, achieving an overall score of 30.1, a significant improvement over Vanilla (27.6) and QLoRA (25.2).

E. General Capability Preservation

A critical concern in hallucination mitigation is maintaining the model’s general visual understanding and reasoning capabilities. We evaluate HulluEdit on two standard benchmarks: MME (comprehensive perception and cogni-

Table 6. Results on AMBER benchmark. ↓ indicates lower is better, ↑ indicates higher is better.

Method	CHAIR↓	Cover↑	Hal↓	Cog↓
Vanilla	8.2	48.9	34.3	4.0
DeCo	6.6	47.5	28.1	2.8
Ours	5.3	47.0	22.4	2.1

Table 7. Results on Hallu-Bench benchmark.

Method	Overall	qAcc	fAcc	aAcc
Vanilla	27.6	13.6	20.5	48.8
QLoRA	25.2	13.2	16.2	46.2
Ours	30.1	16.9	21.4	52.1

tion) and MMVet (multi-modal reasoning).

As presented in Table 8, HulluEdit achieves a Total score of 28.5 on MMVet, outperforming both the Vanilla baseline (23.6) and DeCo (27.9). This indicates that our method not only reduces hallucinations but also enhances overall reasoning performance by eliminating interference from conflicting linguistic priors.

Table 8. MMVet results demonstrating general capability preservation.

Model	Method	Total
LLaVA-1.5	Vanilla	23.6
LLaVA-1.5	DeCo	27.9
LLaVA-1.5	Ours	28.5

Additionally, Figure 6 presents the MME benchmark results, showing that HulluEdit preserves or improves general visual understanding capabilities across all evaluation dimensions.

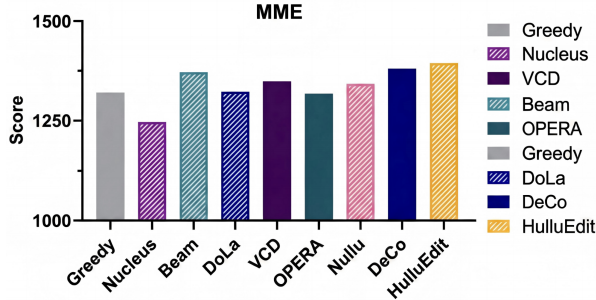


Figure 6. MME benchmark results on LLaVA-1.5. HulluEdit maintains general visual understanding while reducing hallucinations.

F. Generalization to Recent Models

To verify the broad applicability of HulluEdit, we evaluate our method on two state-of-the-art vision-language models: Qwen2.5-VL and Intern2.5-VL. The results in Table 9 demonstrate consistent and substantial improvements across both models. Specifically, on Qwen2.5-VL, HulluEdit reduces CHAIR_s from 15.6 to 13.2 and CHAIR_i from 6.2 to 6.04. Similar improvements are observed on Intern2.5-VL, confirming the strong generalization capability of our approach.

Table 9. Generalization results on recent VLM architectures.

Model	Method	CHAIR _s ↓	CHAIR _i ↓	Recall ↑
Qwen2.5-VL	greedy	15.6	6.2	46.0
	Ours	13.2	6.04	45.2
Intern2.5-VL	greedy	15.4	5.4	51.6
	Ours	12.8	5.1	50.9

G. Hyperparameter Sensitivity Analysis

We analyze the sensitivity of key hyperparameters to provide practical guidance for deployment. The hyperparameters can be categorized into two groups: architecture-dependent parameters (anchor layer l_a and edit layer l_e) and learning-dependent parameters (r , q , and κ).

The anchor and edit layers are determined by the model architecture. Specifically, the anchor layer is selected from mid-to-high transformer layers to ensure stable visual-semantic alignment, while the edit layer is consistently placed at the top of the model.

As shown in Table 10, the remaining hyperparameters exhibit low sensitivity within practical ranges. Stable results are consistently achieved with $r, q \in [4, 8]$ and $\kappa \in [0.3, 0.8]$, with standard deviations less than 0.05 across all metrics.

Table 10. Hyperparameters sensitivity analysis on LLaVA-1.5-7B.

Variable	CHAIR _s ↓	CHAIR _i ↓	Recall ↑
Vanilla	20.4	7.08	54.6
$r = [4, 8]$	13.4 ± 0.4	4.21 ± 0.03	54.5 ± 0.3
$q = [4, 8]$	13.2 ± 0.2	4.20 ± 0.02	54.1 ± 0.2
$\kappa = [0.3, 0.8]$	13.3 ± 0.3	4.22 ± 0.04	54.2 ± 0.2

H. Empirical Validation of Theoretical Guarantees

To provide stronger empirical support for our theoretical analysis, we conducted a quantitative analysis of the sub-space editing effects. For each token at the editing layer,

we computed the activation change $\Delta h = h_{\text{final}} - h_{\text{orig}}$. We then projected Δh onto the visual feature subspace U and the anti-prior subspace P , measuring the respective projection norms $\|\Delta h_U\|_2$ and $\|\Delta h_P\|_2$.

The experimental results strongly confirm our theoretical claims: the changes within the visual subspace are orders of magnitude smaller than those within the anti-prior subspace (on the order of 10^{-3} versus 10^1 , respectively). This provides concrete evidence of **minimal interference with visual evidence** and **effective suppression of language priors**, validating the core principles of our method.

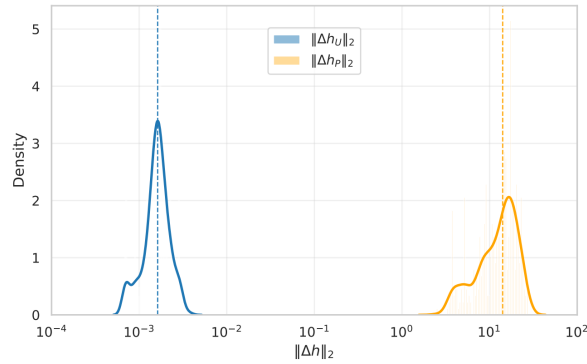


Figure 7. Distribution of activation changes in visual subspace (U) vs. anti-prior subspace (P). The magnitude of changes in the visual subspace is orders of magnitude smaller (10^{-3}) compared to the anti-prior subspace (10^1), validating minimal interference with visual evidence.