SCHRÖDINGER'S PRIOR: QUANTUM-INSPIRED REPRESENTATION INTERVENTION FOR HALLUCINATION MITIGATION IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) achieve strong multimodal performance but remain vulnerable to hallucinations, where language priors can override visually grounded evidence. Existing mitigation methods therefore tend to suppress prior-related information, implicitly treating priors as intrinsically harmful. We challenge this assumption and reveal a **context-conditioned prior duality**: the same prior can provide useful knowledge or induce hallucination depending on the current image, prompt, and decoding state. We further show that, when beneficial and harmful prior effects coexist, no fixed context-independent attenuation can improve both. Based on this finding, we propose **SchröHallu**, a training-free, quantum-inspired inference framework that models unresolved prior utility as a superposition of hallucination-inducing and knowledge-beneficial states. SchröHallu strengthens visual evidence, decomposes hidden representations into evidence, prior, and residual components, and measures whether contextual evidence sufficiently resolves the prior toward a harmful state. Only then is a minimum-norm correction applied along the decision-relevant direction of the prior subspace, preserving grounded evidence and residual capabilities. Extensive experiments across three representative LVLMs consistently reduce hallucinations while largely preserving perceptual and reasoning performance.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) integrate visual perception with large language models and achieve strong performance across image captioning, visual question answering, and multimodal dialogue (Li et al., 2023a; Liu et al., 2023; Zhu et al., 2024; Bai et al., 2023). Despite these advances, LVLMs remain prone to *multimodal hallucination*, generating plausible content that is unsupported by the input image (Li et al., 2023b; Zhou et al., 2024; Leng et al., 2024; Yang et al., 2025a). This mismatch between linguistic plausibility and visual grounding undermines their reliability in perception- and reasoning-intensive applications. Effective mitigation should therefore improve visual faithfulness while preserving the general knowledge and reasoning capabilities.

Recent studies mitigate LVLM hallucinations through increasingly fine-grained inference-time interventions, including contrastive decoding, representation editing, module intervention, visual-information steering, and encoder-side calibration (Zhou et al., 2024; Leng et al., 2024; Yang et al., 2025a; Li et al., 2025; Huang et al., 2026). These methods commonly attribute hallucinations to language-derived priors or internal biases overriding visual evidence, and thus suppress prior-related signals at selected representations or modules (Yang et al., 2025a;b; Lin et al., 2026; Huang et al., 2026). However, model priors can also provide useful parametric knowledge, while excessive suppression may degrade generation quality or introduce new errors (Wu et al., 2025; Fa et al., 2026). Thus, identifying *where* prior information resides does not determine *whether* it should be suppressed, raising a more fundamental question:

Is the harmfulness of a prior an intrinsic property that can be identified and suppressed in advance, or is its utility determined only after interacting with the current visual-linguistic context?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

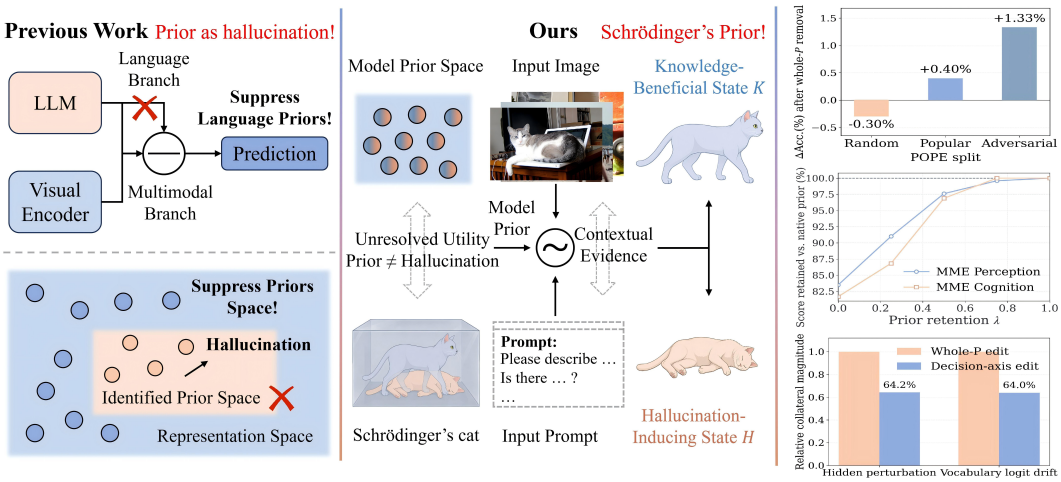


Figure 1: **Motivation and empirical evidence for SchröHallu. Left:** Previous work regards model priors as hallucination sources and suppresses them globally. **Middle:** We instead model prior utility as context-dependent, where a prior may manifest as either a knowledge-beneficial state K or a hallucination-inducing state H under contextual evidence. **Right:** Empirical results show that global prior removal has context-dependent effects, excessive suppression degrades useful perception and cognition, and decision-axis editing induces less collateral perturbation than whole-prior editing.

To investigate this question, we conduct controlled prior interventions and obtain three observations in Fig. 1. The same whole-prior removal yields opposite effects across POPE contexts, while excessive suppression degrades both perception and reasoning, showing that priors are neither uniformly harmful nor dispensable. Moreover, decision-axis editing causes substantially less collateral perturbation than whole-prior editing. These findings reveal a *context-conditioned duality*: the same prior may be hallucination-inducing or knowledge-beneficial depending on the current image, prompt, and decoding state. Accordingly, as illustrated in Fig. 1, we adopt a Schrödinger-inspired view in which prior utility remains unresolved between $|H\rangle$ and $|K\rangle$ before sufficient contextual evidence is available, motivating context-aware and decision-local intervention.

Guided by these observations, we propose **SchröHallu**, a novel quantum-inspired representation intervention framework that operates at inference time without additional training. SchröHallu first strengthens visually grounded evidence, since reliable resolution of prior utility requires sufficiently informative contextual evidence, whereas visual information can be weakened or underutilized during LVLM decoding. It then decomposes the hidden representation into visual-evidence, unresolved-prior, and residual components, and measures whether the current context sufficiently supports resolving the prior toward a harmful state. Only when such evidence is sufficient does SchröHallu intervene, applying a minimum-norm correction along the decision-relevant direction of the prior subspace while leaving visual evidence and residual capabilities unchanged.

In summary, our main contributions are threefold:

- **Empirically grounded insight.** Supported by controlled empirical observations, we reveal a *context-conditioned duality* of model priors: the same prior can be knowledge-beneficial or hallucination-inducing depending on the current multimodal context, showing that fixed context-independent suppression is fundamentally limited.
- **Methodological framework.** We propose **SchröHallu**, a training-free quantum-inspired framework that resolves prior utility from contextual evidence and applies a minimum-norm correction only along the decision-relevant prior direction.
- **Extensive empirical validation.** Extensive experiments across three representative LVLMs and four benchmarks demonstrate consistent hallucination reduction while largely preserving perceptual and reasoning performance.

2 Related Work

2.1 Hallucination Mitigation in LVLMs

Recent studies increasingly focus on training-free inference-time intervention for mitigating LVLM hallucinations. Decoding-based methods improve visual grounding through output contrast, intermediate-layer predictions, or candidate filtering (Leng et al., 2024; Li et al., 2025; Zhang et al., 2026b), while representation-level methods localize and edit hallucination-related directions or modules (Yang et al., 2025b;a; Lin et al., 2026; Zhang et al., 2026a). Complementary approaches strengthen visual evidence by reinforcing intermediate representations, correcting visual attention, or calibrating encoder-side bias (Jiang et al., 2025; Huang et al., 2026; Liu et al., 2026; Yin et al., 2025). Together, these methods establish increasingly fine-grained inference-time correction for improving visual faithfulness without additional training.

2.2 Language Priors and Prior-Oriented Intervention

Language priors are an important source of LVLM hallucination when learned linguistic regularities override image-grounded evidence (Leng et al., 2024; Yang et al., 2025a; Wang et al., 2026). Accordingly, prior-oriented methods suppress, calibrate, or isolate prior-dominated representations during generation (Yang et al., 2025a; Lin et al., 2026). However, recent studies show that priors can also provide useful parametric knowledge, while aggressive suppression may degrade generation quality or introduce image-irrelevant errors (Wu et al., 2025; Fa et al., 2026). This tension suggests that locating prior-related signals alone is insufficient, since the same prior information may play different functional roles under different multimodal contexts. Thus, beyond determining *where* and *how* to edit prior information, a key challenge is *whether* it is harmful in the current context. Our work addresses this gap by explicitly resolving context-conditioned prior utility before intervention.

Summary and Positioning. Existing studies have substantially advanced hallucination mitigation by localizing unreliable representations, strengthening visual grounding, and selectively attenuating prior-related interference. However, most approaches still treat the identified prior contribution as harmful once it conflicts with visual evidence, implicitly assigning prior information a fixed functional role. Our work departs from this static view by treating prior utility as unresolved until conditioned on the current multimodal context. This shifts the focus from merely locating and suppressing prior-related signals to first determining their decision-specific role, motivating the context-conditioned formulation and selective intervention developed in this work.

3 Context-Conditioned Duality of Model Priors

3.1 Problem Formulation

Consider an LVLM at decoding step t with multimodal context $C_t = (I, Q, y_{<t})$, where I, Q, and $y_{<t}$ denote the input image, textual instruction, and previously generated tokens, respectively. Let $h_t \in \mathbb{R}^d$ be the hidden representation for next-token prediction and $W \in \mathbb{R}^{|\mathcal{V}|\times d}$ the language-model head over vocabulary $\mathcal{V}$. For analysis, we conceptually decompose

$$h_t = h_t^U + h_t^P + h_t^R, \tag{1}$$

where h_t^U contains visually grounded evidence, h_t^P prior-related information, and h_t^R the remaining representation. This decomposition identifies the *source* of information rather than its utility; in particular, $h_t^P \neq 0$ does not imply harmfulness.

For two competing tokens $y_t^+, y_t^- \in \mathcal{V}$, let $W_y \in \mathbb{R}^d$ denote the LM-head vector associated with token y. Their decision direction and the corresponding prior contribution are

$$w_t = W_{y_t^+} - W_{y_t^-}, \qquad M_t^P = w_t^\top h_t^P. \tag{2}$$

Definition 3.1.1 (Decision-Specific Prior Utility)**.** Let $s_t \in \{+1, -1\}$ indicate the oracle-preferred alternative, with $s_t = +1$ for y_t^+ and $s_t = -1$ for y_t^-. We define the signed prior utility as

$$u_t^P = s_t M_t^P = s_t w_t^\top h_t^P. \tag{3}$$

The prior is *knowledge-beneficial* if $u_t^P > 0$, *hallucination-inducing* if $u_t^P < 0$, and neutral otherwise.

Thus, prior utility depends on both the prior contribution and the active decision context rather than being an intrinsic property of h_t^P. The oracle variable s_t is used only for theoretical and empirical analysis and is never accessed by our inference-time method.

3.2 Context-Conditioned Prior Duality

The decision-specific utility in Sec. 3.1 implies that the role of prior information should be evaluated relative to the current multimodal context rather than fixed in advance. Let $\mathcal{C}$ denote a set of decoding contexts, where each $C \in \mathcal{C}$ induces a prior component, a decision direction, and hence a signed prior utility $u^P(C)$.

Definition 3.2.1 (Context-Conditioned Prior Duality)**.** For a fixed LVLM and prior-extraction mechanism, model priors exhibit *context-conditioned duality* over $\mathcal{C}$ if there exist $C_a, C_b \in \mathcal{C}$ such that

$$u^P(C_a)\, u^P(C_b) < 0. \tag{4}$$

That is, the same prior mechanism can be knowledge-beneficial in one context but hallucination-inducing in another.

We next examine the implication for context-independent suppression. Consider a fixed attenuation factor $\lambda \in [0, 1]$ applied to the prior component:

$$h_t(\lambda) = h_t^U + \lambda h_t^P + h_t^R. \tag{5}$$

The corresponding signed margin toward the oracle-preferred alternative is $G_t(\lambda) = s_t w_t^\top h_t(\lambda)$.

Proposition 3.2.2 (Effect of Fixed Prior Attenuation)**.** *For any context and any* $\lambda \in [0, 1]$,

$$G_t(\lambda) - G_t(1) = (\lambda - 1) u_t^P. \tag{6}$$

Therefore, any nontrivial attenuation $\lambda < 1$ decreases the correct-decision margin when $u_t^P > 0$, but increases it when $u_t^P < 0$. Consequently, under Definition 3.2.1, no fixed context-independent attenuation can improve all contexts simultaneously. This theoretical result is consistent with the empirical behavior in Fig. 1, motivating prior intervention that first resolves its context-specific role.

3.3 A Schrödinger View of Unresolved Prior Utility

The duality in Sec. 3.2 implies that prior information should not be assigned a fixed beneficial or harmful role before sufficient contextual evidence is available. We therefore introduce a quantum-inspired abstraction for its unresolved decision utility.

Definition 3.3.1 (Schrödinger Prior State)**.** For an unresolved prior contribution h_t^P, we define the abstract state space

$$\mathcal{S}_P = \mathrm{span}\{|H\rangle, |K\rangle\}, \qquad |\psi_t^P\rangle = \alpha_t |H\rangle + \beta_t |K\rangle, \quad |\alpha_t|^2 + |\beta_t|^2 = 1, \tag{7}$$

where $|H\rangle$ and $|K\rangle$ denote the hallucination-inducing and knowledge-beneficial roles, respectively.

Its role is resolved only after interacting with the current multimodal context. Let

$$E_t = \Phi(C_t, h_t^U, h_t^P) \tag{8}$$

denote the contextual evidence available at step t. A contextual measurement rule $\mathcal{M}$ maps E_t to normalized evidential supports (p_t^H, p_t^K) and the corresponding polarization $q_t^H = p_t^H - p_t^K \in [-1, 1]$, where $q_t^H > 0$ favors $|H\rangle$, $q_t^H < 0$ favors $|K\rangle$, and values near zero indicate weak resolution. The concrete construction of Φ and $\mathcal{M}$ is given in Sec. 4.

Accordingly, the presence of h_t^P alone never justifies suppression. Intervention is considered only when contextual evidence resolves the prior toward $|H\rangle$ with sufficient harmful influence. This formulation is a computational analogy to quantum superposition rather than a claim that LVLM representations constitute physical quantum states.

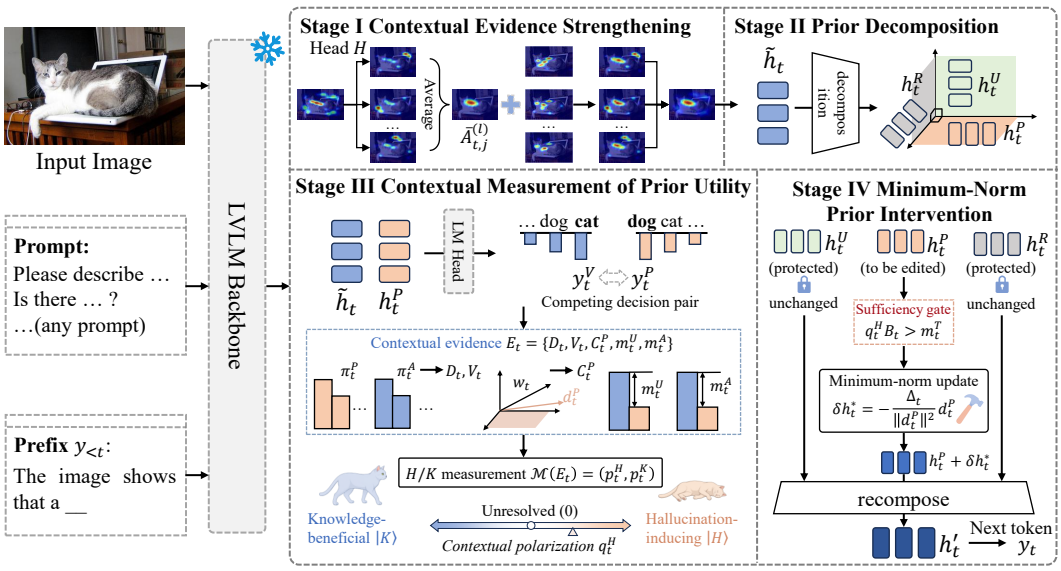


Figure 2: **Illustration of the SchröHallu framework.** (1) **Stage I:** Contextual Evidence Strengthening enhances visual evidence during decoding. (2) **Stage II:** Prior Decomposition separates visual, prior, and residual components. (3) **Stage III:** Contextual Measurement determines whether the prior is knowledge-beneficial or hallucination-inducing. (4) **Stage IV:** Selective Minimum-Norm Prior Intervention suppresses harmful priors while preserving useful information.

4 Methodology

4.1 Contextual Evidence Strengthening

Reliable resolution of prior utility requires sufficiently informative visual evidence. However, visual information can be weakened or underutilized during LVLM decoding (Li et al., 2025; Liu et al., 2026; Huang et al., 2026). We therefore strengthen the visual context before measuring whether a prior should be resolved toward $|H\rangle$ or $|K\rangle$. Following the visual-attention reinforcement principle of Jiang et al. (2025), let $A^{(l,h)}_{t,j}$ denote the pre-softmax attention logit from decoding token t to image token j at layer l and head h. For each selected layer $l \in \mathcal{L}_V$ and image-token set $\mathcal{I}$, we aggregate cross-head visual evidence and reinforce each head by

$$\bar{A}^{(l)}_{t,j} = \frac{1}{H_l}\sum_{h=1}^{H_l}\left|A^{(l,h)}_{t,j}\right|, \qquad \widetilde{A}^{(l,h)}_{t,j} = A^{(l,h)}_{t,j} + \alpha \bar{A}^{(l)}_{t,j}, \quad j \in \mathcal{I}, \tag{9}$$

where H_l is the number of attention heads and $\alpha \geq 0$ controls the reinforcement strength. Non-visual tokens and layers outside $\mathcal{L}_V$ remain unchanged. Propagating the reinforced attention through the remaining LVLM layers yields the strengthened representation $\widetilde{h}_t$, which serves as the visual-context reference for subsequent prior decomposition and contextual measurement rather than triggering intervention itself.

4.2 Evidence-Preserving Prior Decomposition

Given the strengthened representation $\widetilde{h}_t$, we next isolate prior-related information while protecting visually grounded evidence and residual model capabilities. Importantly, decomposition only identifies *where* prior information resides; whether it is harmful remains unresolved until contextual measurement. Following the evidence-preserving decomposition principle of Lin et al. (2026), let $U_t \in \mathbb{R}^{d\times r_U}$ and $P_t \in \mathbb{R}^{d\times r_P}$ denote orthonormal bases of the visual-evidence and prior subspaces, respectively, satisfying

$$U_t^\top U_t = I_{r_U}, \qquad P_t^\top P_t = I_{r_P}, \qquad U_t^\top P_t = 0. \tag{10}$$

Their projectors, together with the residual projector, are $\Pi_t^U = U_t U_t^\top$, $\Pi_t^P = P_t P_t^\top$, and $\Pi_t^R = I_d - \Pi_t^U - \Pi_t^P$. We then decompose

$$h_t^U = \Pi_t^U \widetilde{h}_t, \qquad h_t^P = \Pi_t^P \widetilde{h}_t, \qquad h_t^R = \Pi_t^R \widetilde{h}_t, \quad \widetilde{h}_t = h_t^U + h_t^P + h_t^R. \tag{11}$$

Here, h_t^U provides protected contextual evidence, h_t^P contains the unresolved prior contribution associated with either $|H\rangle$ or $|K\rangle$, and h_t^R preserves information outside these two subspaces. Accordingly, any subsequent update is restricted to $\mathrm{span}(P_t)$, leaving the visual-evidence and residual components unchanged. Whether h_t^P should actually be corrected is determined by the contextual measurement in Sec. 4.3.

4.3 Contextual Measurement of Prior Utility

The decomposition in Sec. 4.2 identifies prior-related information but leaves its utility unresolved. We therefore measure whether the current context supports interpreting h_t^P as hallucination-inducing ($|H\rangle$) or knowledge-beneficial ($|K\rangle$).

Competing decision pair. We first identify the candidates favored by the prior component and the visually strengthened decoding path:

$$y_t^P = \arg\max_{y \in \mathcal{V}} W_y^\top h_t^P, \qquad y_t^V = \arg\max_{y \in \mathcal{V}} W_y^\top \widetilde{h}_t. \tag{12}$$

If $y_t^P = y_t^V$, no actionable conflict exists and the current step remains unchanged. Otherwise, $\mathcal{Y}_t = \{y_t^P, y_t^V\}$ defines $w_t = W_{y_t^P} - W_{y_t^V}$ and its prior-subspace projection $d_t^P = \Pi_t^P w_t$.

Prior-context disagreement and visibility. For $X \in \{P, A\}$, where P denotes the prior component and A the strengthened decoding path, we define $z_{t,y}^P = W_y^\top h_t^P$ and $z_{t,y}^A = W_y^\top \widetilde{h}_t$, and normalize them over $\mathcal{Y}_t$ as $\pi_t^X(y) = \mathrm{softmax}_{\mathcal{Y}_t}(z_{t,y}^X)$. Their disagreement is measured by

$$D_t = 1 - \left(\sum_{y \in \mathcal{Y}_t} \sqrt{\pi_t^P(y) \pi_t^A(y)} \right)^2. \tag{13}$$

To distinguish meaningful conflict from an ambiguous candidate pair, we further define

$$v_t^P = \left| \pi_t^P(y_t^P) - \pi_t^P(y_t^V) \right|, \qquad v_t^A = \left| \pi_t^A(y_t^P) - \pi_t^A(y_t^V) \right|, \qquad V_t = \sqrt{v_t^P v_t^A}. \tag{14}$$

Thus, V_t is large only when both paths express a clear preference over the competing pair.

Decision relevance and contextual support. We measure whether the prior conflict is concentrated along the active decision direction by

$$C_t^P = \frac{|(d_t^P)^\top h_t^P|}{\|d_t^P\|_2 \|h_t^P\|_2 + \epsilon}. \tag{15}$$

Meanwhile, visual support for y_t^V is measured independently from the protected evidence subspace and the strengthened path:

$$m_t^U = (W_{y_t^V} - W_{y_t^P})^\top h_t^U, \qquad m_t^A = z_{t,y_t^V}^A - z_{t,y_t^P}^A. \tag{16}$$

We convert these margins into bounded directional supports and aggregate them as

$$s_t^U = \tanh\left(\frac{m_t^U}{2}\right), \qquad s_t^A = \tanh\left(\frac{m_t^A}{2}\right), \qquad S_t = \frac{s_t^U + s_t^A}{2}. \tag{17}$$

Hence, $S_t > 0$ indicates that visual context collectively supports y_t^V over prior-favored candidate.

Deterministic H/K measurement. Collecting the evidence as $E_t = \{D_t, V_t, C_t^P, m_t^U, m_t^A\}$, we define its reliability and contextual harmful polarization by

$$R_t = D_t V_t C_t^P, \qquad q_t^H = R_t S_t = D_t V_t C_t^P S_t, \tag{18}$$

where $R_t \in [0,1]$ and $q_t^H \in [-1,1]$. The deterministic measurement operator is then

$$\mathcal{M}(E_t) = (p_t^H, p_t^K) = \left(\frac{1+q_t^H}{2}, \frac{1-q_t^H}{2} \right). \tag{19}$$

Thus, $q_t^H > 0$ favors the hallucination-inducing interpretation, whereas $q_t^H < 0$ favors preserving the prior as knowledge-beneficial. The sign of S_t determines which role is favored, while R_t controls how strongly the available evidence resolves the prior. Importantly, $q_t^H > 0$ alone does not trigger an edit; Sec. 4.4 further evaluates intervention sufficiency.

4.4 Selective Minimum-Norm Prior Intervention

A positive polarization q_t^H indicates that the current context favors the hallucination-inducing interpretation, but does not alone justify intervention. We therefore separate prior harmfulness from *intervention sufficiency*. For the competing pair (y_t^P, y_t^V), the prior preference margin is

$$m_t^P = (W_{y_t^P} - W_{y_t^V})^\top h_t^P = w_t^\top h_t^P. \tag{20}$$

Together with the visual counter-evidence margin m_t^A from Eq. (16), we define

$$B_t = \min(m_t^P, m_t^A), \qquad \Delta_t = q_t^H B_t. \tag{21}$$

Since both margins are induced by their respective preferred candidates, B_t is large only when the prior favors y_t^P while the strengthened context independently favors y_t^V. An edit is admitted only if

$$q_t^H B_t > m_t^T, \tag{22}$$

where $m_t^T > 0$ denotes the required intervention margin; otherwise, the prior is preserved.

Once Eq. (22) is satisfied, suppressing the entire prior component is unnecessary. Using the prior-projected decision direction $d_t^P = \Pi_t^P w_t$, we seek the smallest admissible update:

$$\delta h_t^* = \arg\min_{\delta h_t} \|\delta h_t\|_2^2 \quad \text{s.t.} \quad \delta h_t \in \mathrm{span}(P_t),\ w_t^\top \delta h_t = -\Delta_t, \tag{23}$$

whose unique solution for $d_t^P \neq 0$ is

$$\delta h_t^* = -\frac{\Delta_t}{\|d_t^P\|_2^2} d_t^P, \qquad h_t' = \widetilde{h}_t + \delta h_t^*. \tag{24}$$

Because $\delta h_t^* \in \mathrm{span}(P_t)$, the protected visual-evidence and residual components remain unchanged. Finally, we accept the candidate edit only when its full-vocabulary prediction is y_t^V; otherwise, the original decoding state is retained. The proof and complete intervention procedure are provided in the Appendix B.

5 Experiments

5.1 Experimental Setup

We evaluate our SchröHallu on three representative LVLMs, including LLaVA-1.5-7B, LLaVA-1.5-13B (Liu et al., 2023), and DeepSeek-VL-Chat-7B (Lu et al., 2024), using their publicly released checkpoints without additional training. Experiments are conducted on POPE (Li et al., 2023b), CHAIR (Rohrbach et al., 2018), MME (Fu et al., 2025), and AMBER (Wang et al., 2023), covering discriminative and generative hallucination as well as fine-grained visual perception. We compare against the original LVLMs and representative training-free hallucination mitigation methods under standard benchmark metrics. Detailed model configurations, evaluation protocols, metrics, baselines, and implementation settings are provided in Appendix H.

Table 1: Comparison (%) with state-of-the-art hallucination mitigation methods on POPE and CHAIR across three representative LVLMs. Higher Acc. and F1 and lower C_S and C_I indicate better performance. Best results are highlighted in bold.

Method	Venue	Random		Popular		Adversarial		CHAIR	
		Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	C_S ↓	C_I ↓
LLaVA-1.5 7B									
Vanilla	-	85.1	83.7	83.7	82.4	80.9	80.0	48.0	14.3
OPERA	CVPR'24	89.1	89.0	86.0	86.3	79.1	80.9	44.6	12.8
VTI	ICLR'25	84.5	82.7	83.3	81.7	80.7	79.3	41.4	11.4
Nullu	CVPR'25	89.5	89.2	85.4	85.6	79.4	80.9	50.2	13.8
SVAR	CVPR'25	77.2	81.0	70.0	76.4	63.2	72.5	25.0	6.7
HulluEdit	CVPR'26	90.4	90.5	87.5	87.6	82.5	83.4	51.4	12.9
SHIELD	ICLR'26	90.7	90.3	87.3	87.4	82.7	82.8	36.6	10.3
LVLMs-Saliency	ICLR'26	89.1	88.2	87.1	86.0	83.3	83.1	35.6	8.2
SchröHallu	**Ours**	**91.2**	**91.6**	**87.8**	**88.2**	**83.9**	**83.8**	**21.8**	**6.5**
LLaVA-1.5 13B									
Vanilla	-	85.4	84.2	84.4	83.3	82.0	81.2	48.8	13.1
OPERA	CVPR'24	88.2	88.7	85.2	86.2	78.5	81.1	42.6	13.2
VTI	ICLR'25	85.5	84.2	84.6	83.3	82.2	81.3	32.8	9.2
Nullu	CVPR'25	88.1	88.0	83.7	84.2	78.5	80.1	49.2	13.3
SVAR	CVPR'25	81.0	79.1	81.1	79.3	75.0	74.2	28.8	8.8
HulluEdit	CVPR'26	90.6	90.8	88.0	88.3	82.7	84.0	50.6	13.4
SHIELD	ICLR'26	87.3	86.2	85.4	84.5	81.1	80.8	50.0	15.3
LVLMs-Saliency	ICLR'26	88.9	88.8	87.5	87.7	83.4	84.1	39.8	8.8
SchröHallu	**Ours**	**91.4**	**91.2**	**88.7**	**88.9**	**85.8**	**85.0**	**27.8**	**7.9**
DeepSeek-VL-Chat									
Vanilla	-	87.7	87.1	85.5	85.1	82.1	82.3	25.0	7.1
OPERA	CVPR'24	89.9	89.6	88.3	88.2	84.8	85.2	22.6	6.4
VTI	ICLR'25	90.4	90.3	88.3	88.4	83.9	84.7	24.0	8.4
Nullu	CVPR'25	90.0	89.6	88.6	88.4	85.5	85.6	25.2	6.9
SVAR	CVPR'25	86.9	87.0	83.4	84.1	81.3	82.4	21.4	6.7
HulluEdit	CVPR'26	89.1	88.8	87.7	87.6	84.1	84.5	30.8	8.7
SHIELD	ICLR'26	85.5	86.7	83.3	84.9	76.8	80.2	29.8	8.5
LVLMs-Saliency	ICLR'26	90.0	89.7	88.5	88.4	85.1	85.4	25.8	7.3
SchröHallu	**Ours**	**90.8**	**90.6**	**89.3**	**88.7**	**86.0**	**86.2**	**20.6**	**6.2**

Table 2: Comparison with representative hallucination mitigation methods on AMBER and MME with LLaVA-1.5 7B. Best results are highlighted in bold.

Method	AMBER				MME			
	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓	Existence ↑	Count ↑	Position ↑	Color ↑
Vanilla	8.3	51.8	37.8	4.5	175.6	124.6	114.0	151.0
VTI	6.9	47.2	27.0	1.8	190.0	138.3	131.7	145.0
Nullu	4.7	42.2	18.5	1.3	190.0	121.1	105.6	156.7
DeCo	6.6	47.5	28.1	2.8	175.0	128.3	98.3	125.0
HulluEdit	3.7	42.5	19.5	1.4	195.0	105.0	126.7	160.0
SchröHallu	**3.4**	47.2	**17.0**	**0.9**	**195.0**	**158.3**	**135.0**	**165.0**

5.2 Comparisons with State-of-the-Art

Evaluation on POPE and CHAIR. As shown in Table 1, SchröHallu consistently improves both discriminative and generative hallucination mitigation across three LVLMs. On LLaVA-1.5-7B, it surpasses SHIELD on POPE-Random Accuracy (91.2% vs. 90.7%), HulluEdit on Popular F1 (88.2% vs. 87.6%), and LVLMs-Saliency on Adversarial Accuracy (83.9% vs. 83.3%). Similar gains are observed on LLaVA-1.5-13B in Adversarial Accuracy (85.8% vs. 83.4%) and on DeepSeek-VL-Chat in Adversarial F1 (86.2% vs. 85.6%). On CHAIR, SchröHallu further lowers both C_S and C_I, achieving 21.8%/6.5% on LLaVA-1.5-7B, 27.8%/7.9% on LLaVA-1.5-13B, and 20.6%/6.2% on DeepSeek-VL-Chat, all outperforming the strongest competing results. These results demonstrate robust hallucination reduction across both binary object judgments and open-ended generation.

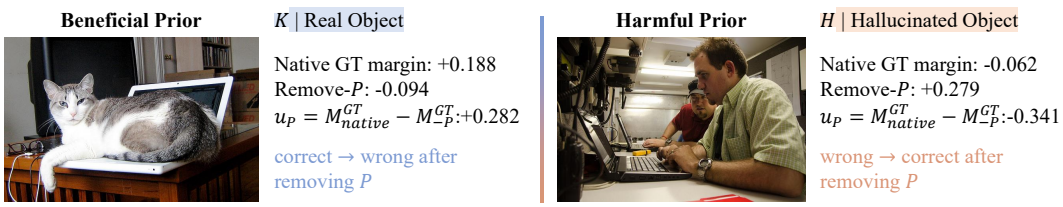


Figure 3: **Case study of the dual role of model priors.** A prior can act as beneficial knowledge (K) when supporting visually grounded content, but become harmful (H) when inducing hallucination. Prior removal correspondingly decreases or increases the ground-truth margin, showing that prior utility is context-dependent.

Evaluation on AMBER and MME. As shown in Table 2, SchröHallu achieves consistent improvements in both comprehensive hallucination evaluation and fine-grained visual perception. On AMBER, it reduces CHAIR relative to HulluEdit (3.4% vs. 3.7%), Hal relative to Nullu (17.0% vs. 18.5%), and Cog relative to Nullu (0.9% vs. 1.3%), while retaining higher Cover than HulluEdit (47.2% vs. 42.5%). On MME, SchröHallu matches HulluEdit on Existence (195.0 vs. 195.0) and substantially improves Count over VTI (158.3 vs. 138.3), Position over VTI (135.0 vs. 131.7), and Color over HulluEdit (165.0 vs. 160.0). These results indicate that SchröHallu reduces hallucinations without sacrificing, and in several cases improving, the fine-grained visual perception ability.

5.3 Case Study

Figure 3 presents representative examples illustrating the dual role of model priors across contexts. In the beneficial case, the prior supports a visually grounded object, yielding a positive ground-truth margin of 0.188. Removing it reverses the margin to -0.094, corresponding to $u_P = +0.282$, and changes a correct prediction into an incorrect one. In the harmful case, the native margin is -0.062, whereas prior removal increases it to 0.279, giving $u_P = -0.341$ and correcting the erroneous decision. These cases show that prior contributions cannot be assigned a fixed harmful or beneficial identity in advance; their utility depends on the current visual evidence and decision context, further supporting the context-conditioned prior formulation of SchröHallu.

5.4 Computational Efficiency

Table 3 compares the inference efficiency of SchröHallu with representative hallucination mitigation methods. SchröHallu achieves higher throughput than SHIELD (27.2 vs. 23.3 TPS) and LVLMs-Saliency (27.2 vs. 1.9 TPS), while using less GPU memory (15.5 GB vs. 20.2 GB and 32.4 GB, respectively). Compared with vanilla decoding, it introduces moderate overhead in throughput (27.2 vs. 50.0 TPS) and memory (15.5 GB vs. 14.5 GB) due to the additional contextual measurement path. Overall, SchröHallu provides effective hallucination mitigation with substantially lower computational overhead than recent inference-time alternatives.

Table 3: Efficiency comparison with SOTA on CHAIR using LLaVA-1.5 7B.

Method	TPS ↑	Mem. (GB) ↓
Vanilla	50.0	14.5
SchröHallu	27.2	15.5
SHIELD	23.3	20.2
LVLMs-Saliency	1.9	32.4

6 Conclusion

In this work, we revisit model priors in LVLM hallucination mitigation and show that their utility is context-conditioned rather than intrinsically harmful or beneficial. Building on this observation, we formulate unresolved prior utility from a Schrödinger-inspired perspective and propose SchröHallu, a training-free inference framework that strengthens contextual visual evidence, measures whether a prior resolves toward hallucination-inducing or knowledge-beneficial behavior, and selectively corrects only the harmful decision-relevant component through minimum-norm intervention. Extensive experiments across representative LVLMs and multiple benchmarks demonstrate consistent hallucination reduction while preserving visual perception and generation quality.

AI USE STATEMENT

Generative AI tools were used in this work solely as assistive tools for language polishing of author-written text and for drafting and refining code used to produce scientific visualizations. We did not use generative AI tools for any research tasks requiring mandatory disclosure under the ICLR 2027 AI Policy. All AI-assisted text was manually reviewed and revised by the authors to preserve the intended technical meaning, ensure consistency with the methodology and experimental results, and verify the accuracy of factual statements and citations. AI-assisted visualization code was likewise inspected and executed by the authors, and the resulting figures were checked against the corresponding experimental outputs to ensure faithful representation of the reported results. Generative AI tools were not treated as sources of scientific evidence or authority. The authors take full responsibility for the final manuscript, scientific claims, code, figures, and other artifacts presented in this work.

ETHICS STATEMENT

This work studies hallucination mitigation in Large Vision-Language Models using publicly available research benchmarks and publicly released model checkpoints. We conduct no human-subject studies, collect no new personal information, and use no private or sensitive data. The models and datasets are used for research purposes subject to their respective licenses and usage terms. The proposed method is intended to improve the faithfulness of LVLM outputs by reducing unsupported generations; however, it does not eliminate hallucinations or other limitations and biases inherited from the underlying models and data. Accordingly, the reported improvements should not be interpreted as a guarantee of factual correctness, reliability, or safety in high-stakes applications. We encourage careful evaluation of the method and the underlying LVLMs before deployment in application-specific settings.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproduction of both the proposed method and the experimental results. Appendix B presents the theoretical derivations underlying the context-independent prior attenuation analysis, the minimum-norm prior-only correction, and the full-vocabulary consistency criterion. Appendix C provides the complete inference procedure in pseudocode. Appendix D reports additional empirical evidence across different LVLMs, while Appendices F and G provide component-wise ablations and parameter analyses, respectively. Detailed model configurations, benchmarks and evaluation metrics, baseline methods, decoding protocols, hyperparameter-selection procedures, and computational settings are documented in Appendix H. All experiments use publicly released LVLM checkpoints, and the proposed method operates entirely at inference time without updating the original model parameters. Together, these materials are intended to make the methodological procedure and experimental evaluation transparent and reproducible. We will release the code upon publication.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

Zhan Fa, Yue Duan, Jian Zhang, Lei Qi, and Yinghuan Shi. One token, two fates: A unified framework via vision token manipulation against mllms hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11106–11115, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427, 2024.

Yiyang Huang, Liang Shi, Yitian Zhang, Yi Xu, and Yun Fu. Shield: Suppressing hallucinations in lvlm encoders via bias and vulnerability defense. In *Proceedings of the Conference on International Conference on Learning Representations (ICLR)*, volume 2026, pp. 125123–125143, 2026.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the Conference on International Conference on Machine Learning (ICML)*, pp. 19730–19742. PmLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023b.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the Conference on International Conference on Machine Learning (ICML)*, volume 267, pp. 35799–35819. PMLR, 2025.

Yangguang Lin, Quan Fang, Yufei Li, Jiachen Sun, Junyu Gao, and Jitao Sang. Hulluedit: Single-pass evidence-consistent subspace editing for mitigating hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11086–11095, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 34892–34916, 2023.

Sheng Liu, Haotian Ye, and James Y. Zou. Reducing hallucinations in large vision-language models via latent space steering. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Hanqing Lu, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *Proceedings of the Conference on International Conference on Learning Representations (ICLR)*, pp. 72849–72876, 2026.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4035–4045, 2018.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 13712–13736, 2025.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Zichuan Wang, Songlin Yang, Bo Peng, Zhenchen Tang, Yang Li, Beibei Dong, and Jing Dong. Same attention, different truths: Put logit-lens over visual attention to detect and mitigate lvlm object hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25315–25325, 2026.

Zongyu Wu, Yuwei Niu, Hongcheng Gao, Minhua Lin, Zhiwei Zhang, Zhifang Zhang, Qi Shi, Yilong Wang, Sike Fu, Junjie Xu, et al. Lanp: Rethinking the impact of language priors in large vision-language models. *arXiv preprint arXiv:2502.12359*, 2025.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14635–14645. IEEE, 2025a.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *Proceedings of the Conference on International Conference on Learning Representations (ICLR)*, volume 2025, pp. 51546–51568, 2025b.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14625–14634, 2025.

Chengsheng Zhang, Chenghao Sun, Xinyan Jiang, Wei Li, and Xinmei Tian. Prefill-time intervention for mitigating hallucination in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25293–25303, 2026a.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Barrett Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. Hallucination begins where saliency drops. In *Proceedings of the Conference on International Conference on Learning Representations (ICLR)*, volume 2026, pp. 15062–15083, 2026b.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *Proceedings of the Conference on International Conference on Learning Representations (ICLR)*, volume 2024, pp. 56969–56998, 2024.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *Proceedings of the Conference on International Conference on Learning Representations (ICLR)*, volume 2024, pp. 18378–18394, 2024.

Appendix

CONTENTS

A Quantum-Theoretic Preliminaries **14**
A.1 Hilbert Space and Basis States 14
A.2 Quantum Superposition and Coherence 14
A.3 Quantum Measurement and Born Rule 15
A.4 State Collapse and Contextual Resolution 16
A.5 Quantum Decoherence and Context Interaction 17
A.6 Quantum Fidelity and State Distinguishability 19
A.7 Mapping Quantum Concepts to Model Priors 20
A.8 Scope of the Quantum Analogy 21

B Theoretical Proof **22**
B.1 Effect of Context-Independent Prior Attenuation 23
B.2 Derivation of the Minimum-Norm Prior-Only Correction 23
B.3 Full-Vocabulary Consistency 24

C Pseudocode of SchröHallu **24**

D Additional Empirical Evidence Across LVLMs **25**

E Discussion **28**

F Ablation Studies **29**

G Parameter Analysis **30**

H Detailed Experimental Setup **30**

I Limitations and Future Work **32**

A Quantum-Theoretic Preliminaries

To facilitate the understanding of the quantum-inspired formulation introduced in the main paper, we briefly review the quantum-theoretic concepts that are directly related to our modeling of context-conditioned prior utility. We first introduce Hilbert spaces and basis states in Sec. A.1 to establish the two-dimensional utility space spanned by the hallucination-inducing state $|H\rangle$ and the knowledge-beneficial state $|K\rangle$. Sec. A.2 then reviews quantum superposition and coherence, which motivate representing prior utility as unresolved before sufficient contextual evidence is available. In Sec. A.3, we introduce quantum measurement and the Born rule as the conceptual basis for resolving such an unresolved state according to the current multimodal context, while Sec. A.4 further explains the connection between measurement-induced state collapse and our notion of context-conditioned utility resolution. Sec. A.5 discusses quantum decoherence and its analogy to the reduction of ambiguity through interaction with contextual evidence, which also motivates strengthening the contextual reference before measuring prior utility. Sec. A.6 introduces quantum fidelity and state distinguishability, providing the mathematical motivation for the agreement and disagreement measures used in our contextual prior measurement. We subsequently summarize the correspondence between these quantum concepts and the concrete components of SchröHallu in Sec. A.7. Finally, Sec. A.8 clarifies the scope and limitations of the analogy. These preliminaries are introduced solely to support the mathematical and conceptual formulation of our method, rather than to claim that LVLMs constitute physical quantum systems.

A.1 Hilbert Space and Basis States

In quantum theory, the state of a system is represented by a vector in a Hilbert space $\mathcal{H}$, i.e., a complete inner-product vector space. A set of mutually orthogonal and normalized vectors can be used as basis states to describe the possible outcomes associated with a particular observable. For a two-dimensional state space, let $\{|0\rangle, |1\rangle\}$ denote an orthonormal basis satisfying

$$\langle 0|0\rangle = \langle 1|1\rangle = 1, \qquad \langle 0|1\rangle = \langle 1|0\rangle = 0. \tag{25}$$

Any valid state defined in this space can therefore be expressed using these basis states. This vector-space formulation provides a compact mathematical representation for multiple possible states before a particular outcome is resolved.

Following this formalism, we introduce a two-dimensional *prior-utility space* for the quantum-inspired modeling used in this work:

$$\mathcal{H}_P = \operatorname{span}\{|H\rangle, |K\rangle\}, \tag{26}$$

where the two orthonormal basis states satisfy

$$\langle H|H\rangle = \langle K|K\rangle = 1, \qquad \langle H|K\rangle = 0. \tag{27}$$

Here, $|H\rangle$ denotes the *hallucination-inducing* functional role of a model prior, whereas $|K\rangle$ denotes its *knowledge-beneficial* role. These states characterize the possible *utilities* of prior information for the current decision, rather than different semantic identities of the prior itself. In particular, the same prior knowledge may correspond to either basis state under different multimodal contexts.

It is important to distinguish the abstract state space $\mathcal{H}_P$ from the neural hidden space $\mathbb{R}^d$ used by an LVLM. The hidden representation $h_{P,t} \in \mathbb{R}^d$ introduced in the main paper specifies the prior-derived information encoded by the network, whereas $|H\rangle$ and $|K\rangle$ describe two possible functional interpretations of that prior contribution with respect to the current decision. Therefore, Eq. (26) is not intended to assert that neural activations physically inhabit a quantum Hilbert space; rather, it provides an auxiliary state space for reasoning about the unresolved utility of model priors.

A.2 Quantum Superposition and Coherence

A fundamental property of quantum theory is that a state can exist as a coherent superposition of multiple basis states before measurement. Given the two-dimensional Hilbert space introduced above, a general normalized state can be written as

$$|\psi\rangle = \alpha|H\rangle + \beta|K\rangle, \qquad |\alpha|^2 + |\beta|^2 = 1, \tag{28}$$

where $\alpha, \beta \in \mathbb{C}$ are the probability amplitudes associated with the two basis states. Importantly, a superposition does not simply mean that the system has already occupied one basis state and that its identity is unknown. Instead, before measurement, the state is represented jointly by the amplitudes of all admissible basis components.

This distinction can be made explicit through the density-matrix representation. For the pure state in Eq. (28), the corresponding density operator is

$$\rho = |\psi\rangle\langle\psi| = \begin{bmatrix} |\alpha|^2 & \alpha\beta^* \\ \alpha^*\beta & |\beta|^2 \end{bmatrix}. \tag{29}$$

The diagonal entries describe the contributions associated with the basis states, whereas the off-diagonal terms

$$\rho_{HK} = \alpha\beta^*, \qquad \rho_{KH} = \alpha^*\beta \tag{30}$$

encode the coherence between them. By contrast, an incoherent classical mixture of the same two basis states takes the form

$$\rho_{\text{mix}} = p_H|H\rangle\langle H| + p_K|K\rangle\langle K| = \begin{bmatrix} p_H & 0 \\ 0 & p_K \end{bmatrix}, \qquad p_H + p_K = 1, \tag{31}$$

where no off-diagonal coherence remains. Thus, quantum superposition provides a mathematical language for describing a state whose basis-level character has not yet been reduced to an incoherent alternative.

This distinction motivates our representation of unresolved model-prior utility. As established in Sec. 3.2, the presence of a prior contribution $h_{P,t}$ does not determine in advance whether that contribution is hallucination-inducing or knowledge-beneficial. Before sufficient evidence from the current image, question, and decoding context is incorporated, we therefore represent its unresolved functional role as

$$|\psi_{P,t}\rangle = \alpha_t|H\rangle + \beta_t|K\rangle, \qquad |\alpha_t|^2 + |\beta_t|^2 = 1. \tag{32}$$

Here, $|H\rangle$ and $|K\rangle$ do not denote two different pieces of prior knowledge. Rather, they represent two possible decision-level utilities of the same prior-derived information. The coefficients α_t and β_t therefore encode the unresolved support for these two functional interpretations before contextual measurement.

The notion of coherence is particularly useful for expressing this unresolved status. In our abstraction, a nonzero cross-state relation between $|H\rangle$ and $|K\rangle$ indicates that the prior has not yet been assigned a single context-independent role. Once sufficient multimodal evidence is incorporated, this unresolved representation can be progressively resolved toward one of the two functional states. This motivates the subsequent measurement and decoherence formulation used in SchröHallu.

Remark A.2.1 (Interpretation of Coherence in SchröHallu). The coherence discussed above is used as a quantum-inspired conceptual abstraction rather than as a physically measured property of neural activations. We do not claim that the hidden representation $h_{P,t}$ possesses physical quantum coherence, nor do we estimate the off-diagonal entries of a physical density matrix during inference. Instead, the superposition formulation expresses the central assumption of our work: prior utility should remain unresolved until sufficient contextual evidence determines whether the prior supports or conflicts with the current decision.

A.3 Quantum Measurement and Born Rule

In quantum theory, measurement provides the mechanism through which an unresolved state is associated with an observable outcome. For the two-dimensional Hilbert space $\mathcal{H}_P = \text{span}\{|H\rangle, |K\rangle\}$, the projective measurement associated with the two basis states can be represented by

$$\Pi_H = |H\rangle\langle H|, \qquad \Pi_K = |K\rangle\langle K|, \tag{33}$$

where the projectors satisfy

$$\Pi_H^2 = \Pi_H, \qquad \Pi_K^2 = \Pi_K, \qquad \Pi_H\Pi_K = 0, \qquad \Pi_H + \Pi_K = I. \tag{34}$$

These properties ensure that $|H\rangle$ and $|K\rangle$ constitute mutually exclusive outcomes under the corresponding measurement.

For a normalized superposition

$$|\psi\rangle = \alpha|H\rangle + \beta|K\rangle, \tag{35}$$

the Born rule assigns the probabilities of observing the two outcomes as

$$\Pr_{\text{Born}}(H) = \langle\psi|\Pi_H|\psi\rangle = |\alpha|^2, \tag{36}$$

and

$$\Pr_{\text{Born}}(K) = \langle\psi|\Pi_K|\psi\rangle = |\beta|^2. \tag{37}$$

Since $|\alpha|^2+|\beta|^2 = 1$, the two outcomes form a normalized measurement distribution. Conceptually, measurement therefore converts the amplitudes of an unresolved state into outcome-specific support with respect to the selected measurement basis.

This measurement perspective motivates how SchröHallu resolves the utility of a model prior. At decoding step t, the relevant multimodal context is

$$\mathcal{C}_t = (I,\, Q,\, y_{<t})\,, \tag{38}$$

where I denotes the input image, Q the textual instruction or question, and $y_{<t}$ the previously generated sequence. From this context and the decomposed representations, SchröHallu constructs contextual evidence

$$\mathcal{E}_t = \Phi\,(I,\, Q,\, y_{<t},\, h_{U,t},\, h_{P,t})\,, \tag{39}$$

which characterizes how the current visual evidence and prior-derived information support competing decisions.

We then use an operational measurement rule

$$\mathcal{M} : \mathcal{E}_t \longmapsto (p_{H,t}, p_{K,t})\,, \qquad p_{H,t} + p_{K,t} = 1, \tag{40}$$

where $p_{H,t}$ and $p_{K,t}$ quantify the contextual evidence supporting the hallucination-inducing and knowledge-beneficial interpretations, respectively. Their relative tendency is summarized by the contextual polarization

$$q_{H,t} = p_{H,t} - p_{K,t} = 2p_{H,t} - 1. \tag{41}$$

Accordingly, $q_{H,t} > 0$ indicates that the available context provides stronger evidence for the hallucination-inducing interpretation, whereas $q_{H,t} < 0$ favors the knowledge-beneficial interpretation. Importantly, this measurement determines how the prior should be interpreted under the current context; it does not by itself imply that an intervention must be performed.

Remark A.3.1 (Born Probabilities vs. Contextual Evidential Supports). The quantities $p_{H,t}$ and $p_{K,t}$ used in SchröHallu should not be interpreted as physical Born probabilities. In standard quantum theory, Eqs. (36)-(37) follow from the amplitudes of a physical quantum state and a specified measurement operator. In our framework, $p_{H,t}$ and $p_{K,t}$ are operational evidential supports derived from multimodal model signals, including prior-visual disagreement, prior-axis concentration, and visually grounded decision margins. Therefore, the Born rule serves as a conceptual template for resolving an unresolved state into outcome-specific support, rather than as a probabilistic law assumed to govern LVLM representations.

A.4 State Collapse and Contextual Resolution

In the standard projective measurement formalism, once a measurement outcome is observed, the pre-measurement superposition is projected onto the corresponding eigenspace. For the binary measurement introduced in Sec. A.3, observing the outcome H transforms

$$|\psi\rangle = \alpha|H\rangle + \beta|K\rangle \tag{42}$$

into

$$|\psi\rangle \xrightarrow{H} \frac{\Pi_H|\psi\rangle}{\sqrt{\langle\psi|\Pi_H|\psi\rangle}} = |H\rangle, \tag{43}$$

whereas observing K yields

$$|\psi\rangle \xrightarrow{K} \frac{\Pi_K|\psi\rangle}{\sqrt{\langle\psi|\Pi_K|\psi\rangle}} = |K\rangle. \tag{44}$$

This measurement-induced transition is commonly referred to as state collapse: an unresolved superposition is associated with a definite outcome under the selected measurement basis.

The central analogy in SchröHallu is that the functional role of a model prior is likewise not assigned independently of the context in which it is used. Let P denote a fixed prior-derived semantic tendency and let

$$\mathcal{C}_t = (I, Q, y_{<t}) \tag{45}$$

denote the current multimodal context. Although the identity of P remains unchanged, its decision-specific utility may differ across contexts. We express this contextual resolution abstractly as

$$|\psi_{P,t}\rangle \xrightarrow{\mathcal{M}(\mathcal{E}_t)} \mathcal{R}_t, \tag{46}$$

where $\mathcal{M}(\mathcal{E}_t)$ denotes the contextual measurement introduced previously and $\mathcal{R}_t$ denotes the resolved functional interpretation of the prior under the current decision context.

Using the contextual polarization

$$q_{H,t} = p_{H,t} - p_{K,t}, \tag{47}$$

the two dominant functional interpretations can be written as

$$\mathcal{R}_t = \begin{cases} H, & q_{H,t} > 0, \\ K, & q_{H,t} < 0. \end{cases} \tag{48}$$

A positive polarization means that the available evidence favors interpreting the prior as hallucination-inducing, whereas a negative polarization favors a knowledge-beneficial interpretation. When $q_{H,t}$ is close to zero or the available contextual evidence is insufficient, SchröHallu does not force a reliable binary interpretation for intervention. Instead, the prior remains operationally uncertain and the original decoding state is preserved unless the subsequent sufficiency criterion is satisfied.

This distinction is important because *prior identity* and *prior utility* are different quantities. Consider the same semantic prior P under two contexts $\mathcal{C}_a$ and $\mathcal{C}_b$. Contextual resolution may produce

$$\mathcal{M}(\mathcal{E}_a) \Rightarrow q_H(\mathcal{C}_a) < 0, \qquad \mathcal{M}(\mathcal{E}_b) \Rightarrow q_H(\mathcal{C}_b) > 0, \tag{49}$$

such that the same prior is beneficial in one decision but harmful in another. This is precisely the operational counterpart of the context-conditioned prior duality established in the main paper.

A simple example is the semantic prior that "bananas are typically yellow." When an image contains a visually supported yellow banana and the question concerns its current color, this prior can reinforce the correct decision and therefore resolve toward $|K\rangle$. If the image instead depicts a non-yellow banana while the same question asks about the observed color, the identical semantic prior may conflict with the visual evidence and resolve toward $|H\rangle$. Moreover, if the question asks about the *typical* color of bananas rather than the color observed in the image, the same prior can again become knowledge-beneficial. Hence, the functional role cannot be determined from the semantic content of the prior alone; it is jointly determined by the prior and the current multimodal decision context.

Remark A.4.1 (Collapse vs. Contextual Resolution). SchröHallu does not implement physical wavefunction collapse. Eqs. (43)-(44) describe the standard quantum-mechanical notion that motivates our abstraction, whereas Eq. (46) denotes an operational inference procedure over neural representations. Moreover, our method retains the continuous polarization $q_{H,t}$ rather than replacing the prior representation with a literal basis state. A prior is modified only when its harmful interpretation is additionally supported by the intervention-sufficiency certificate. Thus, "collapse" provides the conceptual intuition, while *contextual resolution* is the more precise description of the mechanism used in SchröHallu.

A.5 Quantum Decoherence and Context Interaction

Quantum decoherence describes how the coherence of a quantum system is reduced through its interaction with an external environment. Consider an initially isolated two-state system

$$|\psi\rangle = \alpha|H\rangle + \beta|K\rangle, \tag{50}$$

with the corresponding density matrix

$$\rho = \begin{bmatrix} |\alpha|^2 & \alpha\beta^* \\ \alpha^*\beta & |\beta|^2 \end{bmatrix}. \tag{51}$$

As discussed previously, the off-diagonal entries encode the coherence between the two basis states.

To illustrate decoherence, let $|E_0\rangle$ denote the initial state of an external environment. Interaction between the system and the environment can be abstractly written as

$$(\alpha|H\rangle + \beta|K\rangle)\,|E_0\rangle \longrightarrow \alpha|H\rangle|E_H\rangle + \beta|K\rangle|E_K\rangle, \tag{52}$$

where $|E_H\rangle$ and $|E_K\rangle$ are environment states correlated with the two system states. After tracing out the environment, the reduced state of the original system becomes

$$\rho_{\mathrm{red}} = \begin{bmatrix} |\alpha|^2 & \alpha\beta^*\langle E_K|E_H\rangle \\ \alpha^*\beta\langle E_H|E_K\rangle & |\beta|^2 \end{bmatrix}. \tag{53}$$

The overlap

$$\gamma = \langle E_K|E_H\rangle \tag{54}$$

acts as a coherence factor. When the environment interacts only weakly with the system, $|\gamma|$ can remain close to one and substantial coherence is preserved. As the environment becomes increasingly capable of distinguishing the two alternatives, $|\gamma|$ decreases and the off-diagonal terms are suppressed. In the limiting case,

$$|\gamma| \to 0, \tag{55}$$

the reduced state approaches

$$\rho_{\mathrm{red}} \to \begin{bmatrix} |\alpha|^2 & 0 \\ 0 & |\beta|^2 \end{bmatrix}, \tag{56}$$

which behaves as an incoherent mixture in the $\{|H\rangle, |K\rangle\}$ basis.

This environment-dependent loss of coherence provides another useful analogy for context-conditioned prior utility. In SchröHallu, the semantic prior itself can be viewed as initially unresolved with respect to its decision-specific role, while the multimodal context

$$\mathcal{C}_t = (I, Q, y_{<t}) \tag{57}$$

plays the role of an external information source that makes the two functional interpretations increasingly distinguishable. Abstractly, this process can be summarized as

$$|\psi_{P,t}\rangle \xrightarrow{\mathcal{C}_t} \text{contextually distinguishable support for } |H\rangle \text{ and } |K\rangle. \tag{58}$$

The important point is not that the prior representation physically loses quantum coherence, but that sufficient interaction with contextual evidence reduces the ambiguity regarding whether the prior supports or conflicts with the current decision.

This perspective also motivates the ordering of the proposed framework. Reliable prior-utility measurement requires the contextual evidence itself to be sufficiently informative. However, LVLMs do not always exploit the available visual information strongly enough during decoding. SchröHallu therefore first strengthens the visual evidence to obtain a more informative contextual reference,

$$h_t \longrightarrow \widetilde{h}_t, \tag{59}$$

before constructing the measurement evidence $\mathcal{E}_t$. The strengthened context does not directly determine that a prior is harmful. Instead, it improves the evidence against which the unresolved prior contribution is subsequently evaluated. The resulting pipeline can therefore be interpreted as

$$\underbrace{\text{context strengthening}}_{\text{improve the external reference}} \rightarrow \underbrace{\text{prior-utility measurement}}_{\text{resolve } H/K \text{ tendency}} \rightarrow \underbrace{\text{selective intervention}}_{\text{edit only when sufficient}}. \tag{60}$$

Remark A.5.1 (Decoherence as an Analogy). The decoherence formalism above is introduced to explain the role of context in reducing ambiguity about prior utility. SchröHallu does not construct physical environment states $|E_H\rangle$ and $|E_K\rangle$, does not estimate a physical coherence factor γ, and does not claim that multimodal neural representations undergo quantum decoherence. In our framework, the "environment" corresponds conceptually to the multimodal context, while the reduction of coherence corresponds to increasingly discriminative evidence for the H and K interpretations. Therefore, contextual evidence strengthening should be understood as improving the reliability of this functional resolution rather than as implementing a physical decoherence process.

A.6 Quantum Fidelity and State Distinguishability

In quantum information theory, fidelity measures the similarity between two states. For two pure states $|\psi\rangle$ and $|\phi\rangle$, the fidelity is defined as

$$F(\psi, \phi) = |\langle\psi|\phi\rangle|^2. \tag{61}$$

It satisfies

$$0 \leq F(\psi, \phi) \leq 1, \tag{62}$$

where $F = 1$ indicates identical states up to a global phase, whereas $F = 0$ indicates orthogonal and therefore perfectly distinguishable pure states.

For general mixed states represented by density operators ρ and σ, the fidelity can be written as

$$F(\rho, \sigma) = \left[\mathrm{Tr}\left(\sqrt{\sqrt{\rho}\,\sigma\,\sqrt{\rho}}\right)\right]^2. \tag{63}$$

A larger fidelity indicates stronger similarity between the two states, while a smaller fidelity indicates that they are easier to distinguish. Thus, fidelity provides a natural connection between state similarity and distinguishability.

A particularly relevant case arises when the two density matrices are diagonal in the same basis. Let

$$\rho_P = \sum_{y\in\mathcal{Y}} \pi^P(y)|y\rangle\langle y|, \qquad \rho_A = \sum_{y\in\mathcal{Y}} \pi^A(y)|y\rangle\langle y|, \tag{64}$$

where π^P and π^A are two probability distributions over the same outcome set $\mathcal{Y}$. Since ρ_P and ρ_A commute, Eq. (63) reduces to

$$F(\rho_P, \rho_A) = \left(\sum_{y\in\mathcal{Y}} \sqrt{\pi^P(y)\pi^A(y)}\right)^2. \tag{65}$$

This quantity is the squared Bhattacharyya coefficient and is commonly referred to as the classical fidelity between the two distributions. It reaches one when the distributions are identical and decreases as their overlap becomes smaller.

This special case directly motivates the fidelity used in SchröHallu. At decoding step t, we construct a compact candidate set

$$\mathcal{Y}_t = \{y_{P,t}, y_{V,t}\}, \tag{66}$$

where $y_{P,t}$ denotes the candidate favored by the prior-derived representation and $y_{V,t}$ denotes the candidate supported by the strengthened contextual evidence. The corresponding pairwise distributions are denoted by π_t^P and π_t^A, respectively. We then compute

$$F_t = \left(\sum_{y\in\mathcal{Y}_t} \sqrt{\pi_t^P(y)\pi_t^A(y)}\right)^2. \tag{67}$$

When F_t is large, the prior-derived and context-supported decision distributions are similar, indicating that the prior is broadly consistent with the available contextual evidence. Conversely, a small F_t indicates that the two paths assign substantially different support to the competing candidates.

For convenience, SchröHallu uses the complementary disagreement score

$$D_t = 1 - F_t, \tag{68}$$

such that

$$0 \leq D_t \leq 1. \tag{69}$$

Accordingly, $D_t \approx 0$ indicates strong agreement between the two decision distributions, whereas a larger D_t indicates stronger separability between the prior-derived and context-supported predictions. This disagreement signal forms one component of the contextual evidence $\mathcal{E}_t$ used to resolve the functional role of the prior.

Importantly, distributional disagreement alone is insufficient to conclude that the prior is harmful. A disagreement can arise because the prior is misleading, because the contextual evidence is unreliable, or simply because the two paths emphasize different but plausible candidates. SchröHallu therefore combines D_t with visual evidence strength, prior-axis concentration, and decision margins before determining the H/K tendency. The fidelity term should consequently be interpreted as an *agreement indicator*, rather than as an independent hallucination detector.

Remark A.6.1 (Fidelity in SchröHallu). The fidelity in Eq. (67) is computed from classical candidate distributions rather than physical quantum states. Its relation to quantum fidelity follows from the special case in Eq. (65), where classical distributions are represented as commuting diagonal density matrices. Likewise, $D_t = 1 - F_t$ is an operational disagreement score introduced by our method and should not be confused with canonical quantum distances such as trace distance. We adopt fidelity because it provides a bounded and symmetric measure of overlap that naturally expresses the agreement between prior-derived and context-supported decisions.

A.7 Mapping Quantum Concepts to Model Priors

The preceding quantum concepts provide a unified language for describing the central observation of this work: a model prior has a fixed semantic identity but does not necessarily have a fixed decision utility. SchröHallu therefore introduces an auxiliary quantum-inspired utility state to characterize how the same prior contribution may play different functional roles under different multimodal contexts. Importantly, this utility state is conceptually distinct from the neural representation itself. The prior-derived hidden component

$$h_{P,t} \in \mathbb{R}^d \tag{70}$$

encodes prior-related information in the LVLM representation space, whereas

$$|\psi_{P,t}\rangle = \alpha_t|H\rangle + \beta_t|K\rangle \tag{71}$$

is an auxiliary abstraction describing the unresolved *functional utility* of this information for the current decision.

Before sufficient contextual evidence is considered, SchröHallu does not assign $h_{P,t}$ a context-independent harmful or beneficial label. This corresponds to the superposition perspective in Eq. (71), where $|H\rangle$ and $|K\rangle$ represent hallucination-inducing and knowledge-beneficial utilities, respectively. The current multimodal context

$$\mathcal{C}_t = (I, Q, y_{<t}) \tag{72}$$

then plays a role analogous to an external environment: its interaction with the prior contribution provides evidence that makes the two functional interpretations increasingly distinguishable.

This correspondence motivates the three-stage inference process of SchröHallu. First, contextual evidence is strengthened to obtain a more reliable reference for subsequent measurement. The strengthened representation $\widetilde{h}_t$ is then decomposed into

$$\widetilde{h}_t = h_{U,t} + h_{P,t} + h_{R,t}, \tag{73}$$

where $h_{U,t}$ contains visually supported evidence, $h_{P,t}$ contains the prior contribution whose utility remains to be resolved, and $h_{R,t}$ contains residual linguistic and reasoning information that is not directly edited.

Based on the current context and these decomposed signals, SchröHallu constructs measurement evidence

$$\mathcal{E}_t = \Phi\left(I, Q, y_{<t}, h_{U,t}, h_{P,t}\right), \tag{74}$$

and applies the operational measurement

$$\mathcal{M}(\mathcal{E}_t) \longmapsto (p_{H,t}, p_{K,t}), \qquad q_{H,t} = p_{H,t} - p_{K,t}. \tag{75}$$

Here, $p_{H,t}$ and $p_{K,t}$ are contextual evidential supports rather than physical Born probabilities. The resulting polarization $q_{H,t}$ indicates which functional interpretation is better supported by the current context, thereby providing an operational counterpart to measurement-induced state resolution.

The complete correspondence can be summarized as follows:

$$\begin{aligned} \text{Superposition} &\iff \text{unresolved prior utility},\\ \text{Environment interaction} &\iff \text{interaction with multimodal context},\\ \text{Decoherence} &\iff \text{increasing distinguishability of } H/K,\\ \text{Measurement} &\iff \text{contextual evidence evaluation},\\ \text{State resolution} &\iff \text{decision-specific prior interpretation},\\ \text{Fidelity} &\iff \text{prior-context decision agreement}. \end{aligned} \tag{76}$$

Crucially, contextual resolution and representation intervention are two distinct operations in SchröHallu. Even when $q_{H,t} > 0$ indicates that the prior is more likely to play a hallucination-inducing role, the method does not immediately suppress the entire prior component. Instead, intervention is performed only when the harmful tendency is sufficiently strong to affect the current decision:

$$q_{H,t} B_t > m_T, \tag{77}$$

where B_t measures the available decision-relevant prior strength and m_T denotes the required intervention margin. When this certificate is satisfied, SchröHallu modifies only the decision-relevant direction inside the prior subspace through the minimum-norm update

$$\delta h_t^* = -\frac{\Delta_t}{\|d_{P,t}\|_2^2} d_{P,t}, \qquad \Delta_t = q_{H,t} B_t. \tag{78}$$

The visually supported component and residual component remain unchanged. Therefore, the quantum-inspired formulation is used to determine *when and how the utility of a prior should be interpreted*, while the actual correction remains a deterministic and geometry-constrained neural intervention.

Overall, the relationship between the quantum abstraction and the proposed method can be summarized by

$$\underbrace{|\psi_{P,t}\rangle}_{\text{unresolved utility}} \xrightarrow{\mathcal{C}_t} \underbrace{\mathcal{E}_t}_{\text{contextual evidence}} \xrightarrow{\mathcal{M}} \underbrace{q_{H,t}}_{\text{utility resolution}} \xrightarrow{\text{sufficiency}} \underbrace{\delta h_t^*}_{\text{selective intervention}} . \tag{79}$$

This formulation captures the key distinction of SchröHallu: rather than treating prior information as intrinsically harmful and suppressing it globally, the method first resolves its utility under the current context and intervenes only when the resolved harmful contribution is sufficient to alter the decision.

A.8 Scope of the Quantum Analogy

The quantum formulation in this work is intended as a mathematical and conceptual abstraction for reasoning about context-conditioned prior utility. We do *not* claim that LVLMs are physical quantum systems, that their hidden activations correspond to physical quantum states, or that neural inference obeys quantum dynamics. In particular, the Schrödinger equation, Hamiltonian evolution, and other physical mechanisms of quantum systems are not assumed or employed by SchröHallu. The purpose of the analogy is instead to provide a principled language for distinguishing a prior's semantic identity from its unresolved, context-dependent functional utility.

More specifically, the Hilbert space

$$\mathcal{H}_P = \mathrm{span}\{|H\rangle, |K\rangle\} \tag{80}$$

is an *auxiliary utility space*, rather than the physical state space of the neural network. Similarly,

$$|\psi_{P,t}\rangle = \alpha_t |H\rangle + \beta_t |K\rangle \tag{81}$$

should be interpreted as an abstract representation of unresolved prior utility. It expresses that the same prior contribution should not be assigned an intrinsically hallucination-inducing or knowledge-beneficial role before the current multimodal context is considered. It does not imply that the neural representation $h_{P,t}$ is literally a coherent quantum wave function.

The notions of environment interaction, decoherence, measurement, and state collapse are used at the level of structural correspondence. The multimodal context

$$\mathcal{C}_t = (I, Q, y_{<t}) \tag{82}$$

is analogous to an external information source that makes the competing H/K interpretations increasingly distinguishable. Contextual evidence evaluation plays a role analogous to measurement, while contextual resolution describes how the functional role of the prior becomes increasingly determined for the current decision. Nevertheless, SchröHallu neither constructs physical environment states nor performs physical wave-function collapse. Accordingly, terms such as "decoherence" and "collapse" describe the intuition behind context-induced utility resolution rather than literal physical processes.

The quantities $p_{H,t}$ and $p_{K,t}$ should be interpreted under the same restriction. They are operational evidential supports derived from neural signals,

$$\mathcal{M}(\mathcal{E}_t) \longmapsto (p_{H,t}, p_{K,t}), \tag{83}$$

rather than calibrated Born probabilities generated from physical probability amplitudes. Consequently, SchröHallu does not require α_t and β_t to be explicitly instantiated or estimated during inference. The superposition amplitudes provide a conceptual representation of unresolved utility, whereas the practical method directly estimates context-dependent evidence for the two functional interpretations.

Among the quantum concepts adopted in this work, fidelity has a more direct mathematical connection to our implementation. When two classical probability distributions are represented as commuting diagonal density matrices, quantum fidelity reduces exactly to

$$F = \left(\sum_y \sqrt{\pi^P(y)\pi^A(y)} \right)^2, \tag{84}$$

which is the form used by SchröHallu to measure agreement between prior-derived and context-supported candidate distributions. Even in this case, however, the distributions remain ordinary neural decision distributions rather than quantum states.

Finally, the selective intervention performed after utility resolution has no intended quantum-mechanical interpretation. Once contextual evidence indicates a potentially harmful prior and the intervention-sufficiency criterion is satisfied, SchröHallu applies a deterministic minimum-norm update within the prior subspace. Thus, the quantum-inspired formulation and the neural intervention serve two distinct purposes:

$$\underbrace{\text{Quantum-inspired abstraction}}_{\text{interpret context-conditioned prior utility}} \quad + \quad \underbrace{\text{Geometric neural intervention}}_{\text{correct harmful decision influence}}. \tag{85}$$

The former provides a framework for answering *when a prior should be regarded as harmful or beneficial*, while the latter specifies *how a harmful prior contribution should be minimally corrected*.

Therefore, the term *Schrödinger prior* refers specifically to the context-dependent unresolved nature of prior utility:

$$\boxed{\text{fixed prior identity} \not\Rightarrow \text{fixed prior utility}} \tag{86}$$

and

$$\boxed{\text{prior utility} = f(P, I, Q, y_{<t}).} \tag{87}$$

The quantum analogy provides an intuitive and mathematically structured way to express this principle, while all computations and interventions in SchröHallu remain classical neural-network operations.

B THEORETICAL PROOF

This section provides the detailed derivations omitted from the main paper for conciseness. We first establish the effect of context-independent prior attenuation and then derive the minimum-norm prior-only correction used by SchröHallu.

B.1 Effect of Context-Independent Prior Attenuation

We first provide the proof of Proposition 3.2.2. Recall that the attenuated hidden representation is

$$h_t(\lambda) = h_t^U + \lambda h_t^P + h_t^R, \qquad \lambda \in [0, 1], \tag{88}$$

and the signed margin toward the oracle-preferred alternative is

$$G_t(\lambda) = s_t w_t^\top h_t(\lambda). \tag{89}$$

Proof. Substituting the attenuated representation gives

$$G_t(\lambda) = s_t w_t^\top \left(h_t^U + \lambda h_t^P + h_t^R\right) \tag{90}$$

$$= G_t(1) + (\lambda - 1) s_t w_t^\top h_t^P. \tag{91}$$

From Definition 3.1.1, $s_t w_t^\top h_t^P = u_t^P$. Therefore,

$$G_t(\lambda) - G_t(1) = (\lambda - 1) u_t^P, \tag{92}$$

which proves Proposition 3.2.2. □

For any nontrivial attenuation $\lambda < 1$, the factor $\lambda - 1$ is negative. Consequently, attenuation decreases the signed correct-decision margin when $u_t^P > 0$, but increases it when $u_t^P < 0$. This establishes the following consequence.

Corollary B.1.1 (No Uniformly Beneficial Context-Independent Suppression)**.** *If context-conditioned prior duality holds over $\mathcal{C}$, no fixed $\lambda < 1$ can improve the signed correct-decision margin for every $C \in \mathcal{C}$.*

Proof. By Definition 3.2.1, there exist $C_a, C_b \in \mathcal{C}$ such that $u^P(C_a)u^P(C_b) < 0$. Thus, one context has positive prior utility and the other negative prior utility. Proposition 3.2.2 shows that any fixed $\lambda < 1$ necessarily decreases the signed margin in the former while increasing it in the latter. Hence, the same context-independent attenuation cannot benefit both contexts simultaneously. □

Interpretation of the bounded directional support. For a two-candidate softmax with logits z_V and z_P,

$$\pi(y_V) - \pi(y_P) = \frac{e^{z_V} - e^{z_P}}{e^{z_V} + e^{z_P}} = \tanh\left(\frac{z_V - z_P}{2}\right). \tag{93}$$

Therefore, the transformation $s = \tanh(m/2)$ maps a pairwise logit margin directly to the signed difference between the corresponding normalized two-candidate probabilities. Positive values favor the visually supported candidate, whereas negative values favor the prior-supported candidate.

B.2 Derivation of the Minimum-Norm Prior-Only Correction

We next derive the closed-form update used in Sec. 4.4. Once the intervention-sufficiency condition is satisfied, SchröHallu seeks the smallest perturbation restricted to the prior subspace:

$$\delta h_t^* = \arg\min_{\delta h_t} \|\delta h_t\|_2^2 \quad \text{s.t.} \quad \delta h_t \in \operatorname{span}(P_t), \quad w_t^\top \delta h_t = -\Delta_t. \tag{94}$$

Recall that $d_t^P = \Pi_t^P w_t$. Since any admissible update satisfies $\delta h_t = \Pi_t^P \delta h_t$, we have

$$w_t^\top \delta h_t = (\Pi_t^P w_t)^\top \delta h_t = (d_t^P)^\top \delta h_t. \tag{95}$$

Hence, Eq. (94) is equivalent to

$$\min_{\delta h_t} \|\delta h_t\|_2^2 \quad \text{s.t.} \quad (d_t^P)^\top \delta h_t = -\Delta_t. \tag{96}$$

Proposition B.2.1 (Minimum-Norm Prior-Only Correction)**.** *If $d_t^P \neq 0$, the unique solution is*

$$\delta h_t^* = -\frac{\Delta_t}{\|d_t^P\|_2^2} d_t^P. \tag{97}$$

Proof. By the Cauchy-Schwarz inequality,

$$\Delta_t = \left|(d_t^P)^\top \delta h_t\right| \le \|d_t^P\|_2 \|\delta h_t\|_2. \tag{98}$$

Therefore, every feasible update satisfies

$$\|\delta h_t\|_2 \ge \frac{\Delta_t}{\|d_t^P\|_2}. \tag{99}$$

Equality holds only when δh_t is parallel to d_t^P. Because the constraint requires a negative inner product, the minimizing direction is $-d_t^P$. Letting $\delta h_t = -c\, d_t^P$ and enforcing the constraint gives

$$-c\|d_t^P\|_2^2 = -\Delta_t, \qquad c = \frac{\Delta_t}{\|d_t^P\|_2^2}, \tag{100}$$

which yields Eq. (97). □

Since $\delta h_t^* \in \mathrm{span}(P_t)$ and the prior subspace is orthogonal to the protected visual-evidence subspace,

$$\Pi_t^U \delta h_t^* = 0, \qquad \Pi_t^R \delta h_t^* = 0. \tag{101}$$

Therefore, the intervention modifies only the decision-relevant prior direction while leaving the protected visual-evidence and residual components unchanged.

B.3 Full-Vocabulary Consistency

The minimum-norm correction is derived from the local competing pair (y_t^P, y_t^V), whereas next-token decoding is performed over the entire vocabulary $\mathcal{V}$. A local pairwise correction may therefore unintentionally promote a third token that is not involved in the intervention objective.

To prevent such collateral changes, we apply a full-vocabulary consistency guard after obtaining

$$h_t' = \widetilde{h}_t + \delta h_t^*. \tag{102}$$

We evaluate the corrected top-1 prediction

$$\hat{y}_t' = \arg\max_{y \in \mathcal{V}} W_y^\top h_t'. \tag{103}$$

The candidate correction is accepted only if

$$\hat{y}_t' = y_t^V. \tag{104}$$

Otherwise, the update δh_t^* is rejected and the original strengthened state $\widetilde{h}_t$ is retained for decoding.

This guard ensures that an intervention derived from a local prior-visual conflict does not introduce an unintended vocabulary-level decision outside the competing pair.

C Pseudocode of SchröHallu

Algorithm 1 presents the complete inference procedure of Schrödinger Prior. At each decoding step, the method sequentially performs contextual evidence strengthening, evidence-preserving prior decomposition, contextual prior-utility measurement, and selective minimum-norm intervention. The procedure is deliberately conservative: the presence of a prior contribution alone never triggers modification. An intervention is considered only when the prior-favored and visually supported candidates disagree, the corresponding decision direction is sufficiently represented in the prior subspace, the contextual measurement resolves the prior toward the harmful state H, and the resulting harmful evidence satisfies the intervention-sufficiency condition. Otherwise, the current decoding state is preserved. When intervention is activated, the update is restricted to the decision-relevant direction within the prior subspace and is accepted only after a full-vocabulary consistency check. All operations are performed during inference with the original LVLM parameters frozen.

Algorithm 1 Schrödinger Prior for Context-Conditioned Intervention

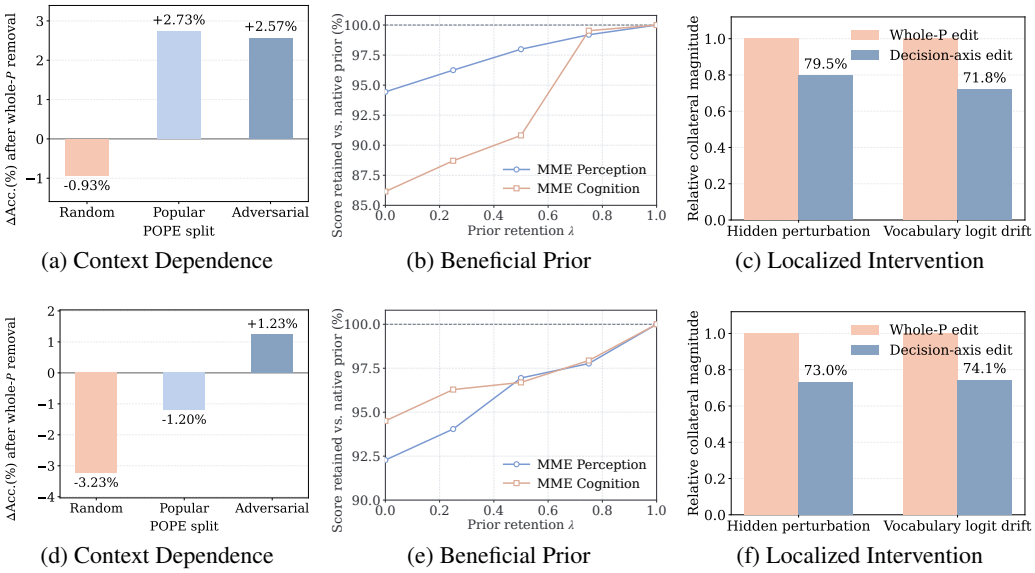
Input: LVLM f_θ, LM head W, image I, query Q, maximum length T
Output: Generated response $\mathbf{y}$

Initialize $\mathbf{y}_{<1} \leftarrow \varnothing$
for $t = 1, \ldots, T$ **do**
  Stage I: Contextual Evidence Strengthening
    Construct $\mathcal{C}_t = (I, Q, \mathbf{y}_{<t})$
    Reinforce visual evidence using Eq. (9)
    Obtain strengthened hidden state $\widetilde{h}_t$ and visual logits z_t^A
  Stage II: Evidence-Preserving Prior Decomposition
    Construct evidence/prior subspaces (U_t, P_t) and projectors $(\Pi_{U,t}, \Pi_{P,t})$
    Decompose $\widetilde{h}_t = h_{U,t} + h_{P,t} + h_{R,t}$ using Eq. (11)
  Stage III: Contextual Prior-Utility Measurement
    Obtain prior-favored candidate $y_{P,t}$ and visual candidate $y_{V,t}$
    Set $h_t^{\text{out}} \leftarrow \widetilde{h}_t$
    if $y_{P,t} \neq y_{V,t}$ **then**
      Construct decision direction $w_t = W_{y_{P,t}} - W_{y_{V,t}}$ and $d_{P,t} = \Pi_{P,t} w_t$
      if $d_{P,t}$ satisfies geometric sufficiency **then**
        Construct contextual evidence $\mathcal{E}_t = \{D_t, V_t, C_{P,t}, m_{U,t}, m_{A,t}\}$
        Measure $(p_{H,t}, p_{K,t}) \leftarrow \mathcal{M}(\mathcal{E}_t)$ and $q_{H,t} = p_{H,t} - p_{K,t}$
  Stage IV: Selective Minimum-Norm Intervention
        Compute $B_t = \min(m_{P,t}, m_{A,t})$
        if $q_{H,t} B_t > m_{T,t}$ **then**
          $\Delta_t \leftarrow q_{H,t} B_t$
          $\delta h_t^* \leftarrow -\dfrac{\Delta_t}{\|d_{P,t}\|_2^2} d_{P,t}$
          $h_t^{\text{cand}} \leftarrow \widetilde{h}_t + \delta h_t^*$
          if $\arg\max_y W_y^\top h_t^{\text{cand}} = y_{V,t}$ **then**
            $h_t^{\text{out}} \leftarrow h_t^{\text{cand}}$
          end if
        end if
      end if
    end if
    Generate $y_t \leftarrow \arg\max_y W_y^\top h_t^{\text{out}}$
    Append y_t to $\mathbf{y}$
    if $y_t = \text{EOS}$ **then**
      break
    end if
end for
return $\mathbf{y}$

D Additional Empirical Evidence Across LVLMs

To examine whether the empirical observations in Fig. 1 are specific to a single LVLM, we further repeat the same analyses on LLaVA-1.5-13B and DeepSeek-VL-Chat-7B. As shown in Fig. 4, both models exhibit the same three qualitative phenomena.

Benchmark rationale and scope. The empirical analyses in Fig. 4 intentionally use different benchmarks because the three observations probe different properties of model priors. For context dependence, we adopt POPE (Li et al., 2023b), whose Random, Popular, and Adversarial subsets share the same binary object-existence formulation but differ in how absent objects are selected. Random samples negatives without additional frequency bias, Popular emphasizes frequently occurring object categories, and Adversarial prioritizes absent objects that commonly co-occur with objects present in the image. These progressively different negative contexts expose the model to different degrees and forms of statistical object prior while retaining a controlled evaluation format. Consequently, applying the same whole-prior removal rule across the three subsets provides a natu-



(a) Context Dependence (b) Beneficial Prior (c) Localized Intervention

(d) Context Dependence (e) Beneficial Prior (f) Localized Intervention

Figure 4: **Additional empirical evidence on LLaVA-1.5-13B (a-c) and DeepSeek-VL-Chat-7B (d-f).** (a,d) Whole-prior removal produces context-dependent accuracy changes across POPE splits. (b,e) Stronger prior suppression degrades both perception and cognition on MME, confirming that priors also retain beneficial knowledge. (c,f) Decision-axis editing induces less hidden-state perturbation and vocabulary-logit drift than whole-prior editing. These results consistently support context-conditioned prior duality and localized intervention across different LVLMs.

ral diagnostic of whether the effect of prior suppression is stable across contexts. The observed sign changes in accuracy therefore support context-conditioned prior utility; importantly, they should be interpreted as evidence that the *effect of the same intervention mechanism* varies with context, rather than as a claim that an identical hidden prior vector is shared by all samples.

For beneficial-prior retention, we instead use MME (Fu et al., 2025), which evaluates LVLMs along complementary perception and cognition dimensions. This structure is well suited to testing whether progressively suppressing prior-related representations removes only hallucination-inducing information or also damages capabilities that depend on useful parametric knowledge. If priors were uniformly harmful, stronger suppression would be expected to improve or at least preserve these capabilities. In contrast, Fig. 4(b,e) shows that reducing the prior-retention factor λ degrades both MME Perception and Cognition across models. We therefore use MME here as a capability-retention stress test rather than as a direct hallucination benchmark: the result indicates that prior-related representations contain functionally useful information, but does not imply that every MME subtask depends on priors to the same extent.

The localized-intervention analysis addresses a third question that is largely independent of benchmark accuracy: given that an intervention is required, how much collateral change does it introduce into the model state? We therefore compare whole-prior and decision-axis editing using representation perturbation and full-vocabulary logit drift. These quantities directly measure the geometric and output-space side effects of an edit, rather than task success alone. Their consistent reduction across additional LVLMs shows that the advantage of decision-local editing is not merely an artifact of improved benchmark accuracy, but follows from modifying a substantially smaller decision-relevant portion of the prior representation.

Taken together, POPE and MME play complementary diagnostic roles rather than serving as interchangeable evaluation datasets. POPE provides controlled variation in hallucination-prone object contexts, MME tests whether useful perceptual and cognitive capabilities survive prior suppression, and the perturbation analysis directly evaluates the locality of representation editing. Accordingly, the three analyses jointly support the motivating claims of context-dependent prior utility, beneficial

prior retention, and localized intervention, while the broader POPE, CHAIR, AMBER, and MME evaluations in Sec. 5 assess the overall effectiveness of SchröHallu beyond these diagnostic settings.

Why not CHAIR or AMBER for the diagnostic analyses? CHAIR and AMBER remain important benchmarks in our main evaluation, but they are less suitable for isolating the specific mechanisms examined in Fig. 4. CHAIR evaluates object hallucination in free-form image captioning by measuring hallucinated object mentions at the sentence and instance levels. Because the generated captions may differ in length, lexical choice, and the set of objects mentioned, changing the prior can simultaneously alter both *what* the model describes and *how much* it generates. Consequently, an observed change in CHAIR may reflect not only a change in prior harmfulness, but also changes in generation dynamics or object coverage. More importantly, CHAIR does not provide controlled context variants analogous to the Random, Popular, and Adversarial splits of POPE. It is therefore well suited for measuring the final reduction of generative object hallucination, but less suitable for diagnosing whether the *same prior intervention* changes its functional effect across controlled contexts.

AMBER provides a broader evaluation by covering both generative and discriminative settings and multiple hallucination dimensions, including existence, attribute, and relation errors (Wang et al., 2023). This breadth makes it valuable for evaluating overall hallucination mitigation, but less appropriate for isolating a single causal factor in our empirical analysis. Its aggregate metrics, such as CHAIR, coverage, hallucination rate, and cognition, reflect several interacting generation and reasoning behaviors rather than a controlled variation of prior strength or contextual bias. Therefore, a performance change under prior suppression is harder to attribute specifically to the removal of beneficial prior knowledge. In contrast, MME explicitly separates perception- and cognition-oriented capabilities, allowing us to use the prior-retention coefficient λ as the only controlled intervention variable and directly examine whether progressively stronger suppression degrades these two capability dimensions.

Accordingly, we use POPE and MME for *mechanism-oriented diagnosis*, while CHAIR and AMBER are retained for *outcome-oriented evaluation*. The former provide more controlled settings for testing context dependence and capability retention, whereas the latter assess whether the resulting intervention ultimately reduces hallucinations under more realistic open-ended and multi-dimensional evaluation protocols. This distinction also prevents the motivating evidence from being inferred solely from the same metrics later used to report the final performance of SchröHallu.

Context-dependent prior effects. Figures 4(a) and 4(d) report the change in POPE accuracy after removing the entire prior component. For LLaVA-1.5-13B, whole-prior removal decreases accuracy on Random by 0.93%, but improves Popular and Adversarial by 2.73% and 2.57%, respectively. DeepSeek-VL-Chat-7B shows a different context profile: removal decreases Random and Popular by 3.23% and 1.20%, while improving Adversarial by 1.23%. Thus, even under the same intervention, the effect varies substantially in both sign and magnitude across evaluation contexts. This further supports our central observation that prior utility is context-conditioned rather than intrinsically harmful.

Retention of beneficial prior knowledge. Figures 4(b) and 4(e) vary the prior-retention factor λ on MME, where $\lambda = 1$ corresponds to the native prior and smaller values impose stronger suppression. For both models, reducing λ progressively degrades perceptual and cognitive performance. The degradation is particularly pronounced under aggressive suppression, showing that prior-related representations also contain information useful for visual perception and reasoning. These results further explain why indiscriminate prior removal can improve hallucination-related decisions in some contexts while simultaneously damaging useful model capabilities.

Benefits of decision-local intervention. Figures 4(c) and 4(f) compare whole-prior editing with the proposed decision-axis editing. Taking whole-prior editing as the reference magnitude, decision-axis editing reduces hidden-state perturbation to 79.5% and vocabulary-logit drift to 71.8% on LLaVA-1.5-13B. The corresponding values on DeepSeek-VL-Chat-7B are 73.0% and 74.1%. Therefore, restricting the intervention to the decision-relevant direction consistently introduces less collateral change than modifying the complete prior component.

Table 4: Ablation study (%) on the CHAIR dataset with LLaVA-1.5 7B.

Visual	Quantum	$C_S \downarrow$	$C_I \downarrow$	F1 ↑
×	×	48.0	14.3	75.6
✓	×	30.2	7.7	75.6
×	✓	42.6	11.8	75.6
✓	✓	**21.8**	**6.5**	**75.7**

Table 5: Ablation study (%) of the key components in SchröHallu.

Method	$C_S \downarrow$	$C_I \downarrow$	F1 ↑
w/o U+A evidence consensus	33.2	7.9	75.0
w/o prior-axis concentration certificate	38.2	10.7	75.6
w/o contextual H/K polarization	32.6	7.8	75.6
Full	**21.8**	**6.5**	**75.7**

Table 6: Performance comparison of different models.

Model	Oracle-resolvable steps	Sign Acc. ↑	AUROC ↑	$F1_H$ ↑	Spearman $\rho(q_t^H, -\bar{u}_t^P)$ ↑
LLaVA-1.5-7B	6,214	84.7	0.912	84.3	0.63
LLaVA-1.5-13B	6,087	85.5	0.921	85.0	0.66
DeepSeek-VL-7B	5,934	82.8	0.897	82.2	0.59

Overall, these additional results reproduce the three observations underlying SchröHallu across model scales and architectures: prior effects are context-dependent, priors retain beneficial knowledge, and localized intervention causes less collateral perturbation. Together with the main-text evidence, they provide further empirical support for resolving prior utility from the current context before applying a selective intervention.

E Discussion

Does q_t^H predict oracle prior utility? We directly evaluate whether the contextual polarization produced by SchröHallu agrees with the oracle prior utility defined in Sec. 3.1. As shown in Table 6, For each oracle-resolvable decision, we define the harmful label as $z_t^H = \mathbb{I}[u_t^P < 0]$ and predict it using $\hat{z}_t^H = \mathbb{I}[q_t^H > 0]$. Across all three LVLMs, q_t^H achieves strong sign agreement and AUROC, and is positively correlated with the magnitude of oracle harmful utility. This provides direct evidence that the deterministic contextual measurement captures the functional role of the prior rather than merely its presence or magnitude.

Controlled context-pair analysis. We further construct counterfactual image pairs that preserve the prompt, decoding prefix, decision direction, and a fixed language-only semantic prior while changing only whether the queried object is visually present. As shown in Table 7, this produces controlled $K \to H$ and $H \to K$ utility transitions. Importantly, the experiment does not merely verify the oracle sign change: we also examine whether the independently computed contextual polarization q_t^H follows the corresponding utility transition. Across the three models, q_t^H follows the oracle flip in approximately 80-84% of paired cases, providing direct evidence that prior utility is resolved by the surrounding visual context.

Failure analysis when the visual candidate is incorrect. The visually strengthened candidate y_t^V is not assumed to be an oracle target and can itself be incorrect. We therefore quantify the potentially harmful case where the prior-favored candidate is correct but y_t^V is incorrect. As shown in Table 8, although such conflicts occur non-negligibly, only a small fraction pass the disagreement, geometric, polarization, and sufficiency gates. Consequently, fewer than 0.3% of all evaluated decisions result in an accepted harmful edit, while more than 95% of accepted interventions favor the oracle-correct candidate. Thus, the conservative decision funnel substantially reduces, but does not completely eliminate, errors caused by an incorrect visual candidate.

Decision-funnel statistics. To characterize how selectively SchröHallu intervenes, we record the fraction of decoding steps surviving each gate on CHAIR. As shown in Table 9, only about one third of steps exhibit disagreement between the prior-favored and visually strengthened candidates. Geometric sufficiency, contextual H/K measurement, and intervention sufficiency progressively reduce this set, and only approximately 5-6% of all decoding steps are ultimately edited. Moreover, roughly

Table 7: Results on paired flip cases.

Model	Paired cases	Oracle $K \leftrightarrow H$	q_t^H follows flip
LLaVA-1.5-7B	300	100%	82.4%
LLaVA-1.5-13B	300	100%	84.1%
DeepSeek-VL-7B	300	100%	80.3%

Table 8: Results on risky conflicts and accepted edits.

Model	y_t^V wrong among conflicts	Accepted within risky conflicts	Harmful edits / all steps	Accepted-edit precision ↑
LLaVA-1.5-7B	9.8%	4.5%	0.18%	96.7%
LLaVA-1.5-13B	8.7%	4.0%	0.15%	97.1%
DeepSeek-VL-7B	10.9%	5.2%	0.24%	95.9%

87% of edits that pass the intervention-sufficiency condition survive the final full-vocabulary consistency check. These statistics confirm that SchröHallu operates as a sparse, conservative intervention rather than continuously modifying the decoding trajectory.

Is the quantum terminology computationally necessary? No. The quantum-inspired formulation provides a structured interpretation of unresolved prior utility, but does not introduce an additional learned module or stochastic quantum computation. If $|H\rangle$ and $|K\rangle$ are simply renamed as harmful and beneficial functional roles while retaining the same deterministic measurement $q_t = D_t V_t C_t^P S_t$ and the same intervention rule, the algorithm is functionally identical and therefore produces exactly the same outputs and performance.

What is lost is not predictive performance but the conceptual formalization: the superposition/measurement view provides a unified language for expressing why prior utility is unresolved before contextual interaction and why intervention should occur only after sufficient evidence resolves its functional role. In contrast, removing the contextual polarization mechanism itself changes the computation and leads to degraded hallucination mitigation, as shown by our component ablation.

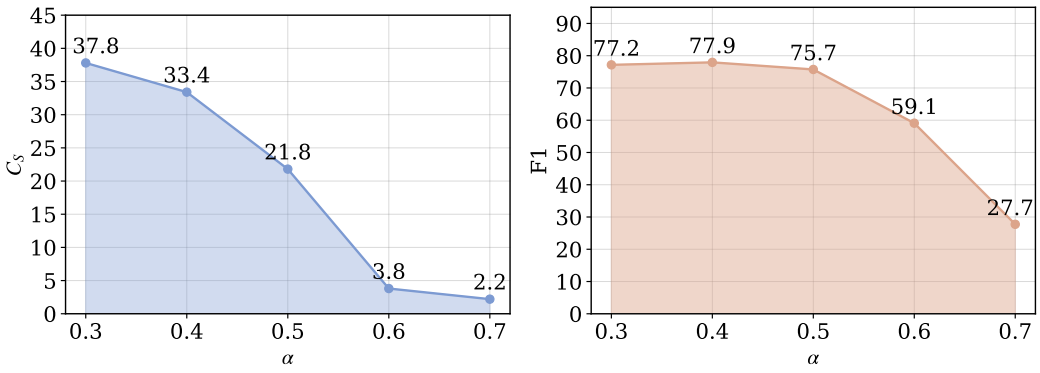
F Ablation Studies

Effect of Main Modules. Table 4 evaluates the contribution of the visual and quantum modules on CHAIR with LLaVA-1.5-7B. Introducing the visual module alone substantially reduces C_S (30.2% vs. 48.0%) and C_I (7.7% vs. 14.3%) while maintaining the original F1 (75.6% vs. 75.6%), confirming that strengthening contextual visual evidence effectively suppresses unsupported generation without sacrificing object coverage. The quantum module alone also improves C_S (42.6% vs. 48.0%) and C_I (11.8% vs. 14.3%), demonstrating the benefit of context-conditioned prior intervention. Combining both modules achieves the best overall performance, further reducing C_S to 21.8% and C_I to 6.5% while slightly improving F1 to 75.7%. These results confirm that contextual evidence strengthening and prior-utility intervention provide complementary benefits in SchröHallu.

Effect of Key Components. Table 5 further investigates the key components involved in contextual prior measurement. Removing the U+A evidence consensus increases C_S (33.2% vs. 21.8%) and reduces F1 (75.0% vs. 75.7%), indicating that agreement between independent visual evidence is important for reliable prior-state resolution. Removing the prior-axis concentration certificate causes a pronounced degradation in both C_S (38.2% vs. 21.8%) and C_I (10.7% vs. 6.5%), showing that verifying whether prior influence is concentrated along the active decision direction is essential for selective intervention. Similarly, removing contextual H/K polarization increases C_S (32.6% vs. 21.8%) and C_I (7.8% vs. 6.5%). The full configuration consistently provides the strongest hallucination reduction, validating the necessary and complementary roles of evidence consensus, geometric certification, and context-conditioned prior resolution.

Table 9: Edit acceptance statistics under different filtering conditions.

Model	$y_t^P \neq y_t^V$	Geom. sufficient	$q_t^H > 0$	$q_t^H B_t > m_t^T$	Accepted edit
LLaVA-1.5-7B	33.2%	24.8%	14.7%	7.0%	6.1%
LLaVA-1.5-13B	30.9%	23.1%	13.5%	6.3%	5.5%
DeepSeek-VL-7B	35.6%	25.7%	15.4%	7.4%	6.4%



(a) Effect of α on C_S

(b) Effect of α on F1

Figure 5: Parameter sensitivity analysis of α on CHAIR with LLaVA-1.5 7B. (a) C_S under different values of α. (b) F1 under different values of α.

G Parameter Analysis

Sensitivity to α. Figures 5a and 5b investigate the effect of the visual evidence reinforcement coefficient α on CHAIR with LLaVA-1.5-7B. As α increases, C_S decreases consistently, whereas an excessively large α substantially degrades F1. In particular, increasing α from 0.4 to 0.5 markedly reduces C_S (21.8% vs. 33.4%) while maintaining a competitive F1 (75.7% vs. 77.9%). Further increasing α to 0.6 yields a lower C_S (3.8% vs. 21.8%) but causes a pronounced F1 drop (59.1% vs. 75.7%), which becomes more severe at $\alpha = 0.7$ (27.7% vs. 59.1%). These results indicate that insufficient reinforcement provides limited contextual evidence, whereas overly strong reinforcement disturbs the balance of the original multimodal representation. Therefore, we set $\alpha = 0.5$ in all experiments, which provides a favorable trade-off between hallucination reduction and generation quality.

Sensitivity to subspace ranks. We further vary the evidence rank r_U and prior rank r_P while keeping all other settings fixed. As shown in Table 10, performance remains stable over a relatively broad range of ranks, with $r_U = 8$ and $r_P = 5$ providing the best overall trade-off. Very small ranks under-represent visual or prior structure, whereas larger ranks introduce additional irrelevant directions without yielding consistent gains. These results indicate that SchröHallu is not sensitive to a narrow choice of subspace dimensionality.

H Detailed Experimental Setup

Models. We evaluate SchröHallu on three representative LVLMs: LLaVA-1.5-7B, LLaVA-1.5-13B (Liu et al., 2023), and DeepSeek-VL-Chat-7B (Lu et al., 2024). The two LLaVA variants allow us to examine the behavior of the proposed intervention across different model scales within the same architecture, while DeepSeek-VL-Chat further evaluates its transferability to a different multimodal backbone. Unless otherwise specified, all experiments directly use the publicly released checkpoints without additional training or parameter updates.

Table 10: Ablation study on r_U and r_P.

r_U	r_P	POPE Avg. Acc. ↑	CHAIR C_S ↓	CHAIR C_I ↓
4	3	87.0	23.6	7.0
8	3	87.4	22.5	6.7
8	**5**	**87.6**	**21.8**	**6.5**
12	5	87.5	22.0	6.6
8	8	87.4	22.3	6.7
12	8	87.3	22.6	6.8

Benchmarks and Metrics. We evaluate hallucination and multimodal capabilities on four widely used benchmarks: POPE, CHAIR, MME, and AMBER. For discriminative object hallucination, we adopt POPE (Li et al., 2023b) and evaluate its Random, Popular, and Adversarial subsets, reporting Accuracy and F1. For open-ended image captioning, we use CHAIR (Rohrbach et al., 2018) and report sentence-level C_S and instance-level C_I, where lower CHAIR scores indicate fewer hallucinated object mentions. We further evaluate object-level perceptual capabilities on MME (Fu et al., 2025), focusing on the Existence, Count, Position, and Color subsets. Finally, we use AMBER (Wang et al., 2023) for a more comprehensive hallucination evaluation and report CHAIR, Cover, Hal, and Cog following its standard protocol. Together, these benchmarks cover binary object-presence judgments, free-form image descriptions, fine-grained visual perception, and comprehensive generative hallucination evaluation.

Baselines. We compare SchröHallu with the original decoding of each LVLM and representative training-free hallucination mitigation methods spanning different intervention mechanisms. Specifically, we include OPERA (Huang et al., 2024), Nullu (Yang et al., 2025a), SVAR (Jiang et al., 2025), VTI (Liu et al., 2025), DeCo (Wang et al., 2025), HulluEdit (Lin et al., 2026), SHIELD (Huang et al., 2026), and LVLMs-Saliency (Zhang et al., 2026b).

Implementation Details SchröHallu operates entirely at inference time and does not update any model parameters. We adopt *greedy decoding* for all experiments and keep the prompt, decoding configuration, and evaluated samples identical across methods on the same benchmark whenever possible. For CHAIR, we use the same image-captioning prompt and generation protocol across all compared methods. The layers used for contextual evidence strengthening are determined according to the visually informative layer range of each backbone, while the corresponding reinforcement strength is fixed after selection. No test annotations are used for hyperparameter selection. All experiments are conducted on a single H200 GPU.

Hyperparameters. We intentionally keep the number of tunable hyperparameters small. For contextual evidence strengthening, we use $\alpha = 0.5$ and apply attention reinforcement over the relative-depth interval $[0.16L, 0.56L]$, corresponding to layers 5-18 for a 32-layer backbone. For subspace construction, we use an evidence rank of $r_U = 8$ and a prior rank of $r_P = 5$, with the representation anchor located at $l_a = \lfloor 0.8125L \rfloor$ (layer 26 for a 32-layer backbone). We use a single numerical constant $\epsilon = 10^{-6}$ throughout all normalization operations. The geometric sufficiency of a candidate pair is measured by

$$G_t = \frac{\|\Pi_{P,t} w_t\|_2}{\|w_t\|_2 + \epsilon}, \tag{105}$$

and we require $G_t \geq \tau_g$, with $\tau_g = 0.10$. No additional hard thresholds are introduced for fidelity, pair visibility, prior-axis concentration, or the H/K measurement; these quantities contribute continuously to the contextual polarization. Moreover, the target correction margin is not a tuned constant but is computed online as

$$m_{T,t} = [w_t^\top \widetilde{h}_t]_+. \tag{106}$$

Online construction of U_t and P_t. Following Lin et al. (2026), the subspaces are constructed online from cached visual and textual representations rather than precomputed dataset-level directions. At the anchor layer l_a, the visual-token matrix $V = [v_1, \ldots, v_{n_v}] \in \mathbb{R}^{d \times n_v}$ is extracted once for each input image and cached throughout decoding. At step t, given the current anchor-layer state g_t, we compute token-wise visual relevance

$$a_{t,i} = \mathrm{softmax}_i \left(\frac{v_i^\top g_t}{\|v_i\|_2 \|g_t\|_2 + \epsilon} \right). \tag{107}$$

Let $A_t = \mathrm{diag}(a_t)$. We perform truncated SVD

$$V A_t^{1/2} = L_t^U \Sigma_t^U (R_t^U)^\top, \qquad U_t = L_t^U[:, 1 : r_U]. \tag{108}$$

Meanwhile, a sliding text cache $T_t \in \mathbb{R}^{d \times n_t}$ stores non-visual hidden states from recent decoding steps. To prevent prior directions from overlapping with protected visual evidence, we first project the cache onto the orthogonal complement of $\mathrm{span}(U_t)$:

$$\bar{T}_t = (I - U_t U_t^\top) T_t. \tag{109}$$

A second truncated SVD gives

$$\bar{T}_t = L_t^P \Sigma_t^P (R_t^P)^\top, \qquad P_t = L_t^P[:, 1 : r_P]. \tag{110}$$

Thus, $U_t^\top P_t = 0$ by construction. The visual-token matrix is cached once per image, whereas the text cache is incrementally updated and the low-rank bases U_t and P_t are re-estimated at each decoding step.

I Limitations and Future Work

Despite its effectiveness, SchröHallu has several limitations. First, our evaluation covers three representative LVLMs and four hallucination/perception benchmarks, leaving its behavior on larger-scale proprietary models, more diverse architectures, and long-form multimodal reasoning to be further examined. Second, the contextual measurement pathway introduces additional inference cost compared with vanilla decoding. Moreover, the reliability of intervention depends on the quality of the visual-evidence/prior decomposition and the local competing-token pair; imperfect subspace identification or ambiguous visual evidence may therefore weaken the contextual resolution of prior utility.

Future work will investigate broader model families and multimodal tasks, more efficient or selectively activated contextual measurement, and more robust prior-evidence decomposition. Another promising direction is to extend the current token-level decision-local intervention toward sequence-level or semantic-unit reasoning, enabling prior utility to be resolved over longer generation trajectories while preserving the training-free nature of the framework.