OMNICONFESS: ELICITING TOKEN CONFESSIONS TO MITIGATE OMNI-MODAL HALLUCINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal large language models (OmniLLMs) unify text, images, audio, and video, yet hallucinate when generation relies on the wrong evidence. Existing inference-time methods can reduce hallucinations, but rarely reveal which evidence sustains a generated commitment. We introduce **OmniConfess**, a training-free method for mitigating omni-modal hallucinations. It fixes a candidate response and re-scores it at token resolution under controlled channel-wise evidence interventions, producing a structured token-by-channel confession that reveals the response's evidential dependence. OmniConfess uses this confession to preserve grounded content and correct commitments driven by irrelevant or contradictory evidence. To evaluate OmniConfess, we construct **OmniHalluBench**, a 3,540-example benchmark built from six datasets spanning text, image, audio, and video settings and both judgment and free-form generation. Experiments show that OmniConfess mitigates hallucinations across heterogeneous modality and task settings. Our code and benchmark are publicly available[1].

1 INTRODUCTION

Omni-modal large language models (OmniLLMs) unify text, images, audio, and video in a single generative model, enabling reasoning over heterogeneous evidence. However, this integration can cause omni-modal hallucination when the model relies on irrelevant evidence or fails to sufficiently use relevant evidence (Leng et al., 2024; Sung-Bin et al., 2025) (see Figure 1). Hallucination control therefore requires assessing both the correctness of the final output and the evidence supporting each generated commitment.

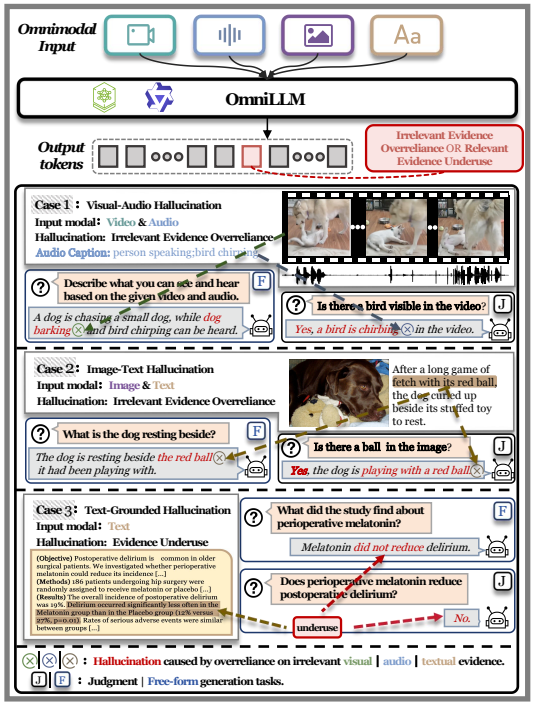


Figure 1: Examples of OmniLLM hallucination.

Existing hallucination-control methods include retrieval augmentation (Qu et al., 2024; Xia et al., 2024), alignment (Yu et al., 2024; Sun et al., 2023), and fine-tuning (Chaubey et al., 2026), often requiring additional training or external resources. Training-free inference-time methods instead follow three paradigms, as illustrated in Figure 2: search-based methods explore or verify candidate generations (Yao et al., 2023; Xu et al., 2025a); evidence-contrast methods steer decoding through evidence-induced distribution shifts (Leng et al., 2023; Jung et al., 2025); and fine-grained methods exploit token-level signals to localize hallucinations (Zhu et al., 2026; Ouyang et al., 2026).

[1] https://anonymous.4open.science/status/OmniConfess-A60C

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.

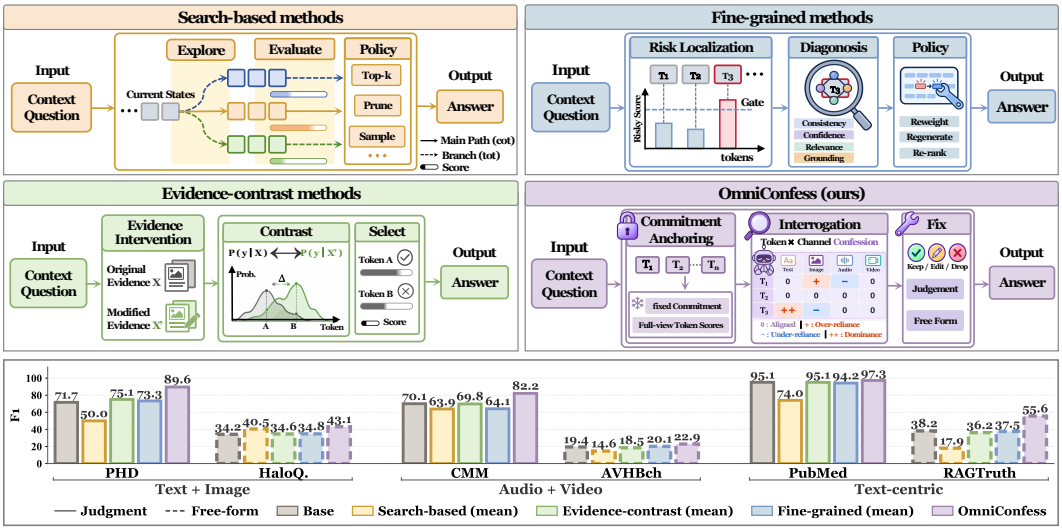


Figure 2: Comparison of three existing inference-time paradigms and OmniConfess, with category-mean F1 across six datasets using Qwen2.5-Omni-7B.

However, three main challenges remain when applied to hallucination control in OmniLLMs. **(i) Trajectory variation.** Regeneration after evidence intervention alters the autoregressive context, confounding evidence effects with sequence differences (Li et al., 2025). **(ii) Entangled cross-channel dependence.** OmniLLMs jointly process text, images, audio, and video, obscuring channel contributions to individual tokens (Parcalabescu & Frank, 2023). **(iii) Insufficient token-level resolution.** Evidential dependence varies across tokens within each channel, making global or modality-level scores insufficient to identify revision targets (Chen et al., 2026). Reliable characterization thus requires token-level channel-wise comparisons along a frozen candidate sequence.

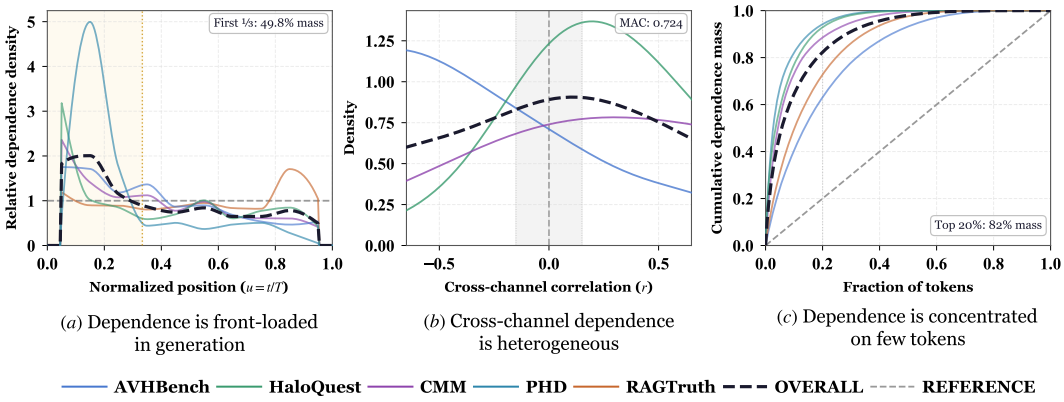
To address these challenges, we propose **`OmniConfess`**, an inference-time hallucination correction method for OmniLLMs, with three key designs: **(1) Commitment Anchoring.** We generate and freeze a candidate response to eliminate trajectory variation across evidence interventions. **(2) Evidence Interrogation.** We intervene on individual evidence channels and re-score the frozen candidate to isolate token-level, channel-wise dependence. **(3) Confession-Guided Correction.** We organize these measurements into a token-by-channel confession to recalibrate judgment scores or locate and revise erroneous free-form spans.Figure 2 contrasts the three existing inference-time paradigms with the commitment-centered workflow of OmniConfess.

To evaluate OmniConfess, we construct **`OmniHalluBench`** with 3,540 hard examples from six public datasets, spanning text, images, audio, and video across judgment and free-form generation tasks, including cross-modal conflicts and text-grounded hallucinations. OmniConfess outperforms the strongest relevant baseline across all six datasets on the primary backbone. Ablations further validate the importance of the confession's token-specific and channel-specific structure.

2 Related Work

Omni-Modal Hallucination. As foundation models evolve from vision-language systems toward unified processing of text, images, audio, and video (Xu et al., 2025b;c; NVIDIA, 2026), hallucination research has expanded from visual grounding failures to encompass cross-channel interference in omni-modal settings (Li et al., 2023; Guan et al., 2024; Sung-Bin et al., 2025; Leng et al., 2024). Recent studies on multimodal hallucination examine modality contributions to generation, differences in cross-modal grounding, and evidential attribution of generated content (Parcalabescu & Frank, 2023; Kim et al., 2026; Yan et al., 2026).

Inference-Time Hallucination Mitigation. Existing inference-time methods mainly follow three paradigms. Search-based methods explore, verify, or select candidates for reliability (Wang et al., 2023; Yao et al., 2023; Xu et al., 2025a). Evidence-contrast methods exploit evidence-induced shifts for decoding, from visual contrast to audio-visual and modality-adaptive decoding (Shi et al., 2023; Leng et al., 2023; Favero et al., 2024; Jung et al., 2025; Chung et al., 2026; Dong et al., 2026).



(a) Dependence is front-loaded in generation

(b) Cross-channel dependence is heterogeneous

(c) Dependence is concentrated on few tokens

Figure 3: Three structural properties of evidential dependence in hallucinated OmniLLM outputs.

Fine-grained methods mitigate hallucinations via token-level or internal signals, including attention patterns, risk, and cross-modal representations (Huang et al., 2024; Zhu et al., 2026; Ouyang et al., 2026; Li et al., 2026). In this work, we introduce OmniConfess to explicitly characterize token-level, channel-wise evidential dependencies in OmniLLMs for hallucination correction.

3 Preliminary

We examine three properties of evidential dependence in hallucinated OmniLLM outputs: front-loaded dependence, cross-channel heterogeneity, and token-level concentration (Figure 3).

For dataset d, let $\mathcal{H}_d$ denote its hallucinated samples. Each $h \in \mathcal{H}_d$ has input X_h, evidence channels $\mathcal{M}_h$, and fixed candidate $y_h = (y_{h,1}, \ldots, y_{h,n_h})$ of length n_h. Let θ denote the OmniLLM parameters. Token i's dependence on channel $m \in \mathcal{M}_h$ is

$$\delta_{h,i}^{(m)} = \log \frac{p_\theta(y_{h,i} \mid y_{h,<i}, X_h)}{p_\theta(y_{h,i} \mid y_{h,<i}, X_h^{-m})}, \tag{1}$$

where p_θ is the OmniLLM conditional token distribution, $y_{h,<i}$ is the fixed prefix, and X_h^{-m} removes channel m. The overall token dependence magnitude is $u_{h,i} = \sum_{m \in \mathcal{M}_h} |\delta_{h,i}^{(m)}|$.

Front-Loaded Dependence. We partition normalized token positions i/n_h into B equal intervals $I_b = ((b-1)/B, b/B]$. The relative dependence in interval b is

$$D_d(b) = \frac{1}{|\mathcal{H}_d|} \sum_{h \in \mathcal{H}_d} \frac{\sum_{i:\, i/n_h \in I_b} u_{h,i}}{B^{-1} \sum_{i=1}^{n_h} u_{h,i}}, \qquad b = 1, \ldots, B. \tag{2}$$

Figure 3(a) shows the first third holds 49.8% of dependence mass. Early token changes alter later contexts, motivating anchoring to separate evidence and trajectory effects.

Cross-Channel Heterogeneity. We measure coupling between distinct channels $m, m' \in \mathcal{M}_h$ through the correlation of their token-level dependence trajectories:

$$r_h^{(m,m')} = \frac{\langle \widetilde{\mathbf{d}}_h^{(m)}, \widetilde{\mathbf{d}}_h^{(m')} \rangle}{\|\widetilde{\mathbf{d}}_h^{(m)}\|_2 \|\widetilde{\mathbf{d}}_h^{(m')}\|_2}, \tag{3}$$

where $\widetilde{\mathbf{d}}_h^{(m)} = (\delta_{h,i}^{(m)} - \bar{\delta}_h^{(m)})_{i=1}^{n_h}$ and $\bar{\delta}_h^{(m)}$ is the mean dependence. Figure 3(b) shows coupled channels with heterogeneous directions, motivating channel-specific attribution.

Token-Level Concentration. The mass captured by the top q fraction of tokens is

$$C_d(q) = \frac{1}{|\mathcal{H}_d|} \sum_{h \in \mathcal{H}_d} \frac{\sum_{j=1}^{\lceil q n_h \rceil} u_{h,(j)}^{\downarrow}}{\sum_{i=1}^{n_h} u_{h,i}}, \qquad q \in [0, 1]. \tag{4}$$

With $u_{h,(j)}^{\downarrow}$ denoting the j-th largest token dependence magnitude. Figure 3(c) shows that the top 20% of tokens capture 82% of the dependence mass on average, motivating token-level localization.

Together, these findings motivate OmniConfess's fixed-candidate anchoring, channel-wise intervention, and token-level rescoring for reliable evidence attribution.

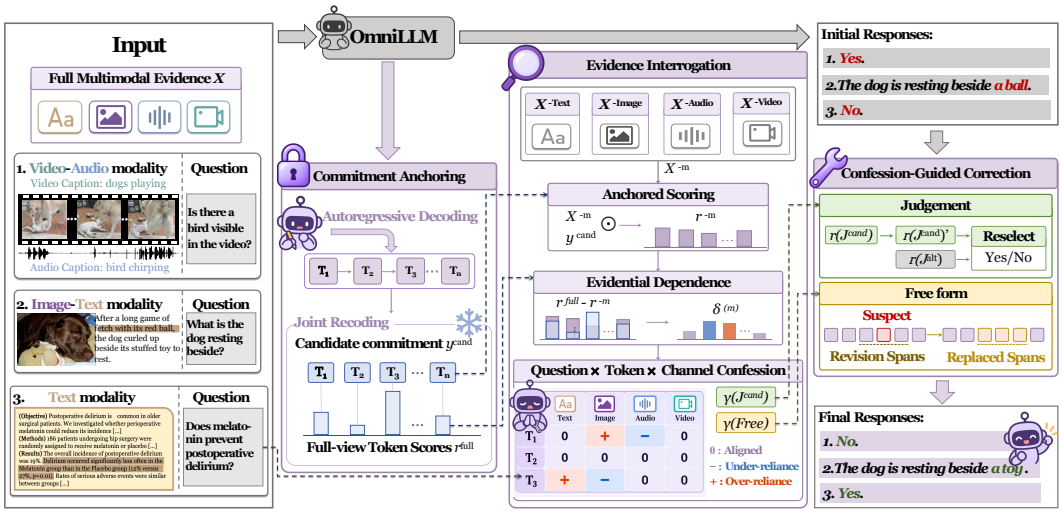


Figure 4: Overview of the OmniConfess framework: commitment anchoring, token-level evidence interrogation, and confession-guided correction for reliable omni-modal outputs.

4 METHODOLOGY: OMNICONFESS

We introduce OmniConfess (Figure 4), comprising Commitment Anchoring, Evidence Interrogation, and Confession-Guided Correction for training-free inference-time hallucination mitigation.

4.1 PROMPTING SETUP

Before the OmniConfess pipeline, we combine evidence-grounded prompting with task-specific constraints. Prompts enforce reliance on provided evidence and appropriate output formats, while task-specific instructions require audible-only judgments in **CMM**, "Yes/No/Maybe" in **PubMedQA**, acknowledgment of absent visual details in **HaloQuest**, and separate audio-visual observations in **AVHBench**.Full generation and correction prompts are provided in Appendix A.1.

4.2 COMMITMENT ANCHORING

For full multimodal evidence $\mathbf{X} = (X^{(1)}, \ldots, X^{(M)})$, OmniConfess records selected-token log-probabilities during autoregressive decoding, obtaining the candidate and full-view scores:

$$\left(y^{\text{cand}}_{1:n}, \mathbf{r}^{\text{full}}_{\theta}\right) \triangleq \left(\text{Decode}_{\theta}(\mathbf{X}), \left[\log p_{\theta}\left(y^{\text{cand}}_{i} \mid y^{\text{cand}}_{<i}, \mathbf{X}\right)\right]_{i=1}^{n}\right) \in \mathcal{V}_{\text{tok}}^{n} \times \mathbb{R}^{n}. \tag{5}$$

where M counts evidence channels $X^{(m)}$, θ denotes OmniLLM parameters, p_{θ} its token distribution, Decode_{θ} autoregressive decoding, $\mathcal{V}_{\text{tok}}$ the vocabulary, n candidate length, and i token position. The candidate $y^{\text{cand}}_{1:n} \equiv y^{\text{cand}}$ has selected token y^{cand}_{i}, prefix $y^{\text{cand}}_{<i}$, and full-view log-probabilities $\mathbf{r}^{\text{full}}_{\theta}$.

The candidate and its prefixes are frozen during interrogation. OmniConfess retains only y^{cand}, $\mathbf{r}^{\text{full}}_{\theta}$, and fixed alternatives a_i with scores: the opposing class for judgment and the top nonselected token for free-form generation. Under altered evidence $\mathbf{X}'$, it computes anchored scores:

$$\mathbf{r}_{\theta}\left(\mathbf{X}'; y^{\text{cand}}\right) \triangleq \left[\log p_{\theta}\left(y^{\text{cand}}_{i} \mid y^{\text{cand}}_{<i}, \mathbf{X}'\right)\right]_{i=1}^{n} \in \mathbb{R}^{n}. \tag{6}$$

where $\mathbf{r}_{\theta}(\mathbf{X}'; y^{\text{cand}})$ denotes the frozen candidate's token-score vector under altered evidence.

Proposition 1. *Motivated by Figure 3(a), commitment anchoring eliminates trajectory variation from token-level comparisons across evidence conditions.*

Proof. We provide experimental results in Sections 5.3 and proofs in Appendix B.1.

4.3 EVIDENCE INTERROGATION

OmniConfess interrogates frozen commitments through channel-wise evidence removal.

Anchored Scoring. OmniConfess reuses full-view scores and sequentially rescores the candidate and alternatives under $\mathbf{X}^{-m}$, with channel m removed. The candidate-to-alternative margin is

$$F_i(\mathbf{Z}) \triangleq \log \frac{p_{\theta}(y^{\text{cand}}_{i} \mid y^{\text{cand}}_{<i}, \mathbf{Z})}{p_{\theta}(a_i \mid y^{\text{cand}}_{<i}, \mathbf{Z})}, \quad \mathbf{Z} \in \{\mathbf{X}\} \cup \{\mathbf{X}^{-m}\}_{m=1}^{M}. \tag{7}$$

Evidential Dependence. For $v \in \{y_i^{\text{cand}}, a_i\}$, define token dependence and margin contribution:

$$\begin{aligned} \delta_i^{(m)}(v) &\triangleq \log \frac{p_\theta(v \mid y_{<i}^{\text{cand}}, \mathbf{X})}{p_\theta(v \mid y_{<i}^{\text{cand}}, \mathbf{X}^{-m})}, \\ \kappa_i^{(m)} &\triangleq F_i(\mathbf{X}) - F_i(\mathbf{X}^{-m}). \end{aligned} \tag{8}$$

Question × Token × Channel Confession. Task rules define channel relevance $r_m(Q,T) \in \{0,1\}$ for question Q and task type T, with $\mathcal{R} = \{m : r_m(Q,T) = 1\}$ and $\mathcal{I} = \{1,\dots,M\}\backslash\mathcal{R}$. Relevance and contribution thresholds assign diagnostic states. The structured confession is

$$\mathbf{C}_\theta(Q,T) \triangleq \left[\left(\delta_i^{(m)}(y_i^{\text{cand}}), \kappa_i^{(m)}, r_m(Q,T), \ell_i^{(m)}\right)\right]_{\substack{i=1,\dots,n \\ m=1,\dots,M}}. \tag{9}$$

where $\ell_i^{(m)} \in \{0,-,+,++\}$ denotes aligned, under-reliant, over-reliant, or dominant dependence, respectively. For retained channels $\mathcal{S} \subseteq \{1,\dots,M\}$, the additive surrogate is

$$\widehat{F}_i(\mathcal{S}) \triangleq F_i(\mathbf{X}) - \sum_{m \notin \mathcal{S}} \kappa_i^{(m)} = \widehat{F}_i(\varnothing) + \sum_{m \in \mathcal{S}} \kappa_i^{(m)}. \tag{10}$$

where $\widehat{F}_i$ is the surrogate margin. Set $P_i^{\varnothing} = \widehat{F}_i(\varnothing)$, $E_i^{\mathcal{R}} = \sum_{m\in\mathcal{R}} \kappa_i^{(m)}$, and $E_i^{\mathcal{I}} = \sum_{m\in\mathcal{I}} \kappa_i^{(m)}$, denoting residual preference, relevant evidence, and irrelevant evidence, respectively.

Task-specific Confession. The resulting judgment correction and free-form token signals are

$$\begin{aligned} \gamma(J^{\text{cand}}) &\triangleq E_d^{\mathcal{R}} - F_d(\mathbf{X}) = -P_d^{\varnothing} - E_d^{\mathcal{I}}, \\ \gamma_i^{\text{free}} &\triangleq \left(E_i^{\mathcal{R}},\ \left[[P_i^{\varnothing}]_+ - [E_i^{\mathcal{R}}]_+\right]_+,\ \sum_{m\notin\mathcal{R}} [\kappa_i^{(m)}]_+ \right). \end{aligned} \tag{11}$$

where J^{cand} is the candidate class, d its decision position, and $[x]_+ = \max(x, 0)$.

Proposition 2. *Motivated by Figure 3(b), the channel-wise surrogate preserves observed margins and yields task-specific evidence signals.*

Proof. We provide experimental results in Sections 5.3 and proofs in Appendix B.2.

4.4 Confession-Guided Correction

OmniConfess applies task-specific signals to judgment recalibration and local free-form correction.

Judgment Commitment Correction. For binary judgment, the candidate and its fixed alternative form $\mathcal{Y} = \{J^{\text{cand}}, a_d\}$. OmniConfess adjusts only the candidate-class score:

$$\widetilde{z}_d(c) \triangleq z_d(c) + \mathbb{I}[c = J^{\text{cand}}]\gamma(J^{\text{cand}}), \quad c \in \mathcal{Y}. \tag{12}$$

where $z_d(c) = \log p_\theta(c \mid y_{<d}^{\text{cand}}, \mathbf{X})$ is the full-view class score and $\mathbb{I}[\cdot]$ the indicator function. The final judgment is $J^{\text{final}} \triangleq \arg\max_{c\in\mathcal{Y}} \widetilde{z}_d(c)$, retaining the candidate in case of a tie.

Free-Form Local Repair. For free-form generation, OmniConfess combines the three confession signals into token-level revision risk:

$$\rho_i \triangleq [-\gamma_{i,1}^{\text{free}}]_+ + \gamma_{i,2}^{\text{free}} + \gamma_{i,3}^{\text{free}}, \quad i = 1,\dots,n. \tag{13}$$

where $\gamma_{i,k}^{\text{free}}$ denotes component k of the free-form confession. Tokens exceeding threshold τ are expanded to complete word boundaries and merged into contiguous revision spans:

$$\mathcal{B}_{\text{rev}} \triangleq \text{Merge}(\text{Expand}\left(\{i : \rho_i > \tau\}\right)) = \{[s_j, e_j)\}_{j=1}^{J}. \tag{14}$$

where $\mathcal{B}_{\text{rev}}$ contains J selected spans with start s_j and exclusive end e_j. Each span is regenerated using the original evidence and structured confession:

$$\begin{aligned} u_j &\triangleq \text{Rev}_\theta\left(\mathbf{X}, y^{\text{cand}}, \mathbf{C}_\theta(Q,T); [s_j, e_j)\right), \\ y^{\text{final}} &\triangleq \text{Replace}\left(y^{\text{cand}}, \{[s_j, e_j) \mapsto u_j\}_{j=1}^{J}\right). \end{aligned} \tag{15}$$

where Rev_θ generates replacement u_j for span j, Replace preserves content outside selected spans.

Proposition 3. *Motivated by Figure 3(c), confession-guided correction yields task-relevant judgment margins and preserves content outside selected revision spans.*

Proof. We provide experimental results in Sections 5.3 and proofs in Appendix B.3.

Table 1: Main results on OMNIHALLUBENCH across three backbones. Best and second-best are in bold and underlined, respectively. Avg. excludes missing entries; — denotes unavailable results.

	Judgment tasks						Free-form tasks						Avg.		
	PHD		CMM		PubMed		RAGTruth		HaloQ.		AVHBch				
Method	**F1**	**Idx**	**F1**	**Idx**	**F1**	**Idx**	**F1**	**GAV**	**F1**	**GAV**	**F1**	**GAV**	**F1**	**Idx**	**GAV**
Qwen2.5-Omni-7B															
Base	71.68	68.35	70.07	60.77	95.14	80.21	38.25	8.77	34.22	7.16	19.42	5.96	54.80	69.78	7.30
SC	76.43	73.18	71.96	71.31	91.25	68.29	35.38	8.89	34.48	7.14	18.29	6.04	54.63	70.93	7.36
GD	21.16	21.73	57.54	63.12	57.37	39.35	11.61	8.59	40.35	6.57	9.80	4.36	32.97	41.40	6.51
PD	32.12	33.01	58.09	63.61	79.35	35.20	12.72	8.57	38.09	5.43	10.85	4.76	38.54	43.94	6.25
ToT	70.25	69.07	67.85	71.91	68.07	43.67	12.08	8.56	42.09	7.38	19.46	6.20	46.64	61.55	7.38
VCD	72.05	68.64	69.90	60.32	95.14	80.21	38.40	8.69	34.62	7.09	19.26	5.99	54.90	69.72	7.26
AVCD	73.81	66.62	69.08	61.93	94.79	86.94	33.78	8.40	33.25	5.28	19.76	5.89	54.08	71.83	6.52
CAD	79.40	73.82	67.37	53.46	95.24	83.46	36.29	8.79	36.06	6.54	18.01	5.94	55.40	70.25	7.09
MAD	—	—	72.80	65.86	—	—	—	—	—	—	16.84	5.88	44.82	65.86	5.88
DoLa	72.30	69.57	70.16	60.99	93.52	72.09	38.74	8.75	33.62	7.05	19.15	6.09	54.58	67.55	7.30
OPERA	74.33	65.71	58.13	62.89	94.91	87.93	36.36	8.53	35.91	5.37	21.11	6.15	53.46	72.18	6.68
OmniConfess	**89.60**	**78.46**	**82.15**	**84.01**	**97.26**	**92.38**	**55.58**	**8.95**	**43.14**	**8.65**	**22.87**	**6.38**	**65.10**	**84.94**	**7.99**
Qwen3-Omni-30B-A3B															
Base	83.40	77.24	74.56	73.69	93.93	71.02	36.30	8.85	33.38	7.25	19.03	6.28	56.77	73.98	7.46
SC	85.19	78.99	**80.72**	**83.39**	94.76	65.82	37.33	8.98	27.37	7.34	16.78	6.24	57.03	76.07	7.52
GD	38.66	39.46	58.97	64.89	77.22	28.46	22.42	8.08	35.19	6.37	7.46	4.32	39.99	44.27	6.26
PD	40.59	41.30	58.50	63.21	67.85	34.70	19.90	8.15	31.71	6.87	7.03	3.66	37.60	46.40	6.23
ToT	83.51	78.26	77.54	79.93	83.38	31.50	22.14	8.07	**41.50**	7.57	18.59	6.43	54.44	63.23	7.36
VCD	84.47	77.65	73.68	72.39	93.93	71.02	35.73	8.96	33.65	6.77	19.07	6.23	56.76	73.69	7.32
AVCD	84.54	77.84	74.32	73.12	93.40	70.74	36.19	9.01	33.21	6.84	19.14	6.19	56.80	73.90	7.35
CAD	85.71	78.00	78.65	80.16	94.88	80.03	38.46	8.96	26.76	7.41	17.74	6.16	57.03	79.40	7.51
MAD	—	—	78.21	80.84	—	—	—	—	—	—	21.49	6.30	49.85	80.84	6.30
DoLa	83.91	78.01	73.71	71.24	93.12	71.87	38.14	8.96	31.43	6.74	18.81	6.21	56.52	73.71	7.30
OPERA	80.11	76.11	74.30	76.24	94.15	74.94	41.30	8.94	37.35	6.95	**21.74**	6.50	58.16	75.76	7.46
OmniConfess	**87.47**	**79.10**	80.24	81.11	**96.76**	**83.84**	**53.20**	**9.06**	39.32	**8.14**	21.66	**6.58**	**63.11**	**81.35**	**7.93**
Nemotron-3-Nano-Omni-30B															
Base	83.14	74.98	76.36	79.46	78.62	66.94	45.10	—	16.48	—	18.67	—	53.06	73.79	—
SC	84.68	76.06	78.82	80.57	72.62	58.72	42.10	9.02	16.37	6.68	17.39	5.94	52.00	71.78	7.21
GD	**86.36**	**81.23**	76.06	78.55	23.72	19.61	26.89	9.11	35.64	8.28	16.63	5.58	44.22	53.26	7.66
PD	82.54	79.05	77.61	80.00	24.08	23.72	26.75	**9.12**	30.43	7.85	17.29	5.97	43.13	54.38	7.65
ToT	85.50	80.53	**80.56**	**82.54**	37.93	21.36	26.48	9.09	30.40	7.53	16.47	5.60	46.22	61.48	7.41
VCD	83.80	74.99	76.83	79.65	72.20	55.08	43.21	9.06	24.32	6.09	21.19	6.25	53.59	69.91	7.13
AVCD	59.03	53.49	56.94	40.02	69.84	63.76	24.92	9.01	25.21	6.20	**22.22**	6.21	43.03	52.42	7.14
CAD	83.04	74.91	78.71	80.43	82.66	69.71	44.08	8.96	32.04	7.03	19.63	6.24	56.69	75.02	7.41
MAD	—	—	73.61	74.70	—	—	—	—	—	—	22.10	**6.30**	47.86	74.70	6.30
DoLa	83.42	74.72	73.72	77.26	62.66	55.45	44.37	8.99	22.14	5.92	21.20	6.18	51.25	69.14	7.03
OPERA	58.84	58.24	76.60	79.38	77.23	63.40	44.90	8.96	30.23	6.58	20.04	6.04	51.31	67.01	7.19
OmniConfess	84.66	75.58	80.50	81.95	**90.63**	**71.65**	**51.71**	9.12	**40.75**	**8.57**	21.14	6.25	**61.57**	**76.39**	**7.88**

5 EXPERIMENTS

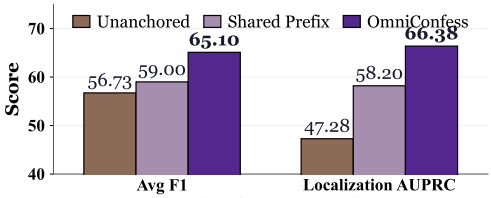
We answer six research questions: **RQ1:** Does OmniConfess outperform existing methods? **RQ2:** Does each stage of OmniConfess work? **RQ3–5:** How are semantic quality, cross-backbone generalization, and inference efficiency?

5.1 EXPERIMENTAL SETUP

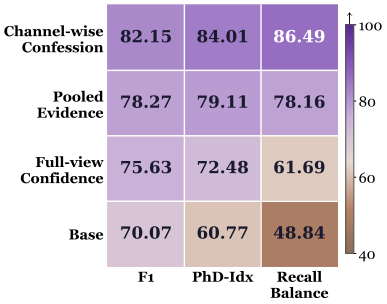
Datasets. We construct OMNIHALLUBENCH, comprising 3,540 examples from six existing benchmarks. Judgment tasks include **CMM** (Leng et al., 2024) (Audio–Visual–Text), **PhD** (Liu et al., 2025) (Image–Text), and **PubMedQA** (Jin et al., 2019) (Text). Free-form tasks include **AVHBench** (Sung-Bin et al., 2025) (Audio–Visual–Text), **HaloQuest** (Wang et al., 2024) (Image–Text), and **RAGTruth** (Niu et al., 2024) (Text). Construction details are provided in Appendix D.

Configuration	Judgment Avg.		HaloQuest		Avg.		
	F1	Idx	F1	GAV	F1	Idx	GAV
w/o Anchoring	80.72	76.74	39.76	8.12	67.06	76.74	8.12
w/o Interrogation	77.25	73.42	37.38	7.83	63.96	73.42	7.83
w/o Correction	70.88	64.56	34.22	7.16	58.66	64.56	7.16
OmniConfess	**85.88**	**81.24**	**43.14**	**8.65**	**71.63**	**81.24**	**8.65**

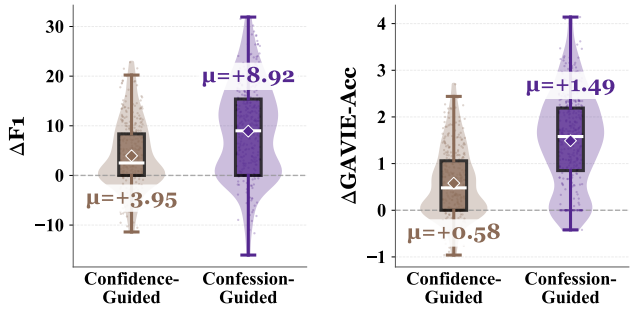
(a) Ablation Study



(b) Anchoring Analysis



(c) Evidence Interrogation



(d) Correction Strategies

Figure 5: (a) Module ablation of OmniConfess. (b) Comparison of commitment anchoring strategies. (c) Comparison of evidence interrogation strategies. (d) Comparison of correction strategies.

Backbones and Baselines. We evaluate OmniConfess on three OmniLLMs: `Qwen2.5-Omni-7B` (Xu et al., 2025b) (default), `Qwen3-Omni-30B-A3B` (Xu et al., 2025c), and `Nemotron-3-Nano-Omni-30B` (NVIDIA, 2026). Baselines include search-based methods (SC (Wang et al., 2023), GD (Xie et al., 2023), PD (Xu et al., 2025a), ToT (Yao et al., 2023)), evidence-contrast methods (VCD (Leng et al., 2023), AVCD (Jung et al., 2025), CAD (Shi et al., 2023), MAD (Chung et al., 2026)), and fine-grained methods (DoLa (Chuang et al., 2024), OPERA (Huang et al., 2024)), alongside the base model. Further details are in Appendix E.

Evaluation Metrics. Judgment tasks use F1 and PhD-Index (Idx) (Liu et al., 2025); free-form tasks use Token-F1 and GAVIE-inspired accuracy (GAV) (GAV) (Liu et al., 2024), scored by an LLM judge on a 0–10 scale using our task-specific evaluation prompt. Average F1 covers six datasets, while average Idx and GAV each cover three.For hallucination localization, we report span-level AUPRC. Metric definitions are in Appendix F.

Implementation Details. OmniConfess requires no additional training. All experiments are conducted on 4 NVIDIA A40 GPUs.

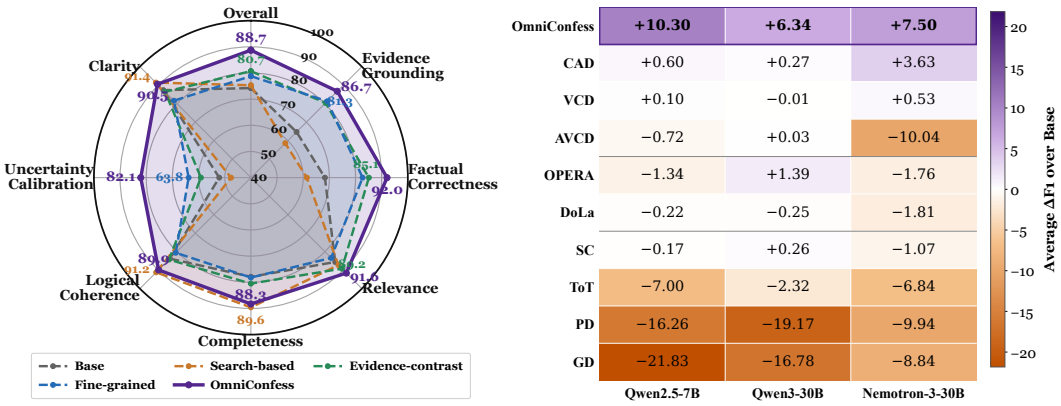
5.2 Main Results (RQ1)

As shown in Table 1, OmniConfess consistently outperforms all baselines across six datasets and all reported metrics on Qwen2.5-Omni-7B. We further identify three key observations.

Hallucination Control Across Evidence Modalities. OmniConfess consistently improves multimodal and text-only tasks. On PHD and CMM, F1 exceeds the strongest baselines by 10.20 and 9.35 points, suggesting channel-wise evidence interrogation curbs reliance on irrelevant channels. On RAGTruth, Token-F1 gains 16.84 points, suggesting confession mitigates underuse of relevant evidence. Despite PubMedQA's strong baseline, F1 rises by 2.02 points.

Hallucination Control Across Generation Tasks. For judgment tasks, improvements in F1 and Index indicate gains in decision accuracy and reduced hallucination-related bias. For free-form generation, gains in Token-F1 and GAVIE reflect gains in lexical consistency and factual accuracy. These results show token-level evidence correction benefits both judgment and generation.

Behavior of Existing Inference-Time Methods. Search-based methods show task sensitivity, with search or iteration sometimes degrading performance under multimodal evidence. Evidence-contrast methods remain close to the base model, as modality interventions may suppress evidence along with distractions. Fine-grained methods lack improvements, suggesting token-level risk signals may not determine whether evidence dependence should be strengthened or reduced.



(a) Generation Quality Comparison (b) Cross-Backbone Performance

Figure 6: Generation quality and cross-backbone portability of OmniConfess. (a) Comparison of OmniConfess and baseline methods across generation quality dimensions. (b) Average F1 changes relative to each backbone's base model across three architectures.

5.3 Ablation Study and Comparative Analysis (RQ2)

Figure 5 presents the ablation study and comparative analysis of OmniConfess.

Module Ablation. Figure 5(a) ablates anchoring, evidence interrogation, and correction. OmniConfess leads all metrics, averaging 71.63 F1. Removing anchoring, interrogation, or correction lowers average F1 to 67.06, 63.96, or 58.66, respectively. All modules contribute; correction's removal has the largest effect, underscoring the value of revision after identifying evidence-related errors.

Anchoring Analysis. Figure 5(b) compares the unanchored setting, shared-prefix anchoring (Shi et al., 2023), and OmniConfess. Shared-prefix anchoring increases average F1 from 56.73 to 59.00 and localization AUPRC from 47.28 to 58.20. OmniConfess improves them to 65.10 and 66.38, respectively. Thus, anchoring benefits both prediction and error localization, and the anchoring used by OmniConfess is more effective than a shared prefix alone.

Evidence Interrogation Strategies. Figure 5(c) compares full-view confidence (Ding et al., 2025), pooled evidence (Chung et al., 2026), and channel-wise confession. It leads in F1 (82.15) and PhD-Index (84.01). We define $\text{Recall Balance} = 100 - |R_{\text{YES}} - R_{\text{NO}}|$, with recall on a 0–100 scale; higher values indicate a smaller YES–NO gap. Channel-wise confession scores 86.49, versus 78.16 for pooled evidence, 61.69 for full-view confidence, and 48.84 for the base model. This reduction helps mitigate the model's bias toward predicting YES for NO-labeled samples.

Correction Strategies. Figure 5(d) compares revision strategies under an equal correction budget. Confession-guided revision yields larger mean gains than confidence-guided revision in F1 ($+8.92$ vs. $+3.95$) and GAVIE-Acc ($+1.49$ vs. $+0.58$). This supports using token- and channel-level evidence dependence to locate spans for correction.

5.4 Analysis of OmniConfess's Impact in Generation Quality (RQ3)

To assess generation quality beyond task metrics, we use GPT-4o (OpenAI et al., 2024) and human experts to cross-evaluate seven quality dimensions and an Overall score. Figure 6(a) compares inference-time baselines with OmniConfess. The prompt and protocol are detailed in Appendix A.3.

Baseline Analysis. Search-based methods retain clarity and coherence but lag in grounding and factual correctness, indicating that refinement need not improve evidential support. Evidence-contrast methods offer balanced quality, whereas Fine-grained interventions strengthen grounding but sacrifice linguistic quality. Both remain weak in uncertainty calibration.

OmniConfess Analysis. OmniConfess achieves the highest Overall score (88.7), with Evidence Grounding (86.7), Factual Correctness (92.0), and Uncertainty Calibration (82.1) exceeding the strongest baselines by 5.4, 6.9, and 18.3 points, respectively. Its competitive clarity and coherence, alongside markedly improved uncertainty calibration, suggest that token- and channel-level evidence interrogation supports reliable correction without sacrificing response quality.

Question: "Beach towels are always spread out on the sand ... kiteboarding, parasailing, and windsurfing ... In case there is an inconsistency between the context and the image content, you should follow the image. Is there a surfboard in the image?"
Ground Truth: **Yes**



	Dense (Qwen2.5-Omni-7B)	MoE (Qwen3-Omni-30B)	Hybrid (Nemotron-3-30B)
Omni Confess	Image Observation: - A person is windsurfing on the water. - Windsurfing uses a board with a sail. **Judgment: \boxed{Yes}**	Image Observation: - There is a person windsurfing in the water. - Windsurfing equipment includes a board and a sail. **Judgment: \boxed{Yes}**	Image Observation: - There are two people engaged in water sports on the water. - One person is kiteboarding, using a board that is not clearly visible due to the water spray. - The other person is windsurfing, which involves a board with a sail attached. - The beach in the background has people, but no clearly identifiable surfboards are spread out on the sand. **Judgment: \boxed{Yes}**
OPERA	Image Observation: - There is a person windsurfing in the water. - Windsurfing equipment includes a board and a sail. **Judgment: \boxed{Yes}**	Image Observation: - There are people doing water sports on the sea. - One person is windsurfing with a sail, and a kite is flying above. - No plain surfboard is clearly visible on the water or on the sand. **Judgment: \boxed{No}**	Image Observation: - There are people doing water sports on the sea. - One person is windsurfing with a sail, and a kite is flying above. - No plain surfboard is clearly visible on the water or on the sand. **Judgment: \boxed{No}**

Figure 7: Case study of OmniConfess and OPERA across three backbones on PHD.

Table 2: Time, output token consumption, and normalized throughput (tokens/sec).

Category	PHD Time · Tokens · Tok/s	PubMedQA Time · Tokens · Tok/s	CMM Time · Tokens · Tok/s	RAGTruth Time · Tokens · Tok/s	HaloQuest Time · Tokens · Tok/s	AVHBench Time · Tokens · Tok/s
BaseModel	2.23 · 23.4 · 10.5	0.81 · 25 · 30.9	20.64 · 22.7 · 1.1	2.69 · 98 · 36.4	0.93 · 10 · 10.8	2.65 · 25 · 9.4
Search-based	34.76 · 789.1 · 22.7	20.28 · 474.8 · 23.4	38.90 · 536.8 · 13.8	30.10 · 838.3 · 27.9	28.39 · 536.3 · 18.9	36.15 · 569.5 · 15.8
Evidence-contrast	7.17 · 21.5 · 3.0	1.19 · 26.3 · 22.1	25.44 · 22.9 · 0.9	2.49 · 69.4 · 27.9	1.32 · 10.5 · 8.0	4.09 · 22.1 · 5.4
Fine-grained	4.17 · 26.7 · 6.4	3.71 · 25.0 · 6.7	15.18 · 25.8 · 1.7	6.48 · 65.7 · 10.1	1.83 · 10.1 · 5.5	5.01 · 24.9 · 5.0
OmniConfess	5.43 · 29.3 · 5.4	1.23 · 25 · 20.3	35.14 · 24.6 · 0.7	4.82 · 89.2 · 18.5	0.82 · 12.5 · 15.2	7.93 · 164.1 · 20.7

5.5 Analysis of OmniConfess's Portability Across Backbone Models (RQ4)

To assess architectural generalizability, we evaluate methods on Qwen2.5-Omni-7B (Dense), Qwen3-Omni-30B-A3B (MoE), and Nemotron-3-Nano-Omni-30B (Hybrid). Figure 6(b) shows that OmniConfess improves average F1 over their respective base models by 10.30, 6.34, and 7.50 points.

Unlike OmniConfess, baseline methods either vary across backbones or perform consistently poorly. AVCD, OPERA, and SC switch between gains and losses, improving only on MoE but declining on Dense and Hybrid, while GD, PD, and ToT reduce F1 on all three backbones. Even CAD, which improves all three, yields smaller gains than OmniConfess. These results highlight OmniConfess's consistent advantage across Dense, MoE, and Hybrid architectures.

Case Study. Figure 7 shows larger backbones do not guarantee better grounding. OPERA succeeds on Dense, but its longer reasoning on 30B backbones narrows the target category and overrides visual evidence. Architecture-dependent reasoning may amplify multimodal hallucinations. OmniConfess remains correct across architectures, showing its robustness.

5.6 Analysis of OmniConfess's Inference Efficiency (RQ5)

We evaluate inference time, generated tokens, and normalized throughput in Table 2.

Evidence Interrogation Overhead. Judgment outputs are short, yet OmniConfess takes 5.43 seconds on PHD and 35.14 seconds on audio-visual CMM, compared with 2.23 and 20.64 seconds for BaseModel. CMM generates only 24.6 tokens, suggesting that repeated channel-wise evidence rescoring, rather than output length alone, accounts for much of its latency.

Token–Latency Trade-off. Compared with search-based methods, OmniConfess uses about 27× and 22× fewer generated tokens on PHD and CMM, respectively, but reduces inference time by only 6.4× and 1.1×. On AVHBench, generated tokens fall from 569.5 to 164.1, while time falls from 36.15 to 7.93 seconds. Anchoring a candidate limits generation cost, but multimodal evidence interrogation still incurs substantial latency on some tasks.

6 Discussion and Conclusion

In this work, we propose OmniConfess, a training-free framework for mitigating omni-modal hallucinations through token- and channel-level evidence interrogation. Structured confessions expose evidence dependencies and guide correction in both judgment and free-form generation. Experiments across six benchmarks show improved factual reliability and evidence grounding, with F1 gains of up to 13.3 points and consistent gains across backbone architectures.

AI USE STATEMENT

We used generative AI tools to assist with English translation and polishing the manuscript, including the presentation of experimental results. The authors reviewed the AI-assisted text, verified its claims against the experimental records, and take responsibility for the final content.

REFERENCES

Ashutosh Chaubey, Jiacheng Pang, and Mohammad Soleymani. Mod-dpo: Towards mitigating cross-modal hallucinations in omni llms using modality decoupled preference optimization. 2026. URL https://arxiv.org/abs/2603.03192.

Ruoyu Chen, Xiaoqing Guo, Kangwei Liu, Siyuan Liang, Shiming Liu, Qunli Zhang, Laiyuan Wang, Hua Zhang, and Xiaochun Cao. Where mllms attend and what they rely on: Explaining autoregressive token generation. 2026. URL https://arxiv.org/abs/2509.22496.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. 2024. URL https://arxiv.org/abs/2309.03883.

Sangyun Chung, Se Yeon Kim, Youngchae Chee, and Yong Man Ro. Mad: Modality-adaptive decoding for mitigating cross-modal hallucinations in multimodal large language models. 2026. URL https://arxiv.org/abs/2601.21181.

Zhikai Ding, Shiyu Ni, and Keping Bi. Do lvlms know what they know? a systematic study of knowledge boundary perception in lvlms. 2025. URL https://arxiv.org/abs/2508.19111.

Zixuan Dong, Jiafu Tang, Zhide Lei, Zhe Cao, Zijie Zhang, Yanghai Wang, Shihao Li, Xiaodong Wang, Baoyun Peng, and Jiaheng Liu. Omnihalluc-l: Counterfactual benchmarking and modality-perturbation reliability calibration for long-form omni hallucination, 2026. URL https://arxiv.org/abs/2606.03614.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. 2024. URL https://arxiv.org/abs/2403.14003.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. 2024. URL https://arxiv.org/abs/2310.14566.

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. 2024. URL https://arxiv.org/abs/2311.17911.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. 2019. URL https://arxiv.org/abs/1909.06146.

Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. Avcd: Mitigating hallucinations in audio-visual large language models through contrastive decoding. 2025. URL https://arxiv.org/abs/2505.20862.

Seunghee Kim, Ingyu Bang, Seokgyu Jang, Changhyeon Kim, Sanghwan Bae, Jihun Choi, Richeng Xuan, and Taeuk Kim. Omhbench: Benchmarking balanced and grounded omni-modal multi-hop reasoning. 2026. URL https://arxiv.org/abs/2508.16198.

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. 2023. URL https://arxiv.org/abs/2311.16922.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. 2024. URL https://arxiv.org/abs/2410.12787.

Li et al. Mitigating multimodal hallucinations through visual attention tracing and origin-point regeneration. *Scientific Reports*, 2026.

Yi Li, Hualiang Wang, Xinpeng Ding, Haonan Wang, and Xiaomeng Li. Token activation map to visually explain multimodal llms. 2025. URL https://arxiv.org/abs/2506.23270.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. 2023. URL https://arxiv.org/abs/2305.10355.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. 2024. URL https://arxiv.org/abs/2306.14565.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. Phd: A chatgpt-prompted visual hallucination evaluation dataset. 2025. URL https://arxiv.org/abs/2403.11116.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. 2024. URL https://arxiv.org/abs/2401.00396.

NVIDIA. Nemotron-3-Nano-Omni-30B-A3B-Reasoning model card. Hugging Face, nvidia/Nemotron-3-Nano-Omni-30B-A3B-Reasoning-BF16, 2026.

OpenAI, :, Aaron Hurst, and Adam Lerer. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Ouyang et al. Taming the phantom: Token-asymmetric filtering for hallucination mitigation in large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

Letitia Parcalabescu and Anette Frank. Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models amp; tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4032–4059. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.223. URL http://dx.doi.org/10.18653/v1/2023.acl-long.223.

Xiaoye Qu, Qiyuan Chen, Wei Wei, Jishuo Sun, and Jianfeng Dong. Alleviating hallucination in large vision-language models with active retrieval augmentation, 2024. URL https://arxiv.org/abs/2408.00555.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. 2023. URL https://arxiv.org/abs/2305.14739.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. 2023. URL https://arxiv.org/abs/2309.14525.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. 2025. URL https://arxiv.org/abs/2410.18325.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. 2023. URL https://arxiv.org/abs/2203.11171.

Zhecan Wang, Garrett Bingham, Adams Yu, Quoc Le, Thang Luong, and Golnaz Ghiasi. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. 2024. URL https://arxiv.org/abs/2407.15680.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models. 2024. URL https://arxiv.org/abs/2407.05131.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. Self-evaluation guided beam search for reasoning. 2023. URL https://arxiv.org/abs/2305.00633.

Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Jun Liu, Qika Lin, and Zhiyong Wu. ϕ-decoding: Adaptive foresight sampling for balanced inference-time exploration and exploitation, 2025a. URL https://arxiv.org/abs/2503.13288.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025b. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025c. URL https://arxiv.org/abs/2509.17765.

Qianqi Yan, Yichen Guo, Ching-Chen Kuo, Shan Jiang, Hang Yin, Yang Zhao, and Xin Eric Wang. Omnitrace: A unified framework for generation-time attribution in omni-modal llms, 2026. URL https://arxiv.org/abs/2604.13073.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. 2023. URL https://arxiv.org/abs/2305.10601.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. 2024. URL https://arxiv.org/abs/2312.00849.

Yifan Zhu, Huiqiang Rong, and Haoran Luo. Token-guard: Towards token-level hallucination control via self-checking decoding. 2026. URL https://arxiv.org/abs/2601.21969.

A INPUTS USED IN OMNICONFESS

A.1 OMNICONFESS PROMPTS

As shown in Figure 8, our generation prompts establish task-specific evidence boundaries rather than relying solely on generic anti-hallucination instructions. These constraints distinguish audible from visible evidence, prevent unsupported cross-modal associations, and require explicit acknowledgment when visual details are absent. Moreover, task-dependent output formats standardize the generated responses, allowing judgment and free-form tasks to preserve their respective answer requirements while remaining grounded in the provided evidence.

– CMM: Cross-Modal Matching –
You are an expert audio-visual reporter. Your task is to provide a single-sentence narrative combining what you see and hear, followed by a final judgment.

STRICT OUTPUT FORMAT:
Description: [One flowing sentence describing the scene and the EXACT sounds heard, like: 'In a room, a man spoke by a window where birds were seen but not heard.']
Judgment: \boxed{Yes} or \boxed{No}

STRICT AUDIT RULES:
1. If a visual object (like a bird) is present but its specific sound (chirping) is NOT in the audio, you MUST state that it was not heard in the Description.
2. The Description must be a single, descriptive sentence.
3. Judgment must be based ONLY on what is audible.

– PubMedQA: Biomedical Question Answering –
You will be given a PubMed-style passage and a Yes/No/Maybe question.
Answer rules:
1. Begin with exactly one of: Yes. / No. / Maybe.
2. Then add ONE short sentence (≤25 words) summarizing the key conclusion from the passage.
3. Preserve key medical terms from the passage; do not replace them with synonyms.
4. Do NOT add recommendations, explanations, or new information.
Example format: Yes. The study found that treatment X significantly reduced symptom Y in patients with Z.

– HaloQuest: Visual Question Answering –
You are an expert visual assistant. Your task is to answer the question based ONLY on the provided image.
CRITICAL: If the question asks about a detail or object that is NOT present in the image, you MUST explicitly state that it is not present or visible.
Answer in a complete, natural sentence (e.g. 'The girl is wearing red sneakers.' or 'There is no cat visible in the image.'). Be strictly factual and do not hallucinate.

– AVHBench: Audio-Visual Captioning –
You are an objective audiovisual forensic analyst. Your task is to decode the video and audio streams into a high-density, expert-level description in ONE SHORT sentence.
CRITICAL ANTI-HALLUCINATION DIRECTIVES:
1. MODALITY INDEPENDENCE: Analyze what you SEE and HEAR independently. DO NOT let the visual scene change your interpretation of the audio, and vice versa.
2. NO SEMANTIC SMOOTHING: If visuals and audio seem contradictory (e.g., a forest visual but an engine sound), report BOTH exactly as they are. Do not invent details like 'chirping birds' to make it natural.
3. AVOID ASSUMPTIONS: Rely on audio to confirm ambiguous visuals (e.g., a 'baaing' sound confirms a goat, do not assume it is a cat).

FORMATTING INSTRUCTIONS:
1. Compression: Pack all objective sensory details into a maximum of 50 words.
2. Directness: Skip all introductory phrases. Start immediately with the main subject.
3. Keywords: Prioritize specific nouns and adjectives.

CORE RULE: Be extremely dense, strictly factual, and under 50 words. Output a plain sentence with NO formatting wrappers.

Figure 8: Representative generation prompts used in OmniConfess.

A.2 GAVIE-INSPIRED EVALUATION PROMPT

For free-form generation, we adopt a GAVIE-inspired evaluation protocol to assess response factual accuracy on a 0–10 scale. As shown in Figure 9, our reference-grounded evaluation prompt uses the reference answer as the sole ground truth and penalizes unsupported or contradictory details, rather than reproducing the original GAVIE procedure.

Reference: {gt}
Model Prediction: {pred}

You are a strict judge. The Reference is the ONLY ground truth; the Model Prediction is the answer under review. Check it only against the Reference, never against your own knowledge, and never reward fluency or plausibility alone.

Score the prediction (0.0–10.0) on accuracy: how many details in the Model Prediction are consistent with the Reference. Any detail that contradicts or is absent in the Reference is a hallucination and lowers accuracy even if plausible.

Use decimal points for partial correctness (e.g., 7.4).

Output strictly in this format: accuracy: Y.Y

Figure 9: GAVIE-inspired evaluation prompt for factual accuracy.

A.3 MULTIDIMENSIONAL QUALITY EVALUATION PROMPTS

For multidimensional quality assessment, we design seven evaluation prompts to assess free-form responses on a 0–10 scale. As shown in Figure 10, these prompts cover evidence grounding, factual

correctness, relevance, completeness, logical coherence, uncertainty calibration, and clarity. Each prompt specifies dimension-specific criteria and inputs to evaluate aspects of response quality.

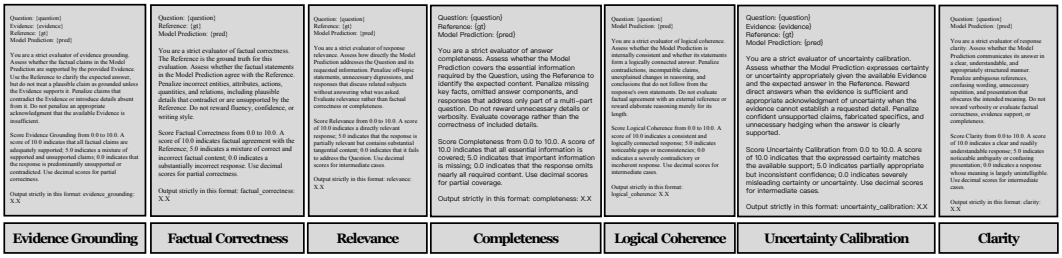


Figure 10: Seven dimension-specific prompts for multidimensional response quality evaluation.

B Theoretical Proof

B.1 Proof of Proposition 1

Proposition 1. Motivated by the front-loaded dependence in Figure 3(a), commitment anchoring eliminates trajectory variation from token-level comparisons across evidence conditions.

Proof. Consider the candidate y^{cand} generated under full evidence $\mathbf{X}$. Without commitment anchoring, removing channel m produces a new autoregressive trajectory $\widetilde{y}^{(m)} = \text{Decode}_\theta(\mathbf{X}^{-m})$. If the two trajectories first differ at position k, their autoregressive prefixes necessarily differ at every subsequent shared position:

$$y_k^{\text{cand}} \neq \widetilde{y}_k^{(m)} \implies y_{<i}^{\text{cand}} \neq \widetilde{y}_{<i}^{(m)}, \quad \forall i > k. \tag{16}$$

An early token change therefore alters the conditioning context of subsequent decisions. This propagation is particularly relevant to the front-loaded dependence observed in Figure 3(a).

Without anchoring, the log-probability contrast between the selected tokens on the two trajectories is

$$\Delta_i^{(m)} \triangleq \log \frac{p_\theta(y_i^{\text{cand}} \mid y_{<i}^{\text{cand}}, \mathbf{X})}{p_\theta(\widetilde{y}_i^{(m)} \mid \widetilde{y}_{<i}^{(m)}, \mathbf{X}^{-m})}. \tag{17}$$

This contrast changes the evidence condition and potentially the selected token and its prefix. To separate these changes, define the trajectory term under the channel-removed condition as

$$\tau_i^{(m)} \triangleq \log \frac{p_\theta(y_i^{\text{cand}} \mid y_{<i}^{\text{cand}}, \mathbf{X}^{-m})}{p_\theta(\widetilde{y}_i^{(m)} \mid \widetilde{y}_{<i}^{(m)}, \mathbf{X}^{-m})}. \tag{18}$$

Adding and subtracting the frozen candidate's log-probability under the channel-removed condition gives the exact decomposition

$$\Delta_i^{(m)} = \delta_i^{(m)}(y_i^{\text{cand}}) + \tau_i^{(m)}. \tag{19}$$

The first term measures the change in the frozen commitment's score under evidence intervention, whereas the second accounts for the difference between the two generated trajectories.

Commitment anchoring fixes the candidate tokens and autoregressive prefixes across evidence conditions. The anchored log-probability contrast is consequently

$$\Delta_{i,\text{anchor}}^{(m)} \triangleq \log \frac{p_\theta(y_i^{\text{cand}} \mid y_{<i}^{\text{cand}}, \mathbf{X})}{p_\theta(y_i^{\text{cand}} \mid y_{<i}^{\text{cand}}, \mathbf{X}^{-m})} = \delta_i^{(m)}(y_i^{\text{cand}}). \tag{20}$$

The trajectory term is thus excluded from the comparison. Moreover, applying autoregressive factorization to the same fixed sequence yields

$$\sum_{i=1}^{n} \delta_i^{(m)}(y_i^{\text{cand}}) = \log \frac{p_\theta(y^{\text{cand}} \mid \mathbf{X})}{p_\theta(y^{\text{cand}} \mid \mathbf{X}^{-m})}. \tag{21}$$

The probabilities on the right denote the autoregressive likelihoods of the complete fixed candidate under the respective evidence conditions. Hence, the token-level differences exactly decompose the evidence-induced log-likelihood change of the same candidate response. Commitment anchoring eliminates generated-trajectory variation from these comparisons, establishing Proposition 1.

B.2 Proof of Proposition 2

Proposition 2. Motivated by Figure 3(b), the channel-wise surrogate preserves observed margins and yields task-specific evidence signals.

Proof. Let $\mathcal{M} = \{1, \dots, M\}$ denote the complete evidence-channel set. For compactness, write $F_i(\mathcal{M}) \equiv F_i(\mathbf{X})$ and $F_i(\mathcal{M}\setminus\{m\}) \equiv F_i(\mathbf{X}^{-m})$ for the observed frozen decision margins. All conditions share the same candidate, alternative, and autoregressive prefix. Their difference therefore determines the conditional channel contribution $\kappa_i^{(m)}$ defined in Section 4.3.

Consider an additive surrogate over the retained channel set:

$$\widehat{F}_i(\mathcal{S}) = b_i + \sum_{m\in\mathcal{S}} w_i^{(m)}, \qquad \mathcal{S} \subseteq \mathcal{M}. \tag{22}$$

where b_i is the intercept and $w_i^{(m)}$ is the additive contribution assigned to channel m. We require consistency with the full-view margin and every observed single-channel intervention:

$$\widehat{F}_i(\mathcal{M}) = F_i(\mathcal{M}), \qquad \widehat{F}_i(\mathcal{M}\setminus\{m\}) = F_i(\mathcal{M}\setminus\{m\}). \tag{23}$$

Subtracting each channel-removed constraint from the full-view constraint eliminates the intercept and all contributions except that of channel m. Substituting the resulting contributions into the full-view constraint determines the intercept:

$$w_i^{(m)} = \kappa_i^{(m)}, \qquad b_i = F_i(\mathcal{M}) - \sum_{m\in\mathcal{M}} \kappa_i^{(m)}. \tag{24}$$

Thus, the $M+1$ observed margins uniquely determine all $M+1$ additive parameters. Substitution gives the reconstruction

$$\widehat{F}_i(\mathcal{S}) = F_i(\mathcal{M}) - \sum_{m\notin\mathcal{S}} \kappa_i^{(m)} = \widehat{F}_i(\varnothing) + \sum_{m\in\mathcal{S}} \kappa_i^{(m)}. \tag{25}$$

Setting $\mathcal{S} = \mathcal{M}$ recovers the full-view margin, while setting $\mathcal{S} = \mathcal{M}\setminus\{m\}$ recovers each observed channel-removed margin. The surrogate therefore preserves every observed frozen decision margin exactly.

Applying the no-channel and task-relevant references from Section 4.3, the same reconstruction yields

$$F_i(\mathcal{M}) = P_i^{\varnothing} + E_i^{\mathcal{R}} + E_i^{\mathcal{I}} = \widehat{F}_i(\varnothing) + \sum_{m\in\mathcal{M}} \kappa_i^{(m)}. \tag{26}$$

The difference between the two references gives $E_i^{\mathcal{R}}$, while the difference between the full-view margin and the task-relevant reference gives $E_i^{\mathcal{I}}$. These signed contributions retain channel-specific effects despite the cross-channel heterogeneity observed in Figure 3(b) and determine the task-specific confession signals.

For judgment, substituting the confession correction into the full-view decision margin gives

$$F_d(\mathcal{M}) + \gamma(J^{\mathrm{cand}}) = F_d(\mathcal{M}) - P_d^{\varnothing} - E_d^{\mathcal{I}} = E_d^{\mathcal{R}}. \tag{27}$$

Thus, the corrected margin equals the net contribution of task-relevant evidence within the surrogate decomposition.

For free-form generation, the first confession component retains the signed relevant contribution, while the second records positive residual preference exceeding positive relevant support. The third component separately accumulates positive support from irrelevant channels:

$$\sum_{m\notin\mathcal{R}} [\kappa_i^{(m)}]_+ \geq \left[\sum_{m\in\mathcal{I}} \kappa_i^{(m)}\right]_+ = [E_i^{\mathcal{I}}]_+. \tag{28}$$

The inequality follows from the subadditivity of the positive-part operator. Consequently, negative contributions from other channels cannot cancel the positive irrelevant support recorded by this component.

Together, the observed-margin consistency, dual-reference decomposition, and properties of both task-specific signals establish Proposition 2.

B.3 Proof of Proposition 3

Judgment Margin Correction. By the additive surrogate in Eq. 10, retaining all channels recovers the observed full-view margin:

$$\begin{aligned} F_d(\mathbf{X}) &= \widehat{F}_d(\{1,\dots,M\}) \\ &= P_d^{\varnothing} + E_d^{\mathcal{R}} + E_d^{\mathcal{I}}. \end{aligned} \tag{29}$$

The judgment correction in Eq. 11 removes the residual preference and irrelevant evidence contribution:

$$\begin{aligned} \widetilde{F}_d &= F_d(\mathbf{X}) + \gamma(J^{\mathrm{cand}}) \\ &= P_d^{\varnothing} + E_d^{\mathcal{R}} + E_d^{\mathcal{I}} - P_d^{\varnothing} - E_d^{\mathcal{I}} \\ &= E_d^{\mathcal{R}}. \end{aligned} \tag{30}$$

Since the alternative-class score remains fixed, the corrected binary decision satisfies

$$J^{\mathrm{final}} = \begin{cases} J^{\mathrm{cand}}, & E_d^{\mathcal{R}} \geq 0, \\ a_d, & E_d^{\mathcal{R}} < 0. \end{cases} \tag{31}$$

Thus, the corrected margin depends only on the task-relevant contribution within the constructed surrogate. This identity does not assume that the resulting decision is necessarily correct.

Non-canceling Revision Risk. Expanding the free-form confession in Eq. 11 gives

$$\begin{aligned} \rho_i = {} & [-E_i^{\mathcal{R}}]_+ \\ & + \left[[P_i^{\varnothing}]_+ - [E_i^{\mathcal{R}}]_+ \right]_+ \\ & + \sum_{m \in \mathcal{I}} [\kappa_i^{(m)}]_+. \end{aligned} \tag{32}$$

Each component is nonnegative. Consequently,

$$\begin{aligned} &\rho_i \geq [-E_i^{\mathcal{R}}]_+, \\ &\rho_i \geq \sum_{m \in \mathcal{I}} [\kappa_i^{(m)}]_+ \geq \left[\sum_{m \in \mathcal{I}} \kappa_i^{(m)} \right]_+ = [E_i^{\mathcal{I}}]_+. \end{aligned} \tag{33}$$

The second inequality follows from the subadditivity of the positive-part operator. Hence, positive irrelevant-channel contributions cannot cancel one another in the revision score, even when their signed aggregate is small.

Local Content Preservation. Let the selected, nonoverlapping revision spans partition the candidate into alternating unselected segments t_j and selected segments b_j, with $\oplus$ denoting sequence concatenation:

$$y^{\mathrm{cand}} = t_0 \oplus b_1 \oplus t_1 \oplus \cdots \oplus b_J \oplus t_J, \quad b_j = y^{\mathrm{cand}}_{[s_j, e_j)}. \tag{34}$$

By the definition of $\mathrm{Replace}$ in Eq. 15, the finalized response is

$$y^{\mathrm{final}} = t_0 \oplus u_1 \oplus t_1 \oplus \cdots \oplus u_J \oplus t_J, \tag{35}$$

where u_j replaces only the corresponding selected segment b_j. Comparing Eqs. 34 and 35, every unselected segment t_j is preserved verbatim and in its original order. If no span is selected, $J = 0$ and $y^{\mathrm{final}} = y^{\mathrm{cand}}$.

Therefore, confession-guided correction yields the task-relevant judgment margin and confines free-form modifications to selected revision spans, establishing the proposition. □

C OMNICONFESS ALGORITHM DETAILS

To illustrate the complete inference procedure of OmniConfess, we present its workflow in pseudocode, comprising commitment anchoring, evidence interrogation, and confession-guided correction.

Algorithm 1 OmniConfess: Confession-Guided Inference

Require: Input X, task T, OmniLLM p_θ, alternatives $\{a_i\}$, relevance rules $r_m(Q, T)$, revision threshold τ
Ensure: Final response y^{final}

// Phase 1: Commitment Anchoring
Generate $y^{\text{cand}} \leftarrow \text{Decode}_\theta(X)$
Freeze y^{cand} and its autoregressive prefixes
// Phase 2: Evidence Interrogation
Obtain full-view token scores $L_i^0(v)$ and margins $F_i(X)$
for $m = 1$ to M **do**
    Construct channel-removed view X^{-m}
    Teacher-force y^{cand} on X^{-m} to obtain $L_i^m(v)$
    Compute token dependence $\delta_i^{(m)}(y_i^{\text{cand}})$
    Compute margin contribution $\kappa_i^{(m)}$
    Release temporary states for X^{-m}
end for
Determine relevant channels $\mathcal{R}$ and irrelevant channels $\mathcal{I}$
Assign diagnostic states $\ell_i^{(m)}$ and assemble $\mathbf{C}_\theta(Q, T)$
Compute $E_i^{\mathcal{R}}$ and $E_i^{\mathcal{I}}$ for all candidate tokens
Compute $P_i^{\varnothing} \leftarrow F_i(X) - E_i^{\mathcal{R}} - E_i^{\mathcal{I}}$
// Phase 3: Confession-Guided Correction
if $T = \text{Judgment}$ **then**
    Identify candidate judgment J^{cand} and decision position d
    $\gamma(J^{\text{cand}}) \leftarrow -P_d^{\varnothing} - E_d^{\mathcal{I}}$
    $\widetilde{F}_d \leftarrow F_d(X) + \gamma(J^{\text{cand}})$
    if $\widetilde{F}_d < 0$ **then**
        return $\text{ReplaceDecision}(y^{\text{cand}}, d, a_d)$
    else
        return y^{cand}
    end if
else if $T = \text{Free-form}$ **then**
    Compute token-level confession γ_i^{free}
    Compute revision risk ρ_i for each token
    $\mathcal{S}_{\text{rev}} \leftarrow \text{Merge}(\text{Expand}(\{i : \rho_i > \tau\}))$
    for each span $[s_j, e_j) \in \mathcal{S}_{\text{rev}}$ **do**
        $u_j \leftarrow \text{Rev}_\theta(X, y^{\text{cand}}, \mathbf{C}_\theta(Q, T); [s_j, e_j))$
    end for
    return $\text{Replace}(y^{\text{cand}}, \{[s_j, e_j) \mapsto u_j\})$
end if

D OMNIHALLUBENCH DETAILS

D.1 BENCHMARK COMPOSITION

We construct OMNIHALLUBENCH from six public hallucination and question-answering benchmarks, covering text, image, and audio-video evidence across judgment and free-form generation tasks:

- **CMM** (Leng et al., 2024): An audio-visual hallucination benchmark evaluating multimodal understanding and evidence-grounded judgment.
- **PhD** (Liu et al., 2025): An image-grounded hallucination benchmark evaluating judgments under different visual context conditions.
- **PubMedQA** (Jin et al., 2019): A biomedical question-answering dataset requiring evidence-based judgments over PubMed abstracts.

- **AVHBench** (Sung-Bin et al., 2025): An audio-visual hallucination benchmark; we use its audio-visual captioning task.
- **HaloQuest** (Wang et al., 2024): A visual question-answering benchmark assessing hallucinations under challenging image-grounded questions.
- **RAGTruth** (Niu et al., 2024): A corpus of retrieval-grounded generation with fine-grained hallucination annotations.

The retained benchmark contains 906 PhD, 721 CMM, 468 PubMedQA, 435 RAGTruth, 365 HaloQuest, and 645 AVHBench examples, totaling 3,540.

D.2 Benchmark Construction

Source Selection and Schema Unification. We aggregate 5,900 candidates from six public benchmarks, pairing judgment and free-form tasks across text, image, and audio-video evidence. Questions, evidence, references, and task formats are organized under a unified schema.

Difficulty-Based Selection. Two independent answerers, GPT-4o and Gemini 2.5 Pro, generate responses, which GLM-4 and an independent researcher assess for hallucination severity. Candidates are globally ranked by difficulty, retaining the hardest 60% to form the 3,540-example benchmark. Neither OmniConfess nor any baseline participates in this method-agnostic selection.

Human Validation. Two independent researchers audit 200 candidates around the selection cutoff, comprising 100 retained examples (ranks 3441–3540) and 100 excluded examples (ranks 3541–3640). They independently rate anonymized responses using a three-level hallucination severity scale: grounded (0), minor hallucination (1), and material hallucination (2). This audit assesses the validity of difficulty-based selection.

Frozen Evaluation Protocol. The selected benchmark is fixed across all methods and backbones. Neither the automatic difficulty ranking nor the human validation uses outputs from OmniConfess or any baseline.

E Backbone and Baseline Details

Backbone Models. We evaluate OmniConfess on three OmniLLMs with different model architectures:

- **Qwen2.5-Omni-7B** (Xu et al., 2025b): A dense omni-modal model supporting text, image, audio, and video understanding and generation. We use it as the default backbone.
- **Qwen3-Omni-30B-A3B** (Xu et al., 2025c): A mixture-of-experts OmniLLM designed for unified multimodal perception and generation.
- **Nemotron-3-Nano-Omni-30B** (NVIDIA, 2026): A hybrid-architecture OmniLLM supporting multimodal reasoning and generation.

Search-based Methods.

- **Self-Consistency (SC)** (Wang et al., 2023): Samples multiple reasoning paths and aggregates their final predictions.
- **Guided Decoding (GD)** (Xie et al., 2023): Uses self-evaluation to guide the generation process.
- **Predictive Decoding (PD)** (Xu et al., 2025a): Anticipates future generation outcomes to guide current decoding decisions.
- **Tree-of-Thought (ToT)** (Yao et al., 2023): Explores and evaluates multiple reasoning branches before selecting a final answer.

Evidence-Contrast Methods.

- **Visual Contrastive Decoding (VCD)** (Leng et al., 2023): Contrasts predictions from original and perturbed visual inputs to mitigate visual hallucination.

- **Audio-Visual Contrastive Decoding (AVCD)** (Jung et al., 2025): Uses audio-visual evidence contrasts to improve multimodal grounding.
- **Context-Aware Decoding (CAD)** (Shi et al., 2023): Contrasts predictions with and without supporting context to strengthen evidence-grounded generation.
- **Modality-Adaptive Decoding (MAD)** (Chung et al., 2026): Adapts decoding to modality-specific evidence contributions.

Fine-Grained Methods.

- **DoLa** (Chuang et al., 2024): Contrasts predictions from different model layers to improve factuality by suppressing prematurely learned linguistic patterns.
- **OPERA** (Huang et al., 2024): Uses attention-based decoding control to mitigate visual hallucination by penalizing over-attended summary tokens.

F Evaluation Details

We evaluate OmniConfess using four metrics across judgment and free-form generation tasks.

(i) Yes-F1. For judgment tasks, we measure the harmonic mean of precision and recall for the positive class:

$$\mathrm{F1}_{\mathrm{Yes}} = \frac{2TP}{2TP + FP + FN}, \tag{36}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

(ii) PhD-Index. Following PhD (Liu et al., 2025), we compute the harmonic mean of Yes and No recall:

$$\mathrm{Idx} = \frac{2R_{\mathrm{Yes}}R_{\mathrm{No}}}{R_{\mathrm{Yes}} + R_{\mathrm{No}}}, \tag{37}$$

where R_{Yes} and R_{No} denote the recall of each judgment class.

(iii) Token-F1. For free-form tasks, we measure token-level overlap between each prediction and its reference:

$$\text{Token-F1} = \frac{1}{N}\sum_{i=1}^{N}\frac{2|P_i \cap G_i|}{|P_i| + |G_i|}, \tag{38}$$

where P_i and G_i are the token multisets of the prediction and reference, respectively, and N is the number of evaluated examples.

(iv) GAVIE-Acc. Inspired by GAVIE (Liu et al., 2024), we assess reference-grounded factual accuracy using an LLM judge on a 0–10 scale:

$$\text{GAVIE-Acc} = \frac{1}{N}\sum_{i=1}^{N} s_i, \tag{39}$$

where $s_i \in [0, 10]$ is the factual accuracy score assigned to the i-th response. Our evaluation prompt is provided in Appendix A.2.

(v) Localization AUPRC. For hallucination localization, we measure average precision (AP) over candidate spans ranked by predicted hallucination risk:

$$\mathrm{AUPRC} = \sum_{k=1}^{K}(R_k - R_{k-1})P_k, \tag{40}$$

where P_k and R_k denote precision and recall at the k-th distinct risk threshold, respectively. Hallucinated spans are treated as positives, and all methods are evaluated on the same candidate spans.

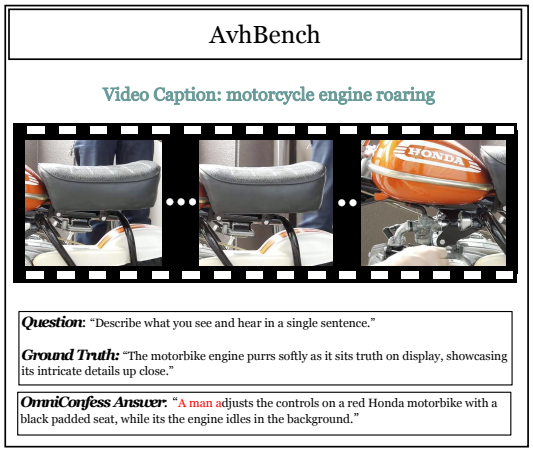


Figure 11: Case study on AVHBench.

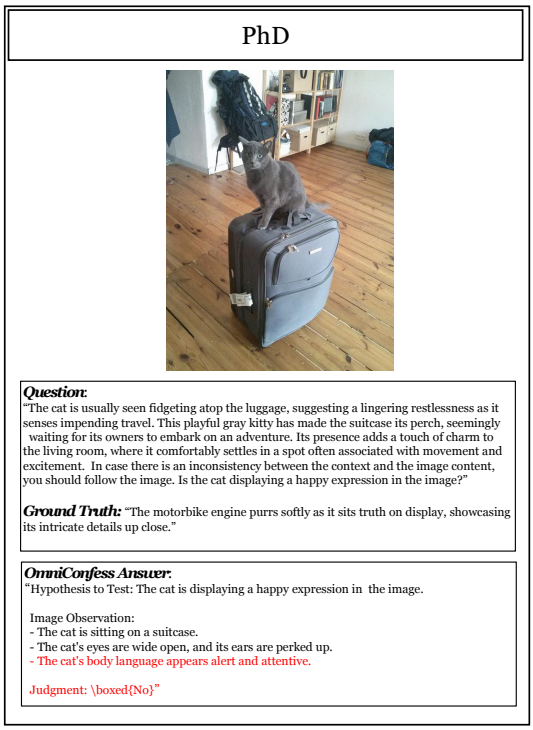


Figure 12: Case study on PHD.

G Error Case Studies

AVHBench Case Study. As shown in Figure 11, The model describes a "person" where the reference uses "man." Although the meaning is preserved, token-level F1 penalizes the lexical mismatch. This case illustrates a limitation of single-reference evaluation: confession can address evidence-related errors, but cannot eliminate penalties for valid paraphrases.

PHD Case Study. As shown in Figure 12,The model correctly observes a cat sitting on a suitcase with open eyes and upright ears, yet answers $\boxed{No}$ to whether it appears happy (ground truth: Yes). The error lies in the final judgment: it treats alertness as evidence against happiness without visual support. OmniConfess does not correct this subjective interpretation.

Together, these cases reveal two limitations: lexical mismatch under single-reference evaluation and incorrect judgments despite accurate visual observations.

H LIMITATIONS AND FUTURE WORK

OmniConfess identifies how predictions depend on evidence, but evidence dependence does not guarantee correct interpretation. A model may attend to the relevant channel yet draw an unsupported conclusion, especially when the judgment is ambiguous. Moreover, correction operates on a generated candidate, so errors embedded in its interpretation may persist even after local revision. Future work could jointly assess evidence dependence and interpretive validity, develop better-calibrated correction for ambiguous cases, and allocate interrogation selectively to reduce inference cost.