

OmniDenoise: Diffusion MultiModal LLMs for Omni-Modal Understanding and Prediction

Xianfeng Wu, Yuhang Chen, Jie Peng, Jinhao Duan, Tianlong Chen

University of North Carolina at Chapel Hill

{xianfeng, tianlong}@cs.unc.edu

Abstract

Omni-modal LLMs aim to answer queries over heterogeneous inputs through a unified interface, but accepting these inputs does not guarantee using them. Existing autoregressive and diffusion multimodal LLMs typically optimize answer likelihood only under the full context, leaving multimodal shortcuts largely unconstrained. We study this problem through the lens of *intervention sensitive omni-modal prediction*, where confidence in the correct answer should decrease when protocol designated evidence is disrupted. We propose *OmniDenoise*, a diffusion MLLM framework built around the *Evidence Denoising Gap* (EDG), a diffusion native objective that compares gold token denoising losses under full and intervened contexts. EDG uses the token-by-timestep denoising trajectory as both a training signal and a diagnostic, without requiring token level evidence annotations. OmniDenoise further applies consistency distillation for efficient few step bidirectional denoising. Across tri-modal, clinical, and out-of-distribution omni-modal benchmarks, OmniDenoise substantially improves over diffusion baselines and remains competitive with strong autoregressive omni models. On OmniBench, the distilled model achieves 45.27% accuracy at 0.19 seconds per sample. On EHRXQA, EDG improves H-Acc from 36.9% to 43.9%, while the distilled student retains 42.6% with a $3.9\times$ speedup.

1 Introduction

Omni-modal understanding and prediction require a model to answer questions from heterogeneous conditioning signals, such as images, audio, text, and structured records. This requirement is stronger than merely accepting many input streams through a unified interface. A model may process all modalities yet still answer from a language prior, a dominant modality, or a dataset specific shortcut, a failure mode long studied in visual question answering and multimodal robustness (Goyal

et al., 2017; Agrawal et al., 2018; Yang et al., 2024; Liu et al., 2025a). This issue is visible in omni-modal benchmarks that combine image, audio, and text (LI et al., 2025; Chen et al., 2025), and in clinical question answering where questions are paired with both chest X-rays and structured electronic health records (Bae et al., 2023). Recent autoregressive omni-modal models have made substantial progress toward universal multimodal interfaces (Hurst et al., 2024; Singh et al., 2025; Co-manici et al., 2025; AI et al., 2025; Li et al., 2025; Yu et al., 2025; Team et al., 2025; Team, 2026). However, high full context accuracy alone does not show whether a model’s predictions depend on the required modalities.

A useful omni-modal predictor should be sensitive to interventions on informative evidence. If a modality contains evidence needed for the answer, disrupting that modality should reduce the model’s confidence in the correct answer. Standard training objectives do not directly enforce this behavior. Autoregressive MLLMs optimize left-to-right likelihood under the observed context without explicitly comparing predictions under full and intervened contexts. Missing modality robustness addresses a related but distinct goal by encouraging graceful behavior when inputs are absent or corrupted. For evidence seeking omni-modal tasks, however, invariance to removing informative evidence may also indicate shortcut reliance. Intervention sensitivity and missing modality robustness are therefore complementary objectives: the former asks whether predictions benefit from informative evidence, while the latter asks how gracefully performance degrades when evidence is incomplete.

Diffusion MLLMs offer a natural way to model intervention sensitivity (Nie et al., 2025; Zhu et al., 2025; Ye et al., 2025c; You et al., 2025; Liu et al., 2025b; Ye et al., 2025b; You et al., 2026; Li et al., 2026; Kim et al., 2026). In masked diffusion, answer tokens are predicted through iterative denois-

ing. At each timestep, the model produces a distribution over every masked answer token conditioned on the multimodal context, the partially denoised answer state, and the timestep. This creates a token-by-timestep confidence trajectory. Existing diffusion omni-modal models mainly use this trajectory for reconstruction or unified generation. We instead use it to ask a diagnostic question: does the denoising confidence for gold answer tokens decrease when protocol designated evidence is intervened upon?

We propose *OmniDenoise*, a diffusion MLLM framework for intervention sensitive omni-modal prediction. The key component is the *Evidence Denoising Gap* (EDG), a intervention gap objective that measures the increase in gold token denoising loss between full and modality intervened contexts. EDG compares the two contexts on the same masked answer tokens under the same corrupted answer state, providing a matched token-by-timestep signal for whether answer denoising depends on the intervened modality. EDG does not require token level evidence annotations and does not by itself establish fine-grained causal evidence use; rather, it serves as a training signal and offline diagnostic for protocol level intervention sensitivity.

OmniDenoise is trained in two stages. First, we optimize masked answer denoising together with an aggregate EDG objective over designated modality interventions. This encourages accurate full context prediction while maintaining a loss gap when informative modalities are disrupted. Second, we adapt consistency distillation to few step omni-modal denoising, enabling faster inference while preserving bidirectional answer token refinement.

We evaluate OmniDenoise on OmniBench (LI et al., 2025), EHRXQA (Bae et al., 2023), and a filtered subset of UNO-Bench (Chen et al., 2025). OmniDenoise substantially improves over diffusion baselines and remains competitive with strong autoregressive omni-modal models in several settings. On OmniBench, the distilled model reaches 45.27% accuracy at 0.19 seconds per sample. EDG raises OmniBench accuracy from 37.48% to 46.41% and EHRXQA H-Acc from 36.9% to 43.9%, while consistency distillation preserves most of these gains and reduces latency by 5.3 $\times$ and 3.9 $\times$, respectively. The intervention effect also persists under real input mismatched replacement, indicating that it is not explained solely by artificial null modality cues. A matched autoregressive analogue of the intervention gap objective yields much

smaller marginal changes on OmniBench, suggesting that the largest benefit in our experiments is concentrated in the diffusion backbone.

Our contributions are as follows.

- We formulate *intervention sensitive omni-modal prediction*, where a model should not only answer correctly under full context but also reduce confidence when protocol designated evidence is intervened upon.
- We introduce the *Evidence Denoising Gap*, a diffusion native intervention gap objective and diagnostic that compares matched full context and intervened context denoising losses over masked gold answer tokens.
- We present *OmniDenoise*, an intervention sensitive omni-modal diffusion MLLM with consistency distillation for efficient few step inference.
- We validate OmniDenoise across tri-modal, clinical, and out-of-distribution benchmarks using controlled removal, real input mismatched replacement, and matched autoregressive comparisons, showing improved diffusion model accuracy, stronger selective intervention sensitivity, and better accuracy and latency trade-offs.

2 Related Work

Omni-modal autoregressive MLLMs. Autoregressive multimodal LLMs are increasingly developed as omni-modal assistants that process text, vision, audio, video, and speech through a unified interface (Hurst et al., 2024; Comanici et al., 2025; Li et al., 2025; Xu et al., 2025a,b; AI et al., 2025; Ye et al., 2025a). However, broad modality coverage does not guarantee that the model uses the modality containing the decisive evidence (Goyal et al., 2017; Agrawal et al., 2018; Yang et al., 2024; Ross et al., 2017). Recent omni-modal benchmarks explicitly evaluate cross-modal composition rather than unimodal priors (LI et al., 2025; Chen et al., 2025), and EHRXQA exposes a similar issue in clinical QA, where reliable answers may require jointly grounding structured EHR tables and chest radiographs (Bae et al., 2023). *OmniDenoise* targets this gap by training for intervention sensitivity, the model should remain confident under full evidence but reduce confidence when protocol designated evidence is removed.

Diffusion language and multimodal models.

Diffusion language models replace left-to-right prediction with iterative masked denoising, enabling bidirectional refinement and parallel updates over uncertain tokens (Nie et al., 2025; Zhu et al., 2025; Ye et al., 2025c). Recent multimodal extensions condition the denoising process on visual, video, action, or omni-modal inputs, making diffusion a viable decoder for multimodal understanding and generation (You et al., 2025; Ye et al., 2025b; You et al., 2026; Li et al., 2026; Kim et al., 2026). These models are typically trained to reconstruct targets under the observed context. In contrast, our work treats the token-by-timestep denoising trajectory as an evidence-sensitive signal, by comparing gold token denoising losses under full and modality ablated contexts, EDG directly supervises whether answer confidence depends on designated.

Efficient diffusion decoding and consistency distillation.

A practical limitation of diffusion models is iterative inference. Prior work accelerates sampling through fewer denoising steps, solver-based methods, caching, adaptive decoding, or distillation (Liu et al., 2022; Lu et al., 2022; Salimans and Ho, 2022; Song et al., 2023; Luo et al., 2023; Israel et al., 2025; Liu et al., 2025c; Wei et al., 2025; Wang et al., 2025; Minseo Kim, 2026; Liang et al., 2026). For diffusion language models, this cost is amplified because answer tokens are repeatedly processed in bidirectional refinement. In *OmniDenoise*, consistency distillation is a practical second stage rather than the main scientific claim, it compresses the EDG-trained teacher into a few step student while preserving intervention sensitive denoising behavior.

3 Method

Overview. *OmniDenoise* is a diffusion MLLM framework for intervention sensitive omni-modal prediction. Given a text query and heterogeneous conditioning inputs, the model predicts a textual answer through masked answer denoising rather than left-to-right generation. The core component is the *Evidence Denoising Gap* (EDG), a diffusion native intervention-gap objective that compares gold-token denoising losses under full and modality intervened contexts. EDG uses the matched token-by-timestep denoising trajectory as both a training signal and an offline diagnostic for whether model confidence depends on protocol-designated evidence modalities. After EDG training, we apply

a bidirectional consistency distillation stage to obtain a few step student for faster inference; the full distillation derivation is provided in Appendix B.

Problem Setup. Let (C, y) denote a training example. The conditioning context is

$$C = (q, x_m) \quad m \in \mathcal{M}, \quad (1)$$

where q is the text query, $\mathcal{M}$ is the set of non-query modalities, and x_m is the input from modality m . The modality set may include visual, acoustic, textual, or structured inputs. The answer is a token sequence $y = (y_1, \dots, y_n)$, and the answer span is treated as the denoising target.

$$\mathbb{E}_{(C,y),t,I_t} \left[\frac{1}{|I_t|} \sum_{i \in I_t} -\log p_\theta(y_i \mid C, \tilde{y}_t, t) \right]. \quad (2)$$

Unlike autoregressive decoding, masked denoising predicts all active answer positions in parallel. Thus, each timestep exposes a distribution over every masked answer token, forming a token-by-timestep confidence trajectory. *OmniDenoise* uses this trajectory to compare denoising confidence under full and intervened contexts.

3.1 Evidence Denoising Gap

The full context denoising loss in Eq. 2 trains the model to reconstruct the answer, but it does not test whether reconstruction depends on the modalities that the task expects the model to use. We therefore introduce controlled modality interventions.

For each example, let $R(C) \subseteq \mathcal{M}$ denote the set of non-query modalities designated for intervention by the dataset schema or evaluation protocol. This set is defined at the modality level and does not require token level evidence annotations. For each $m \in R(C)$, we construct an intervened context

$$C^{\setminus m} = g_m(C), \quad (3)$$

where g_m replaces the content of modality m with a modality specific null or mask placeholder while preserving the query, remaining modalities, modality boundary tokens, and serialization template. For non-token inputs such as audio features, the corresponding feature tensors are zeroed when applicable. This keeps the input layout and answer corruption state fixed, so the full and intervened passes differ only in the designated modality content.

For the same corrupted answer $\tilde{y}_t$ and the same masked positions I_t , we compute token level de-

Stage 1 Evidence Denoising Gap

1 Full multimodal context

Q: Which object in the image matches the sound in the audio clip?

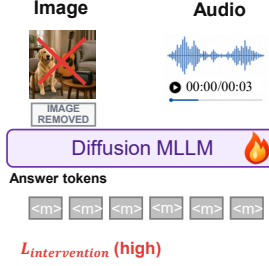


2 Intervention Operators

Q: Which object in the image matches the sound in the audio clip?



Q: Which object in the image matches the sound in the audio clip?



Stage 1 Loss

Stage 2 Bidirectional Consistency Distillation

1 Answer-token trajectory

Start y^0 :

Less denoised state $y(t > t^*)$:

the is .

More denoised state $y^*(t < t^*)$:

the is barking .

Ground truth y_{gt} :

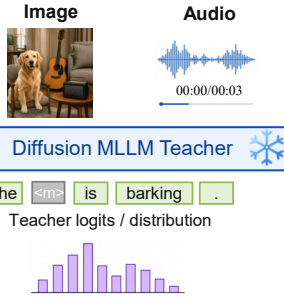
the dog is barking .

Token legend

Ground Truth Masked

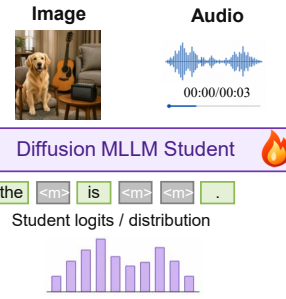
2 Bidirectional teacher

Q: Which object in the image matches the sound in the audio clip?



3 Bidirectional student

Q: Which object in the image matches the sound in the audio clip?



Stage 2 Loss

Figure 1: **Overview of OmniDenoise.** The proposed method uses the diffusion denoising trajectory as an intervention-sensitive training signal. By measuring how answer-token denoising loss changes after removing required modalities, OmniDenoise encourages the model to rely on multimodal evidence rather than shortcuts, and then applies bidirectional consistency distillation for faster inference.

noising losses under the full and intervened contexts:

$$\begin{aligned}\ell_{i,t}^{\text{full}} &= -\log p_{\theta}(y_i | C, \tilde{y}_t, t), \\ \ell_{i,t}^{\setminus m} &= -\log p_{\theta}(y_i | C^{\setminus m}, \tilde{y}_t, t).\end{aligned}\quad (4)$$

The Evidence Denoising Gap for modality m is

$$g_{i,t}^{(m)} = \ell_{i,t}^{\setminus m} - \ell_{i,t}^{\text{full}}. \quad (5)$$

A positive gap indicates that removing modality m increases the denoising loss of the gold token, meaning that the model becomes less confident in the correct answer token under the intervened context.

Since supervision is available at the answer level, we optimize the aggregate gap over masked answer positions:

$$G_t^{(m)} = \frac{1}{|I_t|} \sum_{i \in I_t} g_{i,t}^{(m)}. \quad (6)$$

The EDG loss enforces a margin between full-context and intervened-context denoising:

$$\mathcal{L}_{\text{edg}} = \mathbb{E}_{(C,y),t,I_t} \left[\max_{m \in R(C)} \left[\Delta - G_t^{(m)} \right]_+ \right], \quad (7)$$

where $[a]_+ = \max(a, 0)$ and Δ is a per-token negative log-likelihood margin. The margin prevents unbounded degradation under intervened contexts once sufficient separation from the full context has been reached. The outer maximum focuses training on the weakest designated intervention, so the loss becomes zero only when all designated modality removals produce the desired aggregate gap.

The Stage 1 objective is

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{ce} + \lambda_{\text{edg}} \mathcal{L}_{\text{edg}}. \quad (8)$$

EDG is computed using the same sampled timestep, mask set, and corrupted answer for the full and intervened contexts. This differs from ordinary modality dropout. *OmniDenoise* is not trained to be invariant to missing modalities, but to preserve accurate full-context prediction while reducing gold-token confidence when designated evidence is removed.

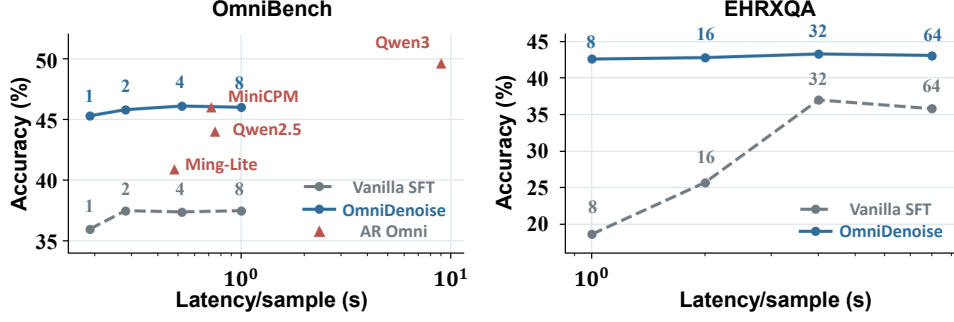


Figure 2: Accuracy latency trade-off on OmniBench and EHRXQA. Latency denotes mean generation time per example on a single NVIDIA A100 80GB GPU.

3.2 Few Step Bidirectional Consistency Distillation

Stage 1 trains an intervention sensitive diffusion model, but iterative denoising can still require many refinement steps. We therefore distill the EDG trained teacher into a few-step student. Let p_{θ_T} be the frozen Stage 1 teacher and p_ϕ be the student initialized from the teacher. The student is trained to map a less denoised answer state to the teacher behavior at a more denoised state while directly recovering newly revealed gold tokens. We retain the bidirectional answer attention of the diffusion backbone and distill only answer token denoising behavior, making this stage narrower than general purpose consistency diffusion language modeling (Minseo Kim, 2026).

The Stage 2 objective is

$$\mathcal{L}_{\text{stage2}} = w_{\text{jump}}\mathcal{L}_{\text{jump}} + w_{\text{kd}}\mathcal{L}_{\text{kd}} + w_{\text{sc}}\mathcal{L}_{\text{sc}}. \quad (9)$$

Here $\mathcal{L}_{\text{jump}}$ trains the student to recover newly revealed answer tokens, $\mathcal{L}_{\text{kd}}$ matches the EDG-trained teacher distribution on unresolved masked positions, and $\mathcal{L}_{\text{sc}}$ enforces student self consistency across denoising states. The full construction of these terms is provided in Appendix B.

3.3 Inference and Diagnostics

At inference time, *OmniDenoise* starts from a fully masked answer span of maximum length N_{max} and performs K bidirectional denoising steps under the full omni-modal context. At step k , the model predicts a distribution over all currently masked answer positions. We use maximum token probability as the default confidence score:

$$s_{i,k} = \max_v p_\phi(v \mid C, \tilde{y}^k, t_k). \quad (10)$$

The sampler finalizes the highest confidence masked positions according to a fixed step schedule

and fills them with their argmax predictions. Positions not selected at step k remain masked for the next refinement step. The same schedule is used across models in the same evaluation setting. The distilled student finalizes more positions per step than the Stage 1 teacher, improving the accuracy latency trade-off while retaining bidirectional answer refinement.

EDG is used during training and offline analysis, not as an additional inference time input. For diagnostics, we compute the full and intervened denoising losses from Eq. 4 on held out examples. We evaluate both the null or mask interventions used by EDG and stronger real-input mismatched replacements, and report whether gold-token confidence decreases under intervention on protocol-designated modalities. These diagnostics measure protocol level intervention sensitivity rather than fine-grained causal attribution.

4 Experiments

We evaluate *OmniDenoise* along six dimensions: full context omni-modal accuracy, accuracy and latency trade-offs, the contributions of Evidence Denoising Gap (EDG) training and consistency distillation (CD), sensitivity to controlled modality interventions, robustness of the intervention signal under real input mismatched replacement, and comparison with a matched autoregressive intervention objective. The evaluation covers three complementary settings: image, audio, and text reasoning; clinical question answering over structured records and radiographs; and out-of-distribution transfer to a held-out omni-modal benchmark.

4.1 Experimental Setup

We evaluate *OmniDenoise* on three complementary omni-modal settings. OmniBench evaluates image, audio, and text reasoning with 1,142 four choice questions; we report overall accuracy and accu-

racy by audio category. EHRXQA evaluates clinical question answering over structured electronic health records and chest X-ray images; following the answerable subset protocol, we evaluate on 526 test examples and report overall heuristic accuracy, Boolean accuracy, list exact match, and list recall. UNO-Bench is used only for out-of-distribution transfer evaluation. We filter its validation set to examples containing a text question, at least one image, and at least one audio clip, yielding 926 examples, and report overall, multiple choice, and multi output accuracy. No UNO-Bench examples are used for training, validation, or model selection. Dataset filtering details, split statistics, and category distributions are provided in Appendix A. Full prompting, answer parsing, decoding, latency measurement, intervention construction, and training configurations are provided in Appendix F.3.

Baselines. We compare with recent autoregressive omni-modal models: Qwen2.5-Omni (Xu et al., 2025a), Qwen3-Omni (Xu et al., 2025b), Ming-Lite-Omni (AI et al., 2025), and MiniCPM-o-4.5 (Yu et al., 2025), and diffusion baselines Omni-Diffusion (Li et al., 2026) and Dynin-Diffusion (Kim et al., 2026). Unless otherwise specified, *OmniDenoise* refers to the Stage 2 EDG-trained, distilled student used for inference. Off-the-shelf diffusion baselines are included as direct-transfer references, while Omni-Diffusion SFT is the primary adapted diffusion baseline trained under the same setting as *OmniDenoise*. We additionally train a matched Qwen2.5-Omni-7B pair: AR-SFT with standard teacher forced likelihood and AR-ELG with an autoregressive analogue of the intervention gap objective. The matched pair uses the same adaptation data, splits, prompts, update budget, decoding protocol, and intervention construction.

Evaluation protocol. All OmniBench models receive the full image, audio, and text context. For EHRXQA, models receive the natural language question, serialized EHR context, and associated CXR image. Invalid, empty, or unparsable responses are counted as incorrect. Latency is measured as mean wall-clock generation time per example. All *OmniDenoise* training runs use 4×NVIDIA A100 80GB GPUs, and latency measurements are conducted on a single NVIDIA A100 80GB GPU. We report denoising steps and output lengths together with latency in Appendix F.3 to make the operating points explicit.

Modality interventions. We evaluate two intervention types. A *null intervention* removes one protocol designated evidence source while preserving the question, remaining modalities, modality boundaries, and serialization structure. OmniBench examples are evaluated with either the image or audio intervened upon, while EHRXQA examples are evaluated with either the EHR context or CXR image intervened upon. A stronger *real input mismatched replacement* substitutes the designated modality with a modality compatible input sampled from another example, excluding self replacement, while preserving tensor and serialization structure. These diagnostics measure protocol level intervention sensitivity and are not intended as token level or per example causal attribution.

4.2 Main Results

Table 1 shows that *OmniDenoise* is the strongest adapted diffusion model on OmniBench. Compared with the Vanilla SFT diffusion baseline trained under the same tri-modal setting, EDG and distillation improve accuracy from 37.48% to 45.27% while reducing latency from 1.01s to 0.19s. The gain is especially large on Music (+29.68 points over SFT), suggesting that EDG is particularly beneficial in settings where acoustic evidence is difficult to replace with text or visual priors. Off-the-shelf diffusion baselines perform poorly under direct transfer, reflecting their limited exposure to simultaneous image, audio, and text conditioning rather than serving as the primary fairness comparison. Among autoregressive models, *OmniDenoise* remains competitive with Qwen2.5-Omni and MiniCPM-o-4.5, while Qwen3-Omni achieves higher absolute accuracy.

Table 2 evaluates structured EHR and CXR reasoning. *OmniDenoise* improves over the adapted SFT diffusion baseline from 36.9% to 42.6% H-Acc after distillation, while the EDG teacher reaches 43.9%. The gains are also visible on Boolean and list style answers. Although these improvements over the diffusion baseline are substantial, the absolute H-Acc remains modest, highlighting the difficulty of structured clinical multimodal reasoning. The matched autoregressive results reported in Section 4.6 further show that *OmniDenoise* does not establish universal accuracy superiority over autoregressive models. These results are intended as benchmark evidence for multimodal clinical QA, not as evidence of clinical deployment readiness.

Table 1: Main results on OmniBench. Accuracy is reported in percent and latency is mean generation time per example in seconds.

Model	Params	Acc. ↑	Music	Speech	Sound	Latency / sample ↓
<i>Autoregressive</i>						
Qwen2.5-Omni (Xu et al., 2025a)	7B	44.05	50.94	40.73	50.94	0.75
Qwen3-Omni (Xu et al., 2025b)	30B-A3B	49.65	57.55	46.30	56.23	8.95
Ming-Lite-Omni (AI et al., 2025)	19B-A2.8B	40.89	53.77	39.43	40.00	0.42
MiniCPM-o-4.5 (Yu et al., 2025)	9B	46.06	59.43	43.19	49.06	0.71
<i>Diffusion</i>						
Omni-Diffusion (Li et al., 2026)	8B	1.49	0.00	2.64	1.30	1.01
Dynin-Diffusion (Kim et al., 2026)	8B	21.19	19.81	19.46	26.79	1.40
Omni-Diffusion SFT	8B	37.48	26.42	36.71	44.15	1.01
<i>OmniDenoise</i>	8B	45.27	56.10	42.22	49.80	0.19

Table 2: Main results on the EHRXQA answerable test subset. Overall H-Acc is computed over all answerable examples; Boolean Acc is computed on Boolean questions; List Exact and List Recall are computed on list-style questions. All metrics are reported in percent.

Model	Params	Overall H-Acc ↑	Boolean Acc ↑	List Exact ↑	List Recall ↑
<i>Autoregressive</i>					
Qwen2.5-Omni (Xu et al., 2025a)	7B	37.6	57.3	10.0	13.8
Qwen3-Omni (Xu et al., 2025b)	30B-A3B	39.7	61.2	9.6	15.9
Ming-Lite-Omni (AI et al., 2025)	19B-A2.8B	32.9	53.1	4.6	6.3
MiniCPM-o-4.5 (Yu et al., 2025)	9B	40.5	65.5	5.5	8.9
<i>Diffusion</i>					
Omni-Diffusion (Li et al., 2026)	8B	18.4	29.6	2.7	3.4
Dynin-Diffusion (Kim et al., 2026)	8B	0.6	1.0	0.0	0.0
Omni-Diffusion SFT	8B	36.9	55.7	10.5	13.7
<i>OmniDenoise</i>	8B	42.6	64.5	11.9	17.8

Table 3: OmniBench component ablation. Accuracy is reported in percent; latency is mean generation time per example in seconds.

Model	Acc. ↑	Music	Speech	Sound	Latency / sample ↓
Vanilla SFT	37.48	26.42	36.71	44.15	1.01
+ EDG	46.41	58.50	43.00	51.50	1.01
+ EDG + CD	45.27	56.10	42.22	49.80	0.19

4.3 Component Analysis

Tables 3 and 4 isolate the contributions of EDG training and CD. EDG is the main source of the accuracy gain: it increases OmniBench accuracy from 37.48% to 46.41% (+8.93) and EHRXQA H-Acc from 36.9% to 43.9% (+7.0). CD introduces a small accuracy trade-off relative to the EDG teacher, decreasing OmniBench accuracy by 1.14 points and EHRXQA H-Acc by 1.3 points. This trade-off yields substantial speedups of $5.3\times$ on OmniBench and $3.9\times$ on EHRXQA. Thus, EDG accounts for the main accuracy improvement, while CD primarily improves the accuracy and latency operating point and preserves most of the intervention sensitive behavior.

4.4 EDG Diagnostic Visualization

Figure 3 directly evaluates the diffusion native diagnostic optimized by EDG. For each held-out example, we compute gold token denoising losses under the full and modality intervened contexts using the same corrupted answer state and masked answer positions. Compared with Vanilla SFT, EDG training substantially increases both the mean denoising loss gap and the fraction of examples with positive aggregate EDG. The consistency distilled student preserves most of this behavior, indicating that the efficiency gain from distillation does not erase the EDG signal. These diagnostics are consistent with the task level intervention results below and should be interpreted as measures of intervention sensitivity rather than fine-grained causal attribution.

4.5 Stronger Intervention Diagnostics

We first evaluate complete modality removal. On OmniBench, Vanilla SFT is nearly invariant to removing audio, changing from 37.48% to 37.5%, and drops only 1.6 points when the image is removed. By contrast, *OmniDenoise* drops 7.2 points without the image and 10.3 points without the audio while maintaining substantially higher full con-

Table 4: EHRXQA component ablation on the answerable test subset. All scores are reported in percent; latency is mean generation time per example in seconds.

Model	Overall H-Acc $\uparrow$	Boolean Acc $\uparrow$	List Exact $\uparrow$	List Recall $\uparrow$	Latency / sample $\downarrow$
Vanilla SFT	36.9	55.7	10.5	13.7	3.95
+ EDG	43.9	66.4	12.4	18.6	3.95
+ EDG + CD	42.6	64.5	11.9	17.8	1.01

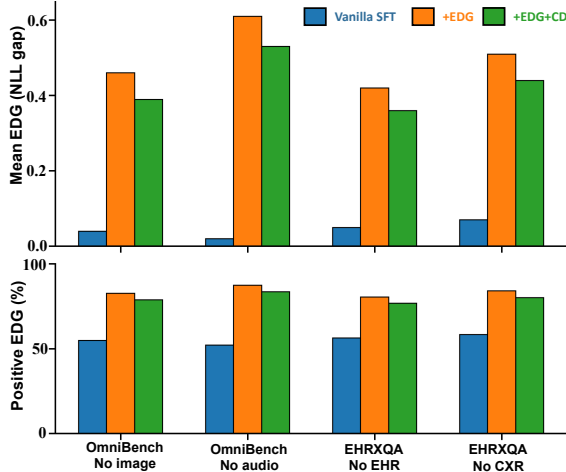


Figure 3: Held-out EDG diagnostics under controlled modality interventions. Top: mean increase in gold token denoising NLL after removing the designated modality. Bottom: percentage of examples with positive aggregate EDG, i.e., $G_t^{(m)} > 0$. Diagnostics are computed offline on held-out examples using matched corrupted answer states and are not used during inference. Higher values indicate stronger protocol level intervention sensitivity.

text accuracy. On EHRXQA, Vanilla SFT drops only 1.2 points without EHR and 2.1 points without CXR, whereas *OmniDenoise* drops 6.2 and 7.7 points, respectively. Detailed per category intervention results are provided in Appendix F.3.

Table 5 shows that the intervention effect persists when null inputs are replaced by real inputs from other examples. On OmniBench, the mismatch drop increases from 1.5 points for Vanilla SFT to 8.4 points after EDG training and remains 7.8 points after distillation. On EHRXQA, the corresponding relevant input mismatch drop increases from 1.9 to 7.2 points and remains 6.6 points after distillation. This substantially reduces the possibility that the observed effect is explained solely by detecting an artificial null placeholder.

On EHRXQA, we additionally evaluate an irrelevant/control intervention. The corresponding drops remain small at 0.8, 1.2, and 1.1 points for Vanilla SFT, +EDG, and +EDG+CD, respectively, while EDG Selectivity increases from 0.04 to 0.35 after EDG training and remains 0.31 after distilla-

Table 5: Full context performance and performance drops under null intervention and real input mismatched replacement. OmniBench reports accuracy and EHRXQA reports H-Acc. Null drops for OmniBench are averaged over image/audio interventions and those for EHRXQA over designated EHR/CXR interventions.

Dataset	Model	Full $\uparrow$	Null Drop $\uparrow$	Shuffle Drop $\uparrow$
OmniBench	Vanilla SFT	37.48	0.8	1.5
	+ EDG	46.41	9.6	8.4
	+ EDG + CD	45.27	8.8	7.8
EHRXQA	Vanilla SFT	36.9	1.7	1.9
	+ EDG	43.9	8.0	7.2
	+ EDG + CD	42.6	7.0	6.6

tion. Thus, EDG produces a substantially stronger response to interventions on protocol relevant evidence than to the control intervention. We interpret this result as *selective protocol level intervention sensitivity*, not as per example causal attribution.

Stronger relative intervention sensitivity does not imply a general collapse under missing modalities. Under complete removal, the average OmniBench accuracy is 36.70 for Vanilla SFT and 36.55 for *OmniDenoise*; on EHRXQA, the corresponding averages are 35.25 and 35.65. Thus, EDG mainly increases the performance benefit obtained from complete multimodal context rather than uniformly reducing absolute missing modality performance. These experiments do not establish robustness to arbitrary partial corruption, noisy-but-present modalities, or naturally missing evidence, which remain complementary objectives.

4.6 Matched Autoregressive Comparison

To separate the general intervention gap principle from its diffusion native realization, we train matched AR-SFT and AR-ELG models using Qwen2.5-Omni-7B. AR-ELG fixes the gold prefix and compares answer token negative log likelihood under full and intervened contexts using the same margin objective as EDG.

On OmniBench, AR-SFT reaches 47.55% full context accuracy with null and shuffle drops of 12.0 and 15.2 points, respectively. AR-ELG reaches 48.77% accuracy with corresponding drops of 12.7 and 15.3 points. Thus, the intervention objective

Table 6: Marginal effect of the intervention objective on OmniBench. Each row reports the difference between training with and without the corresponding intervention gap objective.

Objective effect	Full Acc. Δ	Null Drop Δ	Shuffle Drop Δ
Diffusion (EDG – SFT)	+8.9	+8.8	+6.9
AR (AR-ELG – AR-SFT)	+1.2	+0.7	+0.1

Table 7: OOD results on the filtered UNO-Bench validation subset. MC denotes multiple choice examples and MO denotes multi output examples. Overall is computed over 756 MC and 170 MO examples.

Model	Overall $\uparrow$	MC $\uparrow$	MO $\uparrow$
<i>Autoregressive</i>			
Qwen2.5-Omni (Xu et al., 2025a)	29.30	29.63	27.82
Qwen3-Omni (Xu et al., 2025b)	39.38	40.21	35.71
Ming-Lite-Omni (AI et al., 2025)	28.39	28.57	27.59
MiniCPM-o-4.5 (Yu et al., 2025)	31.94	32.41	<u>29.88</u>
<i>Diffusion</i>			
Omni-Diffusion (Li et al., 2026)	22.06	24.47	11.35
Dynin-Diffusion (Kim et al., 2026)	21.19	24.60	6.00
<i>OmniDenoise</i>	<u>35.75</u>	<u>37.17</u>	29.41

transfers to autoregressive training, but its marginal effect is substantially smaller than for the diffusion backbone.

Table 6 highlights this difference: EDG increases the diffusion model’s null and shuffle drops by 8.8 and 6.9 points, whereas AR-ELG changes them by only 0.7 and 0.1 points. This suggests that teacher forced autoregressive models are already substantially intervention sensitive, while the adapted diffusion baseline benefits much more strongly from explicit intervention gap training.

The EHRXQA matched comparison is more mixed. AR-SFT obtains 47.34% H-Acc, whereas AR-ELG obtains 45.25%. CXR-null sensitivity increases slightly from 6.27 to 6.84 points, but CXR-shuffle sensitivity remains small (0.19 versus 0.57 points), and EHR-null sensitivity decreases. We therefore do not claim a consistent autoregressive improvement or statistical significance from these single run results. Taken together, the objective transfers across architectures, but its largest and most consistent marginal benefit in our experiments is concentrated in the diffusion backbone.

4.7 Out-of-Distribution Evaluation

Table 7 reports transfer to the filtered UNO-Bench subset without additional fine-tuning. This setting evaluates generalization to a held-out omni-modal benchmark with a different data source and question distribution from the training and development benchmarks. *OmniDenoise* achieves 35.75% over-

all accuracy, improving over Omni-Diffusion by 13.69 points and over Dynin-Diffusion by 14.56 points. It is the strongest diffusion model and the second best model overall, within 3.63 points of Qwen3-Omni. The advantage is strongest on the MC subset; on MO, *OmniDenoise* is close to MiniCPM-o-4.5 but slightly lower. This suggests that open or multi output prediction remains more challenging for the current diffusion decoder.

5 Conclusion

We presented *OmniDenoise*, a diffusion MLLM framework for intervention sensitive omni-modal prediction. EDG improves full context performance and strengthens selective intervention sensitivity, including under real input mismatched replacement, while consistency distillation improves the accuracy and latency trade-off. A matched autoregressive analogue shows much smaller marginal gains, suggesting that the largest benefit of explicit intervention gap training is concentrated in the diffusion backbone. EDG should be interpreted as a protocol level intervention sensitivity signal rather than fine grained causal attribution; robustness to partial or noisy inputs and open ended generation remain future directions.

6 Acknowledgments

This work was partially supported by the National Institutes of Health (NIH) under Award No. 1R01EB037101-01 and by the National Science Foundation (NSF) under Award No. IIS-2551752. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH or NSF.

Limitations

OmniDenoise provides protocol level intervention sensitivity rather than fine-grained causal evidence attribution, especially when modalities are redundant or correlated. Real input mismatched replacement reduces concerns about artificial null modality cues, but does not establish causal evidence use or robustness to partial and noisy inputs. The evaluation is limited to benchmark answer prediction, and broader validation is needed for open ended generation, and real world distribution shifts.

References

- Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980.
- Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, and 1 others. 2025. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*.
- Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and 1 others. 2023. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36:3867–3880.
- Chen Chen, ZeYang Hu, Fengjiao Chen, Liya Ma, Jiaxing Liu, Xiaoyu Li, Ziwen Wang, Xuezhi Cao, and Xunliang Cai. 2025. Uno-bench: A unified benchmark for exploring the compositional law between uni-modal and omni-modal in omni models. *arXiv preprint arXiv:2510.18915*.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Daniel Israel, Guy Van den Broeck, and Aditya Grover. 2025. Accelerating diffusion llms via adaptive parallel decoding. In *Advances in neural information processing systems*, volume 38, pages 52870–52888.
- Jaeik Kim, Woojin Kim, Jihwan Hong, Yejoon Lee, Sieun Hyeon, Mintae Lim, Yunseok Han, Donggeun Kim, Hoeun Lee, Hyunggeun Kim, and 1 others. 2026. Dynin-omni: Omnimodal unified large diffusion language model. *arXiv preprint arXiv:2604.00007*.
- Lijiang Li, Zuwei Long, Yunhang Shen, Heting Gao, Haoyu Cao, Xing Sun, Caifeng Shan, Ran He, and Chaoyou Fu. 2026. Omni-diffusion: Unified multimodal understanding and generation with masked discrete diffusion. *arXiv preprint arXiv:2603.06577*.
- Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, and 1 others. 2025. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*.
- Yizhi LI, Yinghao Ma, Ge Zhang, Ruibin Yuan, King Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Moore Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Yidan WEN, Yanghai Wang, Shihao Li, Zhaoxiang Zhang, and 4 others. 2025. Omnibench: Towards the future of universal omni-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Yunxin Li, Xinyu Chen, Shenyuan Jiang, Haoyuan Shi, Zhenyu Liu, Xuanyu Zhang, Nanhao Deng, Zhenran Xu, Yicheng Ma, Meishan Zhang, and 1 others. 2025. Uni-moe-2.0-omni: Scaling language-centric omni-modal large model with advanced moe, training and data. *arXiv preprint arXiv:2511.12609*.
- Yihao Liang, Ze Wang, Hao Chen, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Emad Barsoum, Zicheng Liu, and Niraj K Jha. 2026. Cd4lm: Consistency distillation and adaptive decoding for diffusion language models. *arXiv preprint arXiv:2601.02236*.
- Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. 2022. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*.
- Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, and 1 others. 2025a. Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9087–9097. IEEE.
- Yudong Liu, Spencer Hallyburton, Jiwoo Kim, Yueqian Lin, Yiming Li, Qinsi Wang, Hui Ye, Jingwei Sun, Miroslav Pajic, Yiran Chen, and 1 others. 2025b. Llava: A large language vision driving assistant for explicit reasoning and enhanced trajectory planning. *arXiv preprint arXiv:2512.18211*.
- Zhiyuan Liu, Yicun Yang, Yaojie Zhang, Junjie Chen, Chang Zou, Qingyuan Wei, Shaobo Wang, and Linfeng Zhang. 2025c. dllm-cache: Accelerating diffusion large language models with adaptive caching. *arXiv preprint arXiv:2506.06295*.
- Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787.

- Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.
- Coleman Hooper Harman Singh Ben Athiwaratkun Ce Zhang Kurt Keutzer Amir Gholami Minseo Kim, Chenfeng Xu. 2026. Cdlm: Consistency diffusion language models for faster sampling. In *Ninth Conference on Machine Learning and Systems*.
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.
- Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2662–2670.
- Tim Salimans and Jonathan Ho. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.
- Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency models. In *ICML*, pages 32211–32252.
- Meituan LongCat Team, Bairui Wang, Bin Xiao, Bo Zhang, Bolin Rong, Borun Chen, Chang Wan, Chao Zhang, Chen Huang, Chen Chen, and 1 others. 2025. Longcat-flash-omni technical report. *arXiv preprint arXiv:2511.00279*.
- Qwen Team. 2026. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*.
- Xu Wang, Chenkai Xu, Yijie Jin, Jiachun Jin, Hao Zhang, and Zhijie Deng. 2025. Diffusion llms can do faster-than-ar inference via discrete diffusion forcing. *arXiv preprint arXiv:2508.09192*.
- Qingyan Wei, Yaojie Zhang, Zhiyuan Liu, Puyu Zeng, Yuxuan Wang, Biqing Qi, Dongrui Liu, and Linfeng Zhang. 2025. Accelerating diffusion large language models with slowfast sampling: The three golden principles. *arXiv preprint arXiv:2506.10848*.
- 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. 2025a. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.
- Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.
- Zeun Yang, Yake Wei, Ce Liang, and Di Hu. 2024. Quantifying and enhancing multi-modal robustness with modality preference. *arXiv preprint arXiv:2402.06244*.
- Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, and 1 others. 2025a. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*.
- Jiacheng Ye, Shansan Gong, Jiahui Gao, Junming Fan, Shuang Wu, Wei Bi, Haoli Bai, Lifeng Shang, and Lingpeng Kong. 2025b. Dream-vl & dream-vla: Open vision-language and vision-language-action models with diffusion language model backbone. *arXiv preprint arXiv:2512.22615*.
- Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025c. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.
- Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. 2025. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*.
- Zebin You, Xiaolu Zhang, Jun Zhou, Chongxuan Li, and Ji-Rong Wen. 2026. Llada-o: An effective and length-adaptive omni diffusion model. *arXiv preprint arXiv:2603.01068*.
- Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, and 1 others. 2025. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.
- Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, and 1 others. 2025. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. *arXiv preprint arXiv:2505.19223*.

A Dataset and Filtering Details

This appendix provides dataset statistics, filtering details, additional intervention experiments, matched autoregressive comparisons, and implementation details used in the camera-ready evaluation.

OmniBench. We use OmniInstruct-v1 as the instruction-tuning source for image, audio, and text examples, and evaluate tri-modal reasoning on OmniBench (LI et al., 2025). The local OmniInstruct-v1 mixed split contains 84,580 training examples and 11,525 validation examples. OmniBench contains 1,142 four-choice questions, each paired with an image, an audio clip, and a text question. The evaluated set contains 106 music questions, 771 speech questions, and 265 sound-event questions. We report overall multiple-choice accuracy and accuracy for each audio category.

EHRXQA. EHRXQA evaluates question answering over structured electronic health records and chest X-ray images (Bae et al., 2023). After filtering examples without usable ground-truth answers, the local adaptation split contains 4,135 training examples and 625 validation examples. For test evaluation, we use the 526 answerable examples obtained from the 640 source test questions by excluding 114 records without usable ground-truth answers. The evaluated subset contains 307 Boolean questions and 219 list questions. We report overall heuristic accuracy, Boolean accuracy, list exact match, and list recall.

UNO-Bench. UNO-Bench is used only for out-of-distribution transfer evaluation (Chen et al., 2025). It is not used for training, validation, hyperparameter tuning, or model selection. To match the evidence requirement studied in this paper, we retain validation examples containing a text question, at least one image, and at least one audio clip, and remove video-conditioned or missing-modality examples. This filtering yields 926 examples, including 756 UNO-Bench-MC examples and 170 UNO-Bench-MO examples. We report overall accuracy as well as accuracy on the MC and MO subsets.

B Bidirectional Consistency Distillation Details

This appendix gives the full construction of the few-step bidirectional consistency distillation objective

Dataset	Train	Validation	Test/Eval	Main metrics
OmniBench + OmniInstruct-v1	84,580	11,525	1,142	Acc.
EHRXQA	4,135	625	526	H-Acc., Bool., List
UNO-Bench filtered	—	—	926	Acc., MC, MO

Table 8: Dataset statistics used in our experiments. UNO-Bench is used only for transfer evaluation and is not used during training, validation, or model selection.

used in Stage 2. We use the convention that larger t corresponds to a more corrupted answer state. Let p_{θ_T} be the frozen EDG-trained teacher and p_ϕ be the student initialized from the teacher.

We first sample a less-denoised answer state $\tilde{y}^0$ at timestep t_0 and define the currently masked positions as

$$M_0 = \{i : \tilde{y}_i^0 = [\text{MASK}]\}. \quad (11)$$

We then sample a subset $U \subseteq M_0$ and reveal their ground-truth tokens to construct a more-denoised state $\tilde{y}^1$:

$$\tilde{y}_i^1 = \begin{cases} y_i, & i \in U, \\ \tilde{y}_i^0, & i \notin U. \end{cases} \quad (12)$$

Let $S = M_0 \setminus U$ denote positions that remain masked in $\tilde{y}^1$. We associate $\tilde{y}^1$ with timestep $t_1 < t_0$, reflecting that it is more denoised than $\tilde{y}^0$.

For the newly revealed positions U , the student is trained with a direct denoising loss:

$$\mathcal{L}_{\text{jump}} = \frac{1}{|U|} \sum_{i \in U} -\log p_\phi(y_i \mid C, \tilde{y}^0, t_0). \quad (13)$$

For the positions S that remain masked, we distill the teacher distribution at the more-denoised state into the student distribution at the less-denoised state:

$$\mathcal{L}_{\text{kd}} = \frac{1}{|S|} \sum_{i \in S} D_{\text{KL}} \left(\text{sg}[p_{\theta_T}(\cdot \mid C, \tilde{y}^1, t_1)] \parallel p_\phi(\cdot \mid C, \tilde{y}^0, t_0) \right), \quad (14)$$

where $\text{sg}[\cdot]$ stops gradients through the teacher. The KL term is computed only on S , which prevents the teacher target from being conditioned on the gold token at the same position.

We also add a student self-consistency loss on the remaining masked positions:

$$\mathcal{L}_{\text{sc}} = \frac{1}{|S|} \sum_{i \in S} D_{\text{KL}} \left(\text{sg}[p_\phi(\cdot \mid C, \tilde{y}^1, t_1)] \parallel p_\phi(\cdot \mid C, \tilde{y}^0, t_0) \right). \quad (15)$$

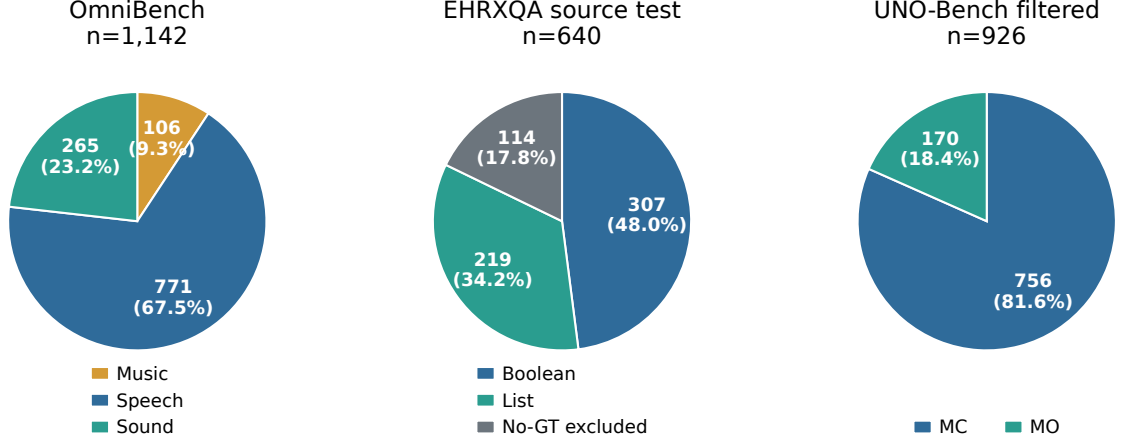


Figure 4: Effective evaluation data. OmniBench is grouped by audio category; EHRXQA reports the answerable test subset after removing examples without usable ground-truth answers; UNO-Bench reports the filtered image, audio, and text question validation subset used only for transfer evaluation.

The complete Stage 2 objective is

$$\mathcal{L}_{\text{stage2}} = w_{\text{jump}}\mathcal{L}_{\text{jump}} + w_{\text{kd}}\mathcal{L}_{\text{kd}} + w_{\text{sc}}\mathcal{L}_{\text{sc}}. \quad (16)$$

If U or S is empty, the corresponding term is omitted. This objective teaches the student to recover more answer tokens per denoising step while matching the EDG-trained teacher on unresolved positions.

C Additional Modality Intervention Experiments

This section provides detailed results for the modality-intervention experiments summarized in the main paper. We distinguish between the null/mask intervention used during EDG training and a stronger real-input mismatched replacement used for evaluation.

C.1 Intervention Construction

For a designated modality m , the null intervention modifies only its content while preserving the query, all remaining modalities, modality boundary tokens, tensor layout, and serialization template. Image and audio inputs are zeroed where applicable, while textual or structured spans are replaced by the corresponding null or mask representation. The full and intervened EDG passes use the same timestep, mask set, and corrupted answer state.

For real-input mismatched replacement, the designated image, audio, EHR, or CXR input is instead replaced by a real input of the same modality drawn from another example. Self-replacement is

excluded. This preserves the modality type and input structure while disrupting the correspondence between the designated evidence and the target answer.

C.2 Detailed Complete-Removal Results

Tables 9 and 10 provide the detailed modality-removal results that complement the compact analysis in the main paper.

C.3 Real-Input Mismatched Replacement

Table 11 compares null removal with real-input mismatched replacement on OmniBench. Drops are computed as full-context accuracy minus intervened accuracy and are averaged over image and audio interventions.

The shuffle drop increases from 1.5 points for Vanilla SFT to 8.4 points after EDG training and remains 7.8 points after consistency distillation. The effect therefore persists when the intervened modality contains a structurally valid real input rather than an artificial null placeholder.

On the 526-example EHRXQA answerable subset, we additionally compare interventions on protocol-relevant evidence with an irrelevant/control intervention.

Relevant real-input shuffling produces a 7.2-point drop after EDG training, compared with 1.9 points for Vanilla SFT, while the control-intervention drops remain small. EDG Selectivity increases from 0.04 to 0.35 and remains 0.31 after consistency distillation. We interpret this pattern as *selective protocol-level intervention sensitivity*, rather than fine-grained causal attribution.

Table 9: OmniBench performance under complete modality removal. Each cell reports the intervened score and the absolute point change relative to the corresponding full-context score.

Model	No Image				No Audio			
	Overall	Music	Speech	Sound	Overall	Music	Speech	Sound
<i>Autoregressive</i>								
Qwen2.5-Omni	29.3 (-14.8)	49.1 (-1.8)	24.6 (-16.1)	34.7 (-16.2)	33.4 (-10.7)	44.3 (-6.6)	32.6 (-8.1)	31.3 (-19.6)
Qwen3-Omni	36.3 (-13.4)	50.9 (-6.7)	31.7 (-14.6)	44.2 (-12.0)	38.5 (-11.2)	58.5 (+1.0)	36.7 (-9.6)	35.8 (-20.4)
Ming-Lite-Omni	25.4 (-15.5)	39.6 (-14.2)	23.0 (-16.4)	26.8 (-13.2)	34.5 (-6.4)	50.9 (-2.9)	32.0 (-7.4)	35.1 (-4.9)
MiniCPM-o-4.5	37.0 (-9.1)	55.7 (-3.7)	32.9 (-10.3)	41.5 (-7.6)	37.0 (-9.1)	49.1 (-10.3)	36.8 (-6.4)	32.5 (-16.6)
<i>Diffusion</i>								
Omni-Diffusion SFT	35.9 (-1.6)	26.4 (+0.0)	34.8 (-1.9)	43.0 (-1.2)	37.5 (+0.0)	25.5 (-0.9)	36.7 (+0.0)	44.5 (+0.4)
OmniDenoise	38.1 (-7.2)	53.4 (-2.7)	34.6 (-7.6)	39.2 (-10.6)	35.0 (-10.3)	42.8 (-13.3)	32.9 (-9.3)	34.7 (-15.1)

Table 10: EHRXQA performance under complete modality removal. Each cell reports the intervened score and the absolute point change relative to the corresponding full-context score.

Model	No EHR				No CXR			
	Overall	Boolean	List Exact	List Recall	Overall	Boolean	List Exact	List Recall
<i>Autoregressive</i>								
Qwen2.5-Omni	35.0 (-2.6)	55.4 (-1.9)	6.4 (-3.6)	10.6 (-3.2)	34.8 (-2.8)	56.0 (-1.3)	5.0 (-5.0)	6.3 (-7.5)
Qwen3-Omni	37.1 (-2.6)	55.4 (-5.8)	11.4 (+1.8)	16.0 (+0.1)	32.7 (-7.0)	51.8 (-9.4)	5.9 (-3.7)	7.7 (-8.2)
Ming-Lite-Omni	31.6 (-1.3)	48.9 (-4.2)	18.3 (+13.7)	11.6 (+5.3)	34.0 (+1.1)	55.7 (+2.6)	3.7 (-0.9)	4.5 (-1.8)
MiniCPM-o-4.5	41.3 (+0.8)	65.5 (+0.0)	7.3 (+1.8)	11.6 (+2.7)	35.4 (-5.1)	57.3 (-8.2)	4.6 (-0.9)	5.3 (-3.6)
<i>Diffusion</i>								
Omni-Diffusion SFT	35.7 (-1.2)	53.1 (-2.6)	11.4 (+0.9)	16.5 (+2.8)	34.8 (-2.1)	48.2 (-7.5)	16.0 (+5.5)	20.8 (+7.1)
OmniDenoise	36.4 (-6.2)	56.8 (-7.7)	7.5 (-4.4)	10.2 (-7.6)	34.9 (-7.7)	54.2 (-10.3)	6.4 (-5.5)	8.7 (-9.1)

Table 11: OmniBench full-context accuracy and performance drops under null intervention and real-input mismatched replacement.

Model	Full Acc. $\uparrow$	Null Drop $\uparrow$	Shuffle Drop $\uparrow$
Vanilla SFT	37.48	0.8	1.5
+ EDG	46.41	9.6	8.4
+ EDG + CD	45.27	8.8	7.8

C.4 Intervention Sensitivity and Missing-Modality Performance

Larger relative drops under intervention do not imply a general collapse in absolute missing-modality performance. On OmniBench, Vanilla SFT obtains 35.9 without the image and 37.5 without audio, for an average of 36.70, whereas *OmniDenoise* obtains 38.1 and 35.0, for an average of 36.55. On EHRXQA, the corresponding removal averages are 35.25 for Vanilla SFT and 35.65 for *OmniDenoise*.

Thus, in the evaluated complete-removal settings, EDG primarily increases the benefit obtained from the full multimodal context without causing a general reduction in absolute missing-modality performance relative to Vanilla SFT. These experiments do not establish robustness to arbitrary partial corruption or noisy-but-present modalities, which remains a complementary objective.

D Matched Autoregressive Intervention Experiments

To separate the general intervention-gap principle from its diffusion-native realization, we construct a matched autoregressive analogue using Qwen2.5-Omni-7B. AR-SFT uses standard teacher-forced likelihood, while AR-ELG fixes the gold prefix and compares answer-token negative log-likelihood under full and intervened contexts using the same margin formulation. The matched pair uses the same data, splits, prompts, update budget, decoding protocol, and intervention construction.

D.1 OmniBench

AR-SFT is already strongly sensitive to both null and shuffled interventions. AR-ELG slightly increases full-context accuracy and produces only small changes in intervention sensitivity.

The intervention-gap objective therefore transfers in form across architectures, but its marginal effect is substantially larger on the diffusion backbone in this experiment. This motivates our characterization of EDG as a *diffusion-native instantiation* of an intervention-gap objective rather than an objective that is exclusive to diffusion models.

Table 12: EHRXQA results under relevant null removal, relevant real-input mismatched replacement, and an irrelevant/control intervention. EDG Selectivity is $\text{mean}(\text{EDG}_{\text{relevant}}) - \text{mean}(\text{EDG}_{\text{irrelevant}})$.

Model	Full H-Acc. $\uparrow$	Relevant Null Drop $\uparrow$	Relevant Shuffle Drop $\uparrow$	Irrelevant Drop $\downarrow$	EDG Selectivity $\uparrow$
Vanilla SFT	36.9	1.7	1.9	0.8	0.04
+ EDG	43.9	8.0	7.2	1.2	0.35
+ EDG + CD	42.6	7.0	6.6	1.1	0.31

Table 13: Matched autoregressive comparison on OmniBench.

Model	Full Acc. $\uparrow$	Null Drop $\uparrow$	Shuffle Drop $\uparrow$
AR-SFT	47.55	12.0	15.2
AR-ELG	48.77	12.7	15.3

Table 14: Per-condition intervention drops for the matched autoregressive models on OmniBench.

Condition	AR-SFT Drop	AR-ELG Drop
Null audio	12.08	12.70
Null image	11.91	12.78
Shuffle audio	11.21	11.73
Shuffle image	19.09	18.83

D.2 EHRXQA

The EHRXQA effect is mixed. AR-ELG slightly increases CXR-null sensitivity, while full-context H-Acc and EHR-null sensitivity decrease. CXR-shuffle drops remain close to zero for both autoregressive models. Because these results are based on single runs, we do not claim statistical significance for these small differences. Overall, the matched comparisons do not support universal superiority of the intervention objective across architectures; its largest and most consistent marginal effect in our experiments occurs on the diffusion backbone.

E Additional Diagnostic and Efficiency Results

E.1 Numerical EDG Diagnostics

Table 17 reports the numerical values corresponding to the held-out EDG diagnostic visualization in the main paper.

EDG training substantially increases both the mean denoising-loss gap and the fraction of examples with positive aggregate EDG. Consistency distillation slightly reduces the magnitude while preserving most of the intervention-sensitive behavior. These values are used as offline diagnostics and are not additional inference-time inputs.

Table 15: Marginal effect of the intervention-gap objective on OmniBench.

Objective effect	Full Acc. Δ	Null-Drop Δ	Shuffle-Drop Δ
Diffusion (EDG – SFT)	+8.9	+8.8	+6.9
AR (AR-ELG – AR-SFT)	+1.2	+0.7	+0.1

Table 16: Matched autoregressive results on the 526-example EHRXQA answerable subset.

Model	Full H-Acc. $\uparrow$	CXR-Null Drop $\uparrow$	CXR-Shuffle Drop $\uparrow$	EHR-Null Drop $\uparrow$
AR-SFT	47.34	6.27	0.19	2.85
AR-ELG	45.25	6.84	0.57	0.57

E.2 Latency Operating Points

All latency measurements report mean wall-clock generation time per example on a single NVIDIA A100 80GB GPU. Within each dataset, diffusion models use the same prompt, answer span, parser, and hardware. We report denoising steps separately from output length because diffusion and autoregressive models operate under different decoding schemes.

Relative to Vanilla SFT, the one-step distilled model yields a $5.3\times$ speedup on OmniBench and a $3.9\times$ speedup on EHRXQA. The autoregressive rows are provided as hardware-matched latency references; their output lengths differ from the fixed answer spans used by the diffusion models.

F Reproducibility and Responsible Research Details

This section consolidates additional implementation and responsible-research details and clarifies how the reported experiments should be interpreted.

F.1 Scientific Artifacts, Licenses, and Intended Use

Artifacts used. The study uses existing multimodal datasets and benchmarks, including OmniInstruct-v1/OmniBench, EHRXQA, and the filtered UNO-Bench validation subset. It also evaluates against publicly described autoregressive and diffusion multimodal models cited in the main pa-

Table 17: Numerical values for the held-out EDG diagnostics. Positive EDG denotes the percentage of examples with positive aggregate EDG.

Intervention	Model	Mean EDG	Positive EDG (%)
OmniBench No Image	Vanilla SFT	0.04	55
	+ EDG	0.46	91
	+ EDG + CD	0.38	88
OmniBench No Audio	Vanilla SFT	0.02	52
	+ EDG	0.61	94
	+ EDG + CD	0.53	92
EHRXQA No EHR	Vanilla SFT	0.05	56
	+ EDG	0.42	88
	+ EDG + CD	0.35	84
EHRXQA No CXR	Vanilla SFT	0.07	60
	+ EDG	0.51	91
	+ EDG + CD	0.44	89

Table 18: Generation latency on OmniBench.

Model	Steps	Latency	Output tokens
<i>OmniDenoise</i>	1	0.19 s	16
Vanilla SFT	8	1.01 s	16
Qwen2.5-Omni	–	0.75 s	1

per. The accompanying implementation includes training scripts, modality-intervention construction, model and optimization configurations, dataset filtering and serialization utilities, evaluation and scoring scripts, offline EDG diagnostics, and latency measurement.

Licenses and terms. The experiments rely on artifacts under their original public licenses, data-use terms, or access conditions. We do not redistribute third-party datasets, patient records, raw benchmark items, or third-party model weights. Users of the released implementation should obtain the underlying datasets and baseline models from their original providers and comply with their respective terms.

Intended use. The artifacts and derived results are intended for research on omni-modal reasoning, diffusion language modeling, and model analysis. The clinical benchmark results are aggregate research evaluations only and are not evidence of clinical deployment readiness.

F.2 Privacy, Sensitive Content, and Data Handling

No new dataset was collected for this work. EHRXQA contains clinical question-answering instances over electronic health records and chest X-ray images from an existing benchmark. We use the answerable subset only for aggregate evalua-

Table 19: Generation latency on EHRXQA under the same hardware protocol.

Model	Steps	Latency	Output tokens
Vanilla SFT	8	3.95 s	128
<i>OmniDenoise</i>	1	1.00 s	128
AR (Qwen-Omni)	–	2.22 s	3.6

tion, do not reproduce individual patient records or images in the paper, and do not attempt re-identification. OmniBench and UNO-Bench may include real-world image or audio content; we use their released benchmark format and report only aggregate metrics. All experiments are conducted under the access conditions of the corresponding artifacts.

F.3 Experimental Setup, Hyperparameters, and Compute

The adapted diffusion baseline and *OmniDenoise* use an 8B Omni-Diffusion backbone. Training runs use 4 NVIDIA A100 80GB GPUs, and all latency measurements are conducted on a single NVIDIA A100 80GB GPU.

The method uses two training stages. Stage 1 optimizes masked answer denoising together with EDG. Stage 2 distills the EDG-trained teacher into a few-step bidirectional denoising student. *OmniDenoise* introduces no architectural modification or additional trainable parameters relative to the Vanilla SFT diffusion baseline; Stage 1 uses the same backbone, optimizer, schedule, and training budget, with the EDG margin objective added to the standard denoising loss.

All supervised fine-tuning runs start from the Omni-Diffusion checkpoint and use the same tokenizer stack: GLM-4-Voice/SenseVoice for audio tokens and MAGVIT-v2 for image tokens. Unless otherwise specified, training uses PyTorch distributed data parallelism, DeepSpeed ZeRO-3 optimizer offload, bf16 model weights, TF32 matrix multiplications, FlashAttention-2, gradient checkpointing, maximum gradient norm 1.0, weight decay 0, AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 10^{-8}$, together with a cosine learning-rate schedule and 3% warmup. The audio encoder is frozen during fine-tuning. Both the training seed and data seed are fixed to 956.

F.4 Result Reporting and Evaluation

Unless otherwise stated, reported accuracies are single-run results on fixed evaluation splits. The

Table 20: Key training and optimization configurations for the OmniBench and EHRXQA experiments.

Setting	OmniBench	EHRXQA
Base model	Omni-Diffusion 8B	Omni-Diffusion 8B
Hardware	4×A100 80GB	4×A100 80GB
Training data	OmniInstruct-v1 (4,491 MCQ + 80,089 open-ended)	EHRXQA answerable func_vqa
Sequence length	3072	8192
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$)	Same
Learning rate / schedule	1×10^{-5} , cosine, 3% warmup	Same
Effective batch size	64	32
Maximum steps / epochs	3000 / 2.27	150 / 1.16
Precision / attention	bf16 / FlashAttention-2	Same
EDG λ_{edg} / margin Δ	0.1 / 0.1	0.1 / 0.1
Required modalities	Image, audio	CXR, EHR
Consistency distillation	Forward KL; reveal prob. 0.5; self-consistency weight 0.5	Same

paper reports aggregate accuracy for OmniBench and UNO-Bench, and H-Acc, Boolean accuracy, list exact match, and list recall for EHRXQA. Invalid, empty, or unparsable responses are counted as incorrect.

Evaluation uses task-specific answer parsing and exact or heuristic scoring rules described in the experimental setup. No new human annotation is introduced. The accompanying implementation provides the intervention, filtering, serialization, decoding, scoring, and latency-evaluation procedures used for the reported experiments.

F.5 Human Subjects and AI Assistance

The work does not recruit participants, hire annotators, or conduct a new human-subjects study. Therefore, participant instructions, recruitment or payment details, and new ethics-board approval for newly collected human-subject data are not applicable. The work relies on existing datasets and follows their original access conditions.

An AI assistant was used for writing support. The authors reviewed and revised AI-assisted text and remain responsible for all claims, experimental results, citations, and final content.