WHAT SHOULD THE MISSING MODALITY SAY? STRUCTURED CANDIDATE RECOVERY FOR INCOMPLETE MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Missing-modality recovery is often underdetermined: the same partial observation can support several task-relevant interpretations of what an absent modality would express. Deterministic reconstruction may collapse these alternatives, while stochastic recovery produces samples without explicitly associating them with distinct semantic interpretations. We propose **HyDiff**, which represents this ambiguity with a finite, data-dependent set of structured candidates. A frozen multimodal reasoner proposes candidates, a learned compatibility scorer assigns operational weights, and candidate- and target-specific retrieval supplies instance-grounded priors. Conditional diffusion then recovers each missing target under the retained candidates, after which task predictions are aggregated with their weights. Candidate-weight entropy is used for selective prediction; the weights are defined only over the reasoner-generated set rather than treated as a human interpretation posterior. The design therefore separates uncertainty over structured interpretations from the realization of each recovered modality and keeps competing semantic alternatives explicit rather than leaving them as anonymous stochastic samples. Experiments on three multimodal sentiment benchmarks evaluate feature- and modality-level missingness with matched-input, sampling, full-anchor, and reduced-anchor controls. When explicit target-facing fields are removed from the shared anchor, HyDiff retains a smaller but consistent advantage over matched P-RMF, separating semantic assistance from the recovery architecture; additional transfer, completion, candidate-coverage, and efficiency diagnostics are reported in the appendix.

1 INTRODUCTION

Multimodal systems are commonly trained with complete inputs but deployed when one or more modalities are unavailable because of sensor failure, occlusion, transmission constraints, or privacy restrictions Baltrušaitis et al. (2019); Yin et al. (2024); Wu et al. (2024). Existing incomplete-multimodal methods therefore either learn robust representations from the available streams or reconstruct the missing ones before prediction Ma et al. (2021); Zhu et al. (2025); Dai et al. (2025). A central difficulty is that the missing modality is not always uniquely determined by what remains. Consider a multimodal sentiment example in which the transcript is missing while the available audio is flat and the face is restrained. The observed evidence can be consistent with more than one sentiment-relevant reading, such as mild approval or quiet frustration. A recovery system that commits to one completion before task prediction hides this ambiguity from the downstream model. This issue is distinct from the expressive capacity of a generator. Under expected squared error, a deterministic regressor minimizes risk at the conditional mean, which can average across separated modes; mixture-density and multiple-hypothesis formulations were developed to avoid this collapse Bishop (1994); Rupprecht et al. (2017). Conditional diffusion models can in principle represent multimodal conditional distributions Ho et al. (2020); Rombach et al. (2022); Esser et al. (2024), but a standard recovery pipeline still has to decide how sampled completions are used by the predictor. Similarly, retrieval can provide instance-specific structure, yet a nearest neighbor need not correspond to the task-relevant interpretation supported by the current evidence Khandelwal et al. (2020); Blattmann et al. (2022). We therefore focus on the representation and use of ambiguity rather than arguing that stochastic generators themselves are incapable of modeling it.

HyDiff represents ambiguity at the level of structured, decision-relevant candidate interpretations. The candidates are not assumed to enumerate all human-plausible explanations, and their normalized

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

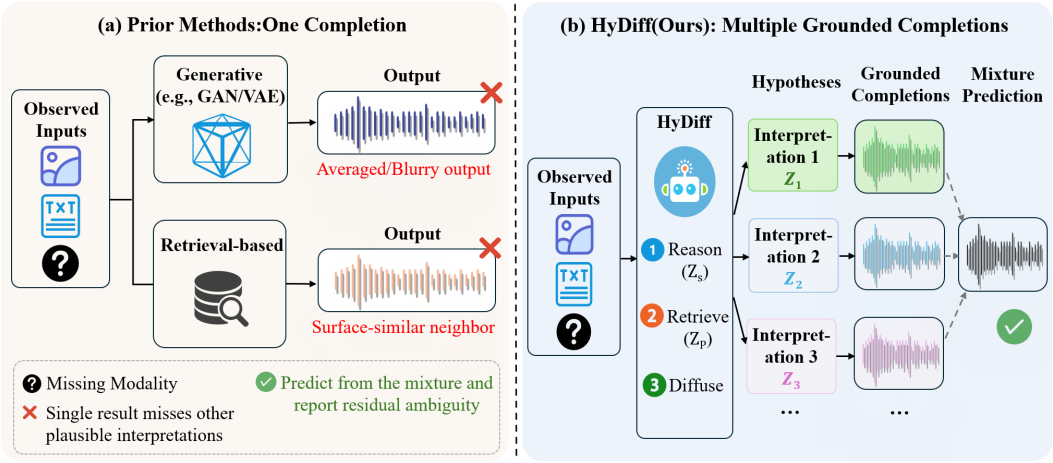

Figure 1: Missing-modality recovery can be underdetermined. (a) Point reconstruction may average distinct valid regions, while stochastic sampling leaves alternatives unstructured. (b) HyDiff retains weighted semantic candidates, performs candidate-conditioned target-wise recovery, and aggregates predictions across branches.

weights are not treated as a ground-truth posterior. Instead, they form a data-dependent candidate set proposed by a frozen multimodal reasoner and ranked by a separately trained compatibility scorer. This representation keeps alternative interpretations explicit: each candidate can condition a separate recovery branch, and the resulting alternatives remain inspectable rather than becoming anonymous samples in feature space. It also enables candidate-level diagnostics of coverage and failure instead of treating branch diversity as latent sampling noise. The connection to classical multiple-choice and multi-hypothesis prediction is discussed in Sec. 2.

HyDiff has three operational stages. First, the frozen reasoner iteratively proposes structured candidates from the available modalities, and a lightweight cross-attention scorer converts candidate–evidence compatibility into temperature-scaled weights. Second, for each retained candidate and missing target modality, HyDiff performs candidate- and target-specific retrieval from the corresponding training memory to obtain instance-specific structure, with a coverage gate falling back to a learned prior when retrieval support is weak. Third, a conditional diffusion model recovers each missing target stream under separate semantic and target-specific structural conditioning, with late guidance from a candidate-compatibility surrogate. For samples with multiple missing modalities, the target-wise recoveries are assembled into one candidate-indexed completion before the shared task head is applied. The downstream predictions are then aggregated with the candidate weights. The normalized entropy of those weights is returned as a candidate-set ambiguity score for selective prediction.

The reasoner has access to raw modalities in the raw-assisted protocol, whereas most baselines operate on pre-extracted features. We therefore separate matched-input and matched-anchor controls from the headline comparison. To address task-aligned semantic shortcuts more directly, we additionally remove `polarity`, `intensity`, and `dominant_affect` and give P-RMF and HyDiff the same reduced-anchor protocol. Under this stricter comparison, HyDiff is higher in all six dataset–protocol settings by 0.45–0.79 points (0.62 on average), with positive differences in $26/30$ paired seed runs. The result indicates that task-aligned semantics explain a substantial portion of the headline gain, while the candidate-indexed recovery architecture contributes a smaller but consistent residual benefit; statistical details are reported in Sec. 4 and Appendix C.6.

We make three **contributions**. *(i)* We formulate incomplete-modality recovery as candidate-indexed conditional generation, separating uncertainty over task-relevant interpretations from the realization of each recovered modality. *(ii)* We instantiate this formulation in **HyDiff**, which combines iterative candidate generation and compatibility weighting, candidate- and target-specific retrieval, and target-wise candidate-conditioned diffusion before decision-level aggregation. *(iii)* We evaluate the framework with matched-input and full/reduced-anchor controls, stochastic controls with approximately matched trajectory counts, completion-fidelity and efficiency diagnostics, and selective prediction, explicitly separating semantic assistance, generic sampling, and the residual contribution of candidate-indexed recovery.

2 Related Work

Incomplete multimodal learning. Methods for incomplete multimodal learning broadly fall into robust representation learning and explicit modality recovery. Representation-oriented approaches improve prediction from partial inputs through distillation, prompt/fusion mechanisms, correlation modeling, reliability weighting, or parameter-efficient adaptation Ma et al. (2021); Zhang et al. (2024); Li et al. (2024); Guo et al. (2024); Lin et al. (2025); Zhao et al. (2026). Recent work broadens this robustness view: BALM addresses modality-dependent missing-rate imbalance through feature calibration and gradient rebalancing Nguyen et al. (2026), while GIML unifies whole-modality absence with intra-modality degradation through dynamic quality perception Meng & Wei (2026). P-RMF additionally learns a stable proxy modality for robust fusion Zhu et al. (2025). Recovery-oriented methods reconstruct or imagine missing information through feature imagination, distribution-consistent recovery, or diffusion Zhao et al. (2021); Wang et al. (2023a;b); Dai et al. (2025), while recent work makes imputation uncertainty or recovery reliability more explicit He et al. (2026a); Zou et al. (2026). These methods improve robustness or recovery quality, but do not expose a finite set of competing semantic interpretations before reconstruction. HyDiff instead makes such alternatives explicit and uses each retained candidate to index target-specific grounding and recovery.

Retrieval and large-model assistance. Retrieval can supplement incomplete observations with training examples or external multimodal context. RAGPT retrieves similar instances to build dynamic prompts for incomplete multimodal transformers Lang et al. (2025), while MissRAG combines multimodal retrieval with prompt engineering for missing-input robustness in multimodal large language models Pipoli et al. (2025). Knowledge Bridger uses structured knowledge extracted by a large multimodal model to guide training-free completion and ranking Ke et al. (2025). RAG4DMC retrieves in-domain and external context to guide generation and then selects among generated completions He et al. (2026b), while ANGA combines retrieval-based reconstruction with semantic prompting Guan et al. (2026); retrieved examples can also provide priors for diffusion generation Blattmann et al. (2022). HyDiff differs in ordering and use: semantic candidates are formed first, each candidate–target pair issues its own retrieval query, and the resulting branches remain distinct through recovery and downstream prediction rather than being reduced to one selected completion.

Multiple hypotheses and uncertainty. Multiple-choice learning, mixture density networks, and multi-hypothesis prediction retain several possible outputs instead of forcing a single point estimate Guzman-Rivera et al. (2012); Bishop (1994); Rupprecht et al. (2017). Conditional generative models, including diffusion, can likewise represent multimodal output distributions Ho et al. (2020); Rombach et al. (2022). HyDiff differs mainly in where the alternatives are represented: short structured interpretations index candidate- and target-specific retrieval and recovery branches rather than anonymous output heads or independent samples. Their normalized compatibility weights are used operationally for prediction and selective prediction Geifman & El-Yaniv (2017); because the candidate set is generated by a frozen reasoner and calibrated against pseudo-references, we do not equate these weights with a complete human interpretation posterior.

3 Method

HyDiff represents missing-modality uncertainty with weighted structured candidates, grounds each candidate–target pair in training prototypes, performs target-wise conditional diffusion, and aggregates task predictions (Fig. 2).

3.1 Problem Formulation

Setup. Let $\mathcal{M}_{\text{obs}}$ and $\mathcal{M}_{\text{miss}}$ denote the observed and missing subsets of modalities $\mathcal{M} = \{v, a, \text{txt}\}$. We write X^{obs} and X^{miss} for their pre-extracted feature streams; only the frozen reasoner may additionally observe available raw streams R^{obs}. Recovery produces $\hat{X}^{\text{miss}}$ for downstream prediction.
Set-valued recovery. Partial evidence can support several task-relevant interpretations of a missing modality, so $p^\star(X^{\text{miss}} \mid X^{\text{obs}})$ need not be concentrated Zeng et al. (2024); Zhang et al. (2025); Dai et al. (2025). Introducing a discrete interpretation variable $z \in \mathcal{Z}_s$ gives

$$p(X^{\text{miss}} \mid X^{\text{obs}}) = \sum_{z \in \mathcal{Z}_s} p(z \mid X^{\text{obs}})\, p(X^{\text{miss}} \mid X^{\text{obs}}, z). \tag{1}$$

HyDiff approximates the first factor with a data-dependent candidate set and, for missing target m, the second with $p_\theta(X_m \mid \bar{z}_s, z_{p,m})$. The anchor and target-specific prior are derived from observed

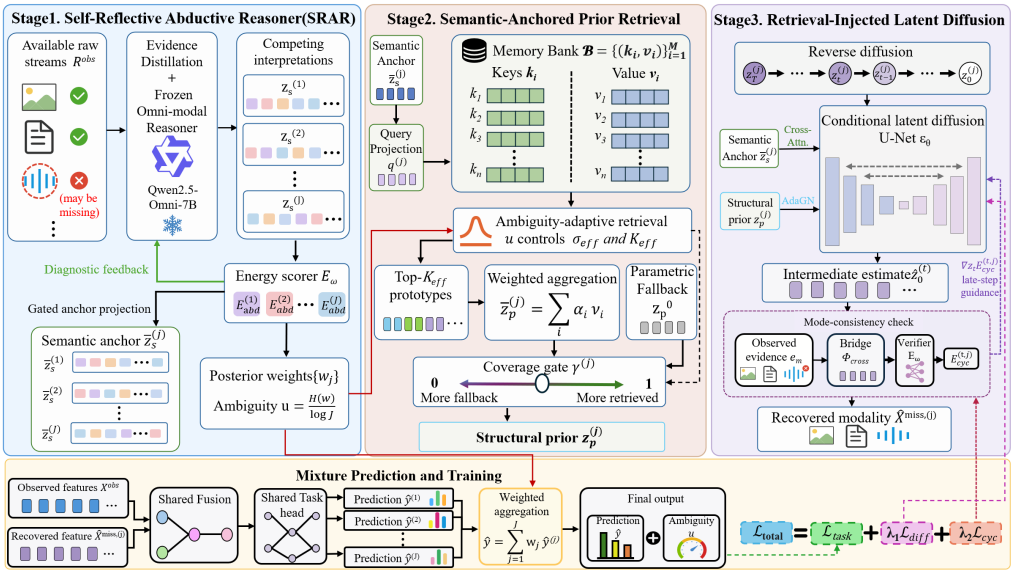

Figure 2: HyDiff has three stages: (a) iterative structured candidate generation and compatibility weighting, (b) candidate- and target-specific prototype retrieval with ambiguity-adaptive breadth and coverage fallback, and (c) candidate-conditioned latent diffusion with compatibility guidance. Recovered streams are combined with observed features, and branch predictions are aggregated using candidate weights.

evidence, so X^{obs} reaches the denoiser through a semantic–structural bottleneck; multiple missing modalities are recovered target-wise.

Proposition 1 (Mean regression). *For a two-component mixture with means μ_1, μ_2, weights $w, 1-w$, and $\|\mu_1 - \mu_2\| \geq 2\delta$, the squared-error estimator $\bar{X} = w\mu_1 + (1-w)\mu_2$ satisfies*

$$\min_{j \in \{1,2\}} \|\bar{X} - \mu_j\| \geq 2\delta \min(w, 1-w). \tag{2}$$

The bound is largest for ambiguous mixtures and applies only to squared-error point estimation; it does not imply a limitation of expressive stochastic density models. It motivates preserving alternatives rather than collapsing them before prediction.

Proposition 2 (Cost of hard commitment). *Let $\ell(d, y) \in [0, B]$, $\ell_j(d) = \mathbb{E}_{y \sim p_j}[\ell(d, y)]$, $j^\star = \arg\max_j w_j$, $d_{\text{hard}} \in \arg\min_d \ell_{j^\star}(d)$, and $d_{\text{mix}} \in \arg\min_d \sum_j w_j \ell_j(d)$. Then*

$$\sum_j w_j \ell_j(d_{\text{hard}}) - \sum_j w_j \ell_j(d_{\text{mix}}) \leq B(1 - w_{j^\star}). \tag{3}$$

This is a worst-case bound, not a guarantee of mixture superiority; it motivates retaining alternatives when candidate weights remain diffuse. Proofs are in Appendix Sec. A.

Design requirements. The formulation motivates **(R1)** retaining weighted interpretations rather than committing early, **(R2)** grounding each interpretation in target-specific structure, and **(R3)** encouraging a recovered branch to remain compatible with its conditioning candidate. For R3, the ideal criterion is

$$z = \arg\max_{z' \in \mathcal{C}} p\big(z' \mid X^{\text{obs}} \cup \hat{X}^{\text{miss}}\big). \tag{4}$$

We use Eq. 4 only as an ideal criterion; exhaustive rescoring is impractical, so Sec. 3.4 uses a differentiable single-candidate surrogate.

3.2 Candidate Set and Weights

Candidate generation. For sentiment, z_s contains dominant affect, intensity, a salient event, and modality-specific evidence; other tasks change only the schema (e.g., genre cues on MM-IMDb Arevalo et al. (2017)). At refinement step ℓ, frozen Qwen2.5-Omni-7B Xu et al. (2025) extracts

evidence from available raw streams and proposes a structured candidate,

$$\mathcal{E}^{(\ell)} = \mathcal{F}_{\text{EXT}}(R^{\text{obs}} \mid \text{fb}^{(\ell-1)}),$$
$$z_s^{(\ell)} \sim p_{\psi_r}(\cdot \mid \mathcal{E}^{(\ell)}; P_{\text{struct}}), \tag{5}$$

with $\text{fb}^{(0)} = \emptyset$. Only this frozen reasoner consumes raw streams; trainable recovery components operate on extracted features.

Compatibility and weights. A cross-attention scorer Vaswani et al. (2017) assigns each candidate an energy $E_{\text{abd}}^{(\ell)} = E_\omega(\text{Enc}_{\text{hyp}}(z_s^{(\ell)}), \{\text{Enc}_m(r_m^{(\ell)})\}_{m\in\mathcal{M}_{\text{obs}}})$, where lower energy indicates stronger candidate–evidence agreement. Weakly supported evidence is returned as feedback for the next refinement Madaan et al. (2023); after schema deduplication, the resulting scores become validation-temperature-scaled candidate weights

$$w_j = \frac{\exp(-E_{\text{abd}}^{(j)}/T)}{\sum_{k\in\mathcal{C}} \exp(-E_{\text{abd}}^{(k)}/T)}, \qquad u = H(w)/\log J \quad (J > 1). \tag{6}$$

We denote the ambiguity score by $u \equiv \nu$ below. A valid singleton uses $u = 0$, while the all-failure null fallback uses $u = 1$ (Appendix B.1). The weights are operational over the reasoner-generated set, not a human interpretation posterior; calibration against full-observation pseudo-references therefore inherits reasoner error. Because candidate identities may change across rounds, the previous weight vector is identity-aligned and zero-padded before computing stability $1 - \text{TV}(w^{(\ell)}, \tilde{w}^{(\ell-1)})$. Refinement stops at τ_{abd} or $L_{\max}$, then prunes low-weight candidates (Appendix B.1).

Anchor projection. Each retained candidate is compressed to a fixed-width semantic anchor $\bar{z}_s^{(j)} = g^{(j)} \odot W_1 \text{Enc}_{\text{hyp}}(z_s^{(j)}) + (1 - g^{(j)}) \odot b$, with gate $g^{(j)} = \sigma(W_g \text{Enc}_{\text{hyp}}(z_s^{(j)}))$. This bottleneck keeps a compact candidate-specific signal for downstream retrieval and generation rather than passing the full reasoner output directly. Prompt, verifier, and matched-anchor details are in Appendix Sec. B.1.

3.3 Candidate- and Target-Specific Grounding

For target modality m, fully observed training samples define a memory $\mathcal{B}_m = \{(k_i, v_{i,m})\}_{i=1}^M$ Sukhbaatar et al. (2015); Khandelwal et al. (2020). Each candidate anchor is projected to a retrieval query, while ambiguity controls how broadly the memory is searched:

$$q^{(j)} = \Phi_{\text{proj}}(\bar{z}_s^{(j)}), \qquad s_i^{(j)} = \left\langle \frac{q^{(j)}}{\|q^{(j)}\|_2}, \frac{k_i}{\|k_i\|_2} \right\rangle / \sigma_{\text{eff}},$$
$$\sigma_{\text{eff}} = \sigma_0 + u\Delta\sigma, \qquad K_{\text{eff}} = K_{\min} + \lfloor u(K_{\max} - K_{\min}) \rfloor. \tag{7}$$

Thus low ambiguity keeps retrieval local, whereas diffuse candidate weights broaden the neighborhood and soften similarity. The selected values are then aggregated and passed through a coverage gate,

$$\alpha_i = \frac{\exp s_i^{(j)}}{\sum_{i'\in\mathcal{N}_{K_{\text{eff}}}} \exp s_{i'}^{(j)}}, \qquad \tilde{z}_p^{(j)} = \sum_{i\in\mathcal{N}_{K_{\text{eff}}}} \alpha_i v_i,$$
$$\gamma^{(j)} = \sigma\Big(W_\gamma[\bar{z}_s^{(j)}; \tilde{z}_p^{(j)}; \max_i s_i^{(j)}]\Big), \qquad z_p^{(j)} = \gamma^{(j)}\tilde{z}_p^{(j)} + (1 - \gamma^{(j)})z_p^0. \tag{8}$$

The gate therefore separates two failure regimes: well-supported queries use instance-specific retrieved structure, while weak coverage falls back toward a learned parametric prior. Retrieval is candidate- and target-specific, so two interpretations of the same incomplete sample may ground themselves in different prototypes, and different missing modalities use different memories. This structure encourages branch separation without assuming that retrieval alone makes the branches distinct. Memory construction, retrieval pre-training, and alignment are detailed in Appendix Sec. B.2.

3.4 Candidate-Consistent Generation

For missing target m, a sequence-aware 1D U-Net Ronneberger et al. (2015) synthesizes X_m from $\bar{z}_s^{(j)}$ and target-specific $z_{p,m}^{(j)}$. With the standard diffusion parameterization Ho et al. (2020); Song et al. (2021), training uses

$$\mathcal{L}_{\text{diff}} = \sum_{j=1}^{J} \text{sg}(w_j)\, \mathbb{E}_{z_0,\epsilon,t}\Big[\|\epsilon - \epsilon_\theta(z_t, t, \bar{z}_s^{(j)}, z_p^{(j)})\|_2^2\Big]. \tag{9}$$

All candidates share the same factual denoising target, so the objective does not manufacture diversity by assigning different ground-truth features to different branches. Instead, branches differ through their candidate-specific semantic anchor and retrieved structural prior. The former enters through cross-attention and provides sequence-level semantic context, whereas the latter modulates GroupNorm through conditional affine parameters Wu & He (2018); Perez et al. (2018). This split is intended to keep semantic interpretation and instance-level structure distinguishable during conditioning, but it is a modeling choice evaluated by ablation rather than a claim that the two pathways contain disjoint information or guarantee branch diversity.

Compatibility guidance. At selected late steps, HyDiff converts the current noisy state to a clean estimate, scores that estimate with the frozen verifier, and nudges the denoising direction toward lower candidate energy:

$$\hat{z}_0^{(t)} = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(z_t, t, \bar{z}_s^{(j)}, z_p^{(j)})}{\sqrt{\bar{\alpha}_t}}, \qquad \tilde{\epsilon}_\theta = \epsilon_\theta + \eta_{\text{guide}}(t)\sqrt{1-\bar{\alpha}_t}\,\nabla_{z_t} E_{\text{cyc}}^{(t,j)}, \tag{10}$$

where $E_{\text{cyc}}^{(t,j)} = E_\omega(\bar{z}_s^{(j)}, \{\Phi_{\text{cross}}(\hat{z}_0^{(t)}, e_m)\}_{m\in\mathcal{M}_{\text{obs}}})$. This is analogous to classifier guidance Dhariwal & Nichol (2021), but the guidance target is candidate compatibility rather than a class label. The bridge Φ_{cross} is fitted on real completions so that the verifier receives inputs in its training space. We apply guidance only late in sampling, when the clean estimate is sufficiently stable; because the verifier scores only the current branch, the update improves local compatibility without establishing the global fixed-point condition in Eq. 4. Scheduling and gradient details are in Appendix Sec. B.3.

Reference-relative hinge. To limit direct pressure to exploit the frozen verifier, training uses

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{t,j}\left[(E_{\text{cyc}}^{(t,j)} - E_{\text{real}}^{(j)} + m)_+\right]. \tag{11}$$

Here $E_{\text{real}}^{(j)}$ is the real-completion energy under the same candidate. The implemented sign convention makes the hinge zero only beyond the reference-relative margin; it limits direct score minimization but does not by itself rule out verifier exploitation. Appendix F.1 therefore evaluates generated energy jointly with feature fidelity and task performance.

3.5 Prediction and Training

Each retained candidate yields target-wise recovery $\hat{X}^{\text{miss,(j)}}$, which is fused with unchanged observed features and passed to task head f:

$$\hat{y} = \sum_j w_j\, f\big(\text{Fuse}(X^{\text{obs}}, \hat{X}^{\text{miss,(j)}})\big). \tag{12}$$

This averages predictions rather than recovered features, avoiding the point-estimation collapse motivating Prop. 1. It is a practical decision-level mixture rather than a universal optimality claim, and u is returned for selective prediction. With the reasoner, verifier, calibration, retrieval keys, and candidate weights frozen, trainable recovery and prediction components optimize

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}}(\hat{y}, y) + \lambda_1 \mathcal{L}_{\text{diff}} + \lambda_2 \mathcal{L}_{\text{cyc}}. \tag{13}$$

Training is staged. The verifier, cross-modal bridge, and retrieval space are first fitted on fully observed samples and then frozen; the generator, projection heads, gates, and task head are optimized afterward. Candidate weights act as stop-gradient sample-wise coefficients, so each branch receives task pressure in proportion to its current compatibility without allowing the task loss to rewrite the candidate distribution itself. Freezing the verifier likewise keeps the guidance energy fixed across generation and evaluation, while the reference-relative hinge limits direct pressure to exploit that frozen score. This separation is important because the same compatibility signal is reused during candidate refinement, late diffusion guidance, and final prediction. At inference, no additional optimization is performed: the frozen reasoner proposes and weights candidates, each candidate retrieves target-specific structure, diffusion recovers the missing stream(s), and the shared head produces branch predictions that are mixed by the retained weights. The same weights also yield the ambiguity score used for selective prediction. Because retrieval and diffusion are executed per retained candidate–target pair, inference cost grows with the retained set and the number of missing targets; pruning and target-specific fallback keep this branching bounded in practice, with measured latency reported in Appendix G.

4 EXPERIMENTS

Setup. We evaluate CMU-MOSI Zadeh et al. (2016), CMU-MOSEI Zadeh et al. (2018), and CH-SIMS v2.0 Liu et al. (2022) with Acc-2/F1, MAE, correlation, and finer-grained accuracies Tsai et al. (2019). **Intra-RA** masks feature coordinates with probability η while allowing the frozen reasoner to inspect available raw streams; **Mod** removes complete modalities. We report the realized-distribution average (Mod-DW) and equal-weight pattern macro-average (Mod-PM), with exact probabilities in Appendix D. BERT, FACET, and COVAREP features are projected to 128 dimensions. SRAR uses frozen Qwen2.5-Omni-7B Xu et al. (2025) with deterministic decoding and at most three refinement rounds; validation-calibrated weights use $T = 1.35$, candidates below 0.08 are pruned, and $J = 2.7$ candidates remain on average. Retrieval uses base $K = 10$ and $\sigma_0 = 0.1$ with ambiguity-adaptive breadth; diffusion uses 1000 training steps and 50-step DDIM Song et al. (2021). The completion objective uses $\lambda_1 = 0.1$, $\lambda_2 = 0.05$, cycle margin $m = 0.12$, and late guidance strength $\eta_0 = 0.25$. Adam Kingma & Ba (2015) starts at 2×10^{-3}; batch sizes are 32 for MOSI/SIMS and 128 for MOSEI. Half the training samples receive dynamic missingness, and all results use seeds $\{13, 17, 23, 41, 89\}$. Full stopping, masking, retrieval, calibration, and latency protocols are in Appendices B–C.7 and G. Baselines include DMD Li et al. (2023), T^{2}DR Lin et al. (2025), LNLN Zhang et al. (2024), CorrKD Li et al. (2024), MPLMM Guo et al. (2024), P-RMF Zhu et al. (2025), DiCMoR Wang et al. (2023a), SMCMSA Sun et al. (2024), IMDer Wang et al. (2023b), MMIN Zhao et al. (2021), and retrieval-assisted MissRAG Pipoli et al. (2025).

Protocol notes. Intra-RA is intentionally reported as a raw-assisted diagnostic rather than a matched-input comparison: the reasoner may inspect the available raw streams while the task backbone receives corrupted extracted features. We therefore include HyDiff-Feat, which disables raw access and receives exactly the same masked BERT/FACET/COVAREP tensors and availability masks as P-RMF. For Mod, the all-missing draw is rejected and redrawn; Mod-DW follows the resulting conditional pattern distribution, whereas Mod-PM gives each non-empty pattern equal weight. The ○○○ pattern requires no recovery, so it is reported for completeness but excluded from recovery-specific interpretation. Candidate-weight ECE uses the fixed full-observation pseudo-reference subset and is treated only as an internal calibration diagnostic, not as calibration to a human interpretation posterior.

4.1 COMPARISON WITH THE STATE OF THE ART

Model	CMU-MOSI							CMU-MOSEI							CH-SIMS						
	oxx	xox	xxo	oox	oxo	xoo	ooo	oxx	xox	xxo	oox	oxo	xoo	ooo	oxx	xox	xxo	oox	oxo	xoo	ooo
DMD	78.52	49.35	51.42	79.85	80.12	53.60	81.45	81.50	56.40	58.62	83.15	82.55	60.12	84.62	75.32	52.15	54.20	76.15	76.82	55.45	78.35
MPLMM	80.15	52.85	53.20	81.22	81.45	55.12	82.40	82.45	59.32	60.15	83.92	83.12	62.35	84.75	76.45	55.30	56.12	77.35	77.15	57.82	79.42
CorrKD	79.15	55.05	56.12	81.10	81.05	57.52	81.42	82.65	62.62	62.75	83.25	83.52	63.75	84.25	76.12	58.12	59.25	77.18	77.35	60.25	78.45
LNLN	85.65	62.10	63.15	85.82	85.45	64.02	86.25	86.05	65.50	65.12	86.22	86.15	65.50	86.40	80.12	63.22	64.10	80.85	81.55	65.15	81.85
T^{2}DR	85.20	62.59	62.90	85.30	85.10	64.55	86.05	85.60	65.85	65.14	85.70	85.72	66.02	85.75	79.52	62.85	63.52	80.45	81.57	64.25	81.35
MMIN	83.82	57.02	55.35	83.85	84.05	60.45	84.62	82.35	59.35	58.92	83.42	83.75	63.52	84.35	77.25	59.45	58.12	78.52	78.42	61.35	79.15
DiCMoR	84.55	62.25	60.52	83.52	85.52	64.05	85.75	84.22	63.62	62.95	84.95	85.05	65.25	85.15	78.45	61.42	60.25	79.82	79.15	63.12	80.12
IMDer	84.82	61.32	62.05	85.55	85.45	63.62	85.78	84.55	63.82	63.82	85.02	85.12	63.55	85.15	79.12	62.25	61.35	80.45	80.15	64.55	80.52
SMCMSA	81.42	60.25	60.25	83.15	81.42	63.35	85.82	80.35	64.25	63.32	75.92	76.55	65.55	80.65	77.15	60.12	59.45	76.35	78.15	62.25	79.12
P-RMF	86.72	62.55	62.08	85.75	85.75	61.15	86.02	84.85	62.95	62.95	85.12	85.35	64.85	85.15	80.25	63.55	63.22	81.12	81.55	65.12	81.95
MissRAG	85.94	63.91	64.03	86.44	85.27	66.05	87.85	86.37	68.00	67.37	84.85	85.02	67.62	86.30	81.07	65.26	67.06	81.44	81.93	66.90	83.28
Ours	**87.85**	**65.12**	**64.55**	**87.15**	**87.25**	**67.45**	**89.62**	**87.25**	**68.75**	**68.95**	**86.85**	**86.62**	**69.55**	**87.82**	**82.45**	**66.82**	**67.55**	**83.15**	**83.42**	**68.12**	**84.55**

Table 1: Acc-2 (%) across the seven non-empty [L,V,A] availability patterns at pre-rejection $\mu = 0.7$; 'o' and 'x' denote observed and missing modalities. Bold/underline mark the best/second-best reported means.

Pattern-level results. HyDiff achieves the highest reported mean in all 21 availability-pattern columns. Relative to MissRAG, the Mod-DW gains are 1.28, 1.31, and 1.24 Acc-2 points on MOSI, MOSEI, and SIMS, with corresponding Mod-PM gains of 1.36, 1.47, and 1.30. The advantage is therefore not driven by a small subset of missingness patterns.

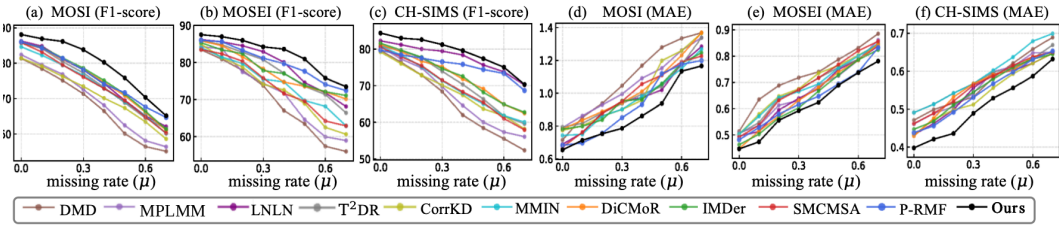

Figure 3: Mod-DW F1 (a–c) and MAE (d–f) as modality-level missingness increases on CMU-MOSI, CMU-MOSEI, and CH-SIMS.

(a) Fairness / semantic controls

Matched-input Acc-2	MOSI	MOSEI	SIMS
P-RMF (matched)	72.35	78.28	73.41
HyDiff-Feat	**73.24**	**79.02**	**74.05**

Anchor-matched mean	P-RMF	HyDiff
Full anchor	76.30	**77.04**
Reduced anchor	74.91	**75.53**

(b) Candidate controls

Mod-PM control	Avg. Acc-2
Single sample	74.57
Three samples	76.16
Repeat top-1 ($J = 3$)	77.06
HyDiff	**78.14**

Pseudo-ref.	Top-1	Set
MOSI	71.8	86.4
MOSEI	76.6	90.3
SIMS	68.4	84.9
Avg.	72.3	87.2

Table 2: Main controls. (a) Matched-input results isolate raw-stream access; anchor-matched means compare P-RMF and HyDiff under the full schema and a shared reduced schema that removes `polarity`, `intensity`, and `dominant_affect` (Appendix C.6). (b) Candidate controls separate generic sampling, repeated top-1 conditioning, and distinct candidate branches; pseudo-reference coverage is an internal diagnostic (Appendix E.2).

Main controls. Under identical masked features, HyDiff-Feat improves over P-RMF by 0.64–0.89 points. Matching the full semantic anchor narrows the six-setting gap to 0.74 points. More stringently, when both methods use the same reduced-anchor protocol without `polarity`, `intensity`, or `dominant_affect`, HyDiff is higher in all six settings by 0.45–0.79 points (0.62 on average), and in 26/30 paired seed runs. Thus the semantic prior explains a material portion of the headline gain, while a smaller architecture-specific advantage remains. Candidate controls further decompose the average gain over single-sample diffusion into +1.59 points from three-sample ensembling, +0.90 from repeated top-1 conditioning, and +1.08 from distinct weighted branches. Retaining the set raises pseudo-reference coverage from 72.3% (top-1) to 87.2%, although 9.7–15.1% remain uncovered. MD^{2}N Dai et al. (2025) is discussed in Related Work but is not included without a protocol-matched rerun.

Interpretation. The controls separate three common alternative explanations. First, the matched-input comparison shows that raw-stream access does not account for the entire gain. Second, the full- and reduced-anchor comparisons show that task-aligned semantics are a substantial contributor and should not be conflated with the recovery architecture; the remaining matched reduced-anchor gain is consistent but modest. Third, the three-trajectory controls show that generic sampling and strong top-1 conditioning explain much of the improvement over a single diffusion sample: distinct candidate-conditioned branches add a further 1.08 points over repeated top-1 conditioning rather than accounting for the full headline gain. The coverage diagnostic is consistent with this mechanism but remains self-referential to the frozen reasoner, so we treat it as internal consistency evidence rather than external validation of human semantic modes.

Additional evidence. On MM-IMDb, HyDiff improves over matched P-RMF by +3.43 Micro-F1 and +3.77 Macro-F1; because the schema uses task-aligned genre candidates, this supports architectural portability under task-specific semantic conditioning, not label-agnostic generalization (Appendix G.1). The measured $J = 1$ latency is 689 ± 15 ms versus 431 ± 9 for P-RMF and 742 ± 17 for DiCMoR; the default serial $J = 2.7$ estimate is about 1.12 s (Appendix G.3). Together with the +1.08-point gain over repeated top-1 conditioning, this makes the additional candidate machinery a modest quality–compute trade-off rather than an efficiency claim. The reference-relative hinge also improves completion fidelity: versus raw energy minimization it lowers NRMSE by 0.200–0.215 and raises M-PM Acc-2 by 2.24–4.30 points across the three datasets (Appendix F.1). **Statistical reliability.** Across the nine primary dataset–protocol comparisons, paired randomization differences remain significant after Holm correction ($p_{\text{Holm}} \leq 0.048$). For the stricter reduced-anchor family, all six raw paired-randomization tests have $p < 0.05$; applying standard Holm correction across these six comparisons leaves MOSEI I-RA significant ($p_{\text{Holm}} = 0.036$), while the other five adjust to 0.055. The mean gain nevertheless remains positive in every setting and in 26/30 paired seed runs; full seed variation and test details are in Appendix F.2.

4.2 Ablation Study

Variant	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	I-RA	M-PM	I-RA	M-PM	I-RA	M-PM
Base, direct completion	61.23	66.17	73.14	72.81	64.86	67.92
R1: candidate set and weighting						
Single anchor, $J = 1$	70.18	71.94	78.12	76.63	73.27	72.65
Argmax instead of mixture	69.60	74.33	75.09	75.93	70.95	73.17
Uncalibrated weights	68.63	75.25	75.46	75.25	70.37	73.82
R2: grounding						
No retrieval	68.49	75.46	76.88	78.57	71.53	75.19
No coverage gate	70.50	74.84	74.57	76.55	70.29	73.54
No ambiguity adaptation	71.95	76.47	78.11	77.82	72.86	74.93
Class-mean prior	67.83	74.44	75.64	77.11	70.99	74.47
R3: consistency						
Single injection path	72.31	76.68	79.04	78.72	73.66	75.74
No cycle training loss	69.72	74.91	75.22	75.34	71.50	74.48
No inference guidance	70.41	74.12	76.18	76.79	71.25	74.40
Raw energy, no hinge	69.37	74.13	75.18	75.48	71.27	74.34
HyDiff	**73.68**	**78.43**	**79.55**	**79.40**	**74.62**	**76.58**

Table 3: Requirement ablations (Acc-2, %). I-RA uses feature corruption at $\eta = 0.9$ and M-PM is the pattern-macro score at $\mu = 0.7$.

All R1–R3 interventions reduce performance (Table 3); full HyDiff exceeds direct completion by 9.36 points on average across the six settings. Under R1, collapsing to one anchor drops the mean by 3.25 points, while argmax prediction and uncalibrated weights drop it by 3.87 and 3.91 points, supporting both candidate retention and calibrated mixing. Under R2, removing the coverage gate or replacing instance-specific prototypes with a class mean costs 3.66 and 3.63 points; disabling ambiguity adaptation is milder at 1.69 points. Under R3, removing the hinge, cycle loss, or inference guidance costs 3.75, 3.52, and 3.18 points, while the single-injection variant is only 1.02 points lower, indicating that the dual conditioning path is helpful but not the dominant source of gain. Appendix E reports further sensitivity, completion, and verifier diagnostics.

5 Conclusion

HyDiff represents incomplete-modality uncertainty with a weighted set of structured candidates, grounds each candidate–target pair by retrieval, and performs candidate-conditioned recovery before decision aggregation. Across three sentiment benchmarks it improves over strong baselines under diverse missingness patterns and matched-input controls. The matched reduced-anchor experiment further clarifies the source of the gain: task-aligned semantic assistance contributes substantially, while candidate-indexed recovery retains a smaller but consistent residual advantage when the three most target-facing fields are removed. Candidate controls and latency measurements likewise show that distinct branches add quality beyond repeated top-1 conditioning, but at nontrivial inference cost; HyDiff is therefore a quality–compute trade-off rather than an efficiency improvement. MM-IMDb supports architectural portability only when task-aligned semantic conditioning is available, so generalization without such a schema remains open. Finally, the frozen reasoner's pretraining data are not fully auditable, and useful interpretations may still be omitted upstream. Future work should strengthen external validation of candidate ambiguity, improve candidate recall, and share computation more efficiently across branches.

AI Use Statement

Generative AI tools were used to assist with literature retrieval and discovery, including identifying potentially relevant prior work. All references suggested by AI tools were independently verified by the authors against the original sources. The authors take full responsibility for the accuracy and appropriateness of all cited references.

Reproducibility Statement

An anonymized code package is provided with the submission through the conference submission system to support reproducibility of the reported results.

References

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *International Conference on Learning Representations Workshop*, 2017.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Christopher M. Bishop. Mixture density networks. Technical Report NCRG/94/004, Aston University, 1994.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24507–24517, 2025.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. COVAREP: A collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 960–964, 2014. doi: 10.1109/ICASSP.2014.6853739.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zhi-Hao Guan, Longfei Huang, and Yang Yang. Anchor-guided gradient alignment for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37896–37905, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1726–1736, 2024. doi: 10.18653/v1/2024.acl-long.94.

Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. In *Advances in Neural Information Processing Systems*, volume 25, pp. 1808–1816, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kang He, Yuzhe Ding, Rao Fu, Yukang Feng, Kaipeng Zhang, Yiming Liu, Fei Li, Chong Teng, and Donghong Ji. Uncertainty-calibrated elastic alignment for multimodal sentiment analysis with missing modalities. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5268–5288, 2026a. doi: 10.18653/v1/2026.findings-acl.260.

Ningxin He, Yongheng Deng, Sheng Yue, Yongjian Fu, Zehui Zhang, and Tiegang Gao. RAG4DMC: Retrieval-augmented generation for data-level modality completion. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25864–25873, 2025. doi: 10.1109/CVPR52734.2025.02409.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18035–18043, 2025. doi: 10.1609/aaai.v39i17.33984.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12458–12468, 2024.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6631–6640, 2023.

Han Lin, Xiu Tang, Huan Li, Wenxue Cao, Sai Wu, Chang Yao, Lidan Shou, and Gang Chen. T^2DR: A two-tier deficiency-resistant framework for incomplete multimodal learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8602–8616, 2025. doi: 10.18653/v1/2025.findings-acl.452.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 247–258, 2022. doi: 10.1145/3536221.3556630.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021. doi: 10.1609/aaai.v35i3.16330.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xiangyu Meng and Shicai Wei. General incomplete multimodal learning via dynamic quality perception. In *Computer Vision – ECCV 2026*, pp. 76–95, 2026. doi: 10.1007/978-3-032-37595-7_5.

Phuong-Anh Nguyen, Tien Anh Pham, Duc-Trong Le, and Cam-Van Thi Nguyen. BALM: A model-agnostic framework for balanced multimodal learning under imbalanced missing rates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30246–30256, 2026.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671.

Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. MissRAG: Addressing the missing modality challenge in multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3215–3224, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3591–3600, 2017. doi: 10.1109/ICCV.2017.388.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Sabrina Stöckli, Michael Schulte-Mecklenbeck, Stefan Borer, and Andrea C. Samson. Facial expression analysis with AFFDEX and FACET: A validation study. *Behavior Research Methods*, 50:1446–1460, 2018. doi: 10.3758/s13428-017-0996-1.

Sainbayar Sukhbaatar, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Yuhang Sun, Zhizhong Liu, Quan Z. Sheng, Dianhui Chu, Jian Yu, and Hongxiang Sun. Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 110:102454, 2024. doi: 10.1016/j.inffus.2024.102454.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019. doi: 10.18653/v1/P19-1656.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22025–22034, 2023a.

Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

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. *arXiv preprint arXiv:2503.20215*, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024. doi: 10.1093/nsr/nwae403.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. doi: 10.1109/MIS.2016.94.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018. doi: 10.18653/v1/P18-1208.

Zhilin Zeng, Zelin Peng, Xiaokang Yang, and Wei Shen. Missing as masking: Arbitrary cross-modal feature reconstruction for incomplete multimodal brain tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15008 of *Lecture Notes in Computer Science*, pp. 424–433, 2024. doi: 10.1007/978-3-031-72111-3_40.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Hengrui Zhang, Liancheng Fang, Qitian Wu, and Philip S. Yu. DiffPuter: Empowering diffusion models for missing data imputation. In *International Conference on Learning Representations*, 2025.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2608–2618, 2021. doi: 10.18653/v1/2021.acl-long.203.

Shu Zhao, Nilesh Ahuja, Tan Yu, Tianyi Shen, and Vijaykrishnan Narayanan. MoRA: Missing modality low-rank adaptation for visual recognition. In *The Fourteenth International Conference on Learning Representations*, 2026.

Han Zhou, Jordy Van Landeghem, Teodora Popordanoska, and Matthew B. Blaschko. A novel characterization of the population area under the risk coverage curve (AURC) and rates of finite sample estimators. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79226–79253. PMLR, 2025.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22123–22138, 2025. doi: 10.18653/v1/2025.acl-long.1075.

Shihao Zou, Wei Wei, and Yongshuo Zhang. DEAR: Distributional error-aware reliability for robust multimodal sentiment analysis with missing modalities. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30344–30361, 2026. doi: 10.18653/v1/2026.findings-acl.1517.

A Theoretical Analysis

This appendix proves the claims stated in Sec. 3.1, adds a recovery-error decomposition that separates the three design requirements and links them to empirical diagnostics, and closes with the relation between the conditional likelihood and the denoising surrogate actually optimized.

A.1 Notation

Let X^{obs} denote the observed modalities and X^{miss} the missing ones. We write $\mathcal{Z}_s$ for the countable candidate-interpretation space, $p(z \mid X^{\text{obs}})$ for the corresponding conditional weight, and $p(X^{\text{miss}} \mid X^{\text{obs}}, z)$ for the interpretation-indexed conditional completion distribution. We write $\mathcal{C} \subset \mathcal{Z}_s$ for the finite candidate set retained by SRAR for a given input. When analyzing a fixed X^{obs}, we suppress this conditioning argument and write $p(\cdot \mid z)$ for readability. Estimated quantities carry hats, so $\hat{w}$ is the candidate-set weighting produced by SRAR and $\hat{p}(\cdot \mid z)$ is the distribution realized by retrieval-injected diffusion. Total variation between distributions P and Q is $\mathrm{TV}(P, Q) = \frac{1}{2}\|P - Q\|_1$.

A.2 Interpretation-Indexed Mixture Identity

Lemma A.1 (Conditional mixture identity)**.** *For any discrete interpretation variable $z \in \mathcal{Z}_s$,*

$$p(X^{\text{miss}} \mid X^{\text{obs}}) = \sum_{z \in \mathcal{Z}_s} p(z \mid X^{\text{obs}}) p(X^{\text{miss}} \mid X^{\text{obs}}, z). \tag{A.1}$$

Proof. Marginalizing the joint distribution over z and applying the chain rule gives

$$\begin{aligned} p(X^{\text{miss}} \mid X^{\text{obs}}) &= \sum_{z} p(X^{\text{miss}}, z \mid X^{\text{obs}}), \\ &= \sum_{z} p(z \mid X^{\text{obs}}) p(X^{\text{miss}} \mid X^{\text{obs}}, z). \end{aligned} \tag{A.2}$$

No conditional-independence assumption is required. □

Equation A.1 is an exact probabilistic identity. Approximation enters when HyDiff replaces the full conditional $p(X^{\text{miss}} \mid X^{\text{obs}}, z)$ with the implemented semantic–structural bottleneck. For a missing target modality m, the implemented generator is $p_\theta(X_m \mid \bar{z}_s, z_{p,m})$, where $z_{p,m}$ is retrieved from the target-specific memory $\mathcal{B}_m$ using a query derived from the observed-evidence anchor. When several modalities are missing, this conditional is invoked target-wise and the recovered streams are collected as $\hat{X}^{\text{miss}}$; this operational construction is not a claim that missing modalities are conditionally independent. HyDiff also replaces the full interpretation space with a finite reasoner-generated candidate set. The experiments therefore evaluate the usefulness of this approximation rather than claiming that the structured schema is a sufficient statistic for the missing modality.

A.3 Proof of Proposition 1

Proposition A.1 (Restatement of Proposition 1)**.** *Let the mixture in Eq. A.1 have two components with means μ_1, μ_2, weights w and $1 - w$, and separation $\|\mu_1 - \mu_2\| \geq 2\delta$. Any estimator minimizing the expected squared recovery error returns $\bar{X} = w\mu_1 + (1 - w)\mu_2$ and satisfies*

$$\min_{j \in \{1,2\}} \|\bar{X} - \mu_j\| \;\geq\; 2\delta \min(w, 1 - w). \tag{A.3}$$

Proof. For a fixed X^{obs}, the minimizer of $\mathbb{E}\big[\|\hat{X} - X^{\text{miss}}\|_2^2\big]$ over point estimates is the conditional mean, so $\hat{X} = \mathbb{E}[X^{\text{miss}} \mid X^{\text{obs}}] = w\mu_1 + (1 - w)\mu_2 = \bar{X}$. Substituting,

$$\begin{aligned} \|\bar{X} - \mu_1\| &= \|w\mu_1 + (1 - w)\mu_2 - \mu_1\| \\ &= (1 - w)\|\mu_2 - \mu_1\| \;\geq\; 2\delta(1 - w), \end{aligned} \tag{A.4}$$

and symmetrically $\|\bar{X} - \mu_2\| = w\|\mu_1 - \mu_2\| \geq 2\delta w$. Taking the minimum of the two bounds gives Eq. A.3. □

Remark A.1 (Behavior of the bound)**.** *The right side is maximized at $w = \frac{1}{2}$, where the barycenter sits at distance at least δ from both components. It vanishes as $w \to 0$ or $w \to 1$, which is the regime in which the posterior concentrates on one reading and a point estimate can be adequate. The result is a conditional geometric statement about a two-component mixture. It does not by itself predict the relative accuracy margins of different missingness patterns in Table 1, which also depend on posterior estimation, grounding, generation, and task prediction.*

Remark A.2 (More than two components)**.** *Eq. A.3 does not extend verbatim to $J > 2$, because the barycenter can coincide with a component mean. Taking collinear means $\mu_1 = -e$, $\mu_2 = 0$, $\mu_3 = e$ with equal weights gives $\bar{X} = 0 = \mu_2$ and a vanishing left side. The statement holds for general J under the additional condition $\bar{X} \notin \{\mu_j\}_{j=1}^{J}$, in which case $\min_j \|\bar{X} - \mu_j\| > 0$ and the displacement is again governed by how far the weights are from a vertex of the simplex. We state the two-component case because it is provable without side conditions and because the candidate-weight distributions we observe concentrate on two or three candidates.*

A.4 Proof of Proposition 2

Proposition A.2 (Restatement of Proposition 2)**.** *Let ℓ be a task loss bounded in $[0, B]$, let $\ell_j(d) = \mathbb{E}_{y\sim p_j}[\ell(d, y)]$ be the risk of decision d under component j, and let $w_{j^\star} = \max_j w_j$. Let d_{mix} minimize $\sum_j w_j \ell_j(d)$ and let d_{hard} minimize $\ell_{j^\star}(d)$. Then*

$$\sum_j w_j \ell_j(d_{\mathrm{hard}}) - \sum_j w_j \ell_j(d_{\mathrm{mix}}) \;\le\; B\big(1 - w_{j^\star}\big). \tag{A.5}$$

Proof. Split the risk of the committed decision by component,

$$\begin{aligned} \sum_j w_j \ell_j(d_{\mathrm{hard}}) &= w_{j^\star} \ell_{j^\star}(d_{\mathrm{hard}}) + \sum_{j \neq j^\star} w_j \ell_j(d_{\mathrm{hard}}) \\ &\le w_{j^\star} \ell_{j^\star}(d_{\mathrm{mix}}) + B\big(1 - w_{j^\star}\big), \end{aligned} \tag{A.6}$$

where the first term uses optimality of d_{hard} for component $j^\star$ and the second uses $\ell_j \le B$ together with $\sum_{j\neq j^\star} w_j = 1 - w_{j^\star}$. Nonnegativity of the loss gives the matching lower bound

$$\sum_j w_j \ell_j(d_{\mathrm{mix}}) \;\ge\; w_{j^\star} \ell_{j^\star}(d_{\mathrm{mix}}). \tag{A.7}$$

Subtracting Eq. A.7 from Eq. A.6 yields Eq. A.5. □

Remark A.3 (Tightness and consequences)**.** *The bound is attained when the non-dominant components assign maximal loss to d_{hard}, which can occur when competing readings imply opposite labels. It depends directly on the residual candidate weight $1 - w_{j^\star}$ and vanishes when the candidate weights are concentrated on one component. The proposition therefore motivates retaining multiple components when the candidate weights are not concentrated, but it does not prove that the normalized entropy u is equivalent to $1 - w_{j^\star}$ or that the arithmetic aggregation in Eq. 12 is Bayes-optimal for every task loss. We use u as a smooth summary of the full candidate-weight vector and validate its usefulness for selective prediction empirically; $1 - \max_j w_j$ is a compatible alternative concentration measure.*

A.5 Recovery Error Decomposition

The three design requirements are intended to address conceptually distinct sources of error. The following decomposition makes that separation explicit under a common-support convention, while the subsequent experiments provide diagnostics rather than direct estimates of the total-variation terms.

Proposition A.3 (Three-term decomposition)**.** *Represent the reference and estimated mixtures on the common interpretation index set $\mathcal{I} = \mathrm{supp}(w) \cup \mathrm{supp}(\hat{w})$, extending absent weights by zero. Assume that the conditional model $\hat{p}(\cdot \mid z_j)$ is defined for every $z_j \in \mathcal{I}$. Let $\hat{p}(X^{\mathrm{miss}} \mid X^{\mathrm{obs}}) = \sum_{j\in\mathcal{I}} \hat{w}_j \, \hat{p}(X^{\mathrm{miss}} \mid z_j)$ be the completion distribution realized by HyDiff, and let p be the reference mixture of Eq. A.1. Let $\bar{p}(\cdot \mid z_j)$ denote the distribution produced by an ideal conditional generator when conditioned on the grounding retrieved by HyDiff for z_j. Thus $\bar{p}$ retains grounding error but removes generator approximation error. Then*

$$\begin{aligned} \mathrm{TV}\big(\hat{p}, p\big) \;\le\; & \underbrace{\mathrm{TV}(\hat{w}, w)}_{\varepsilon_{\mathrm{int}}\ (\mathrm{R1})} + \underbrace{\textstyle\sum_j w_j \, \mathrm{TV}\big(\bar{p}(\cdot \mid z_j), p(\cdot \mid z_j)\big)}_{\varepsilon_{\mathrm{gr}}\ (\mathrm{R2})} \\ & + \underbrace{\textstyle\sum_j w_j \, \mathrm{TV}\big(\hat{p}(\cdot \mid z_j), \bar{p}(\cdot \mid z_j)\big)}_{\varepsilon_{\mathrm{gen}}\ (\mathrm{R3})}. \end{aligned} \tag{A.8}$$

Proof. Insert the intermediate mixture $\sum_j w_j \hat{p}(\cdot \mid z_j)$ and apply the triangle inequality,

$$\begin{aligned} \mathrm{TV}(\hat{p}, p) \le \mathrm{TV}\left(\sum_j \hat{w}_j \hat{p}_j, \sum_j w_j \hat{p}_j\right) \\ + \mathrm{TV}\left(\sum_j w_j \hat{p}_j, \sum_j w_j p_j\right), \end{aligned} \tag{A.9}$$

For the first term, linearity gives

$$\left\| \sum_j (\hat{w}_j - w_j) \hat{p}(\cdot \mid z_j) \right\|_1 \le \sum_j |\hat{w}_j - w_j| = 2\,\mathrm{TV}(\hat{w}, w), \tag{A.10}$$

since each $\hat{p}(\cdot \mid z_j)$ has unit mass, so the term is bounded by $\mathrm{TV}(\hat{w}, w)$ in total-variation units. For the second term, convexity of total variation over a fixed mixing measure gives $\sum_j w_j\, \mathrm{TV}\big(\hat{p}(\cdot \mid z_j), p(\cdot \mid z_j)\big)$, and inserting $\bar{p}(\cdot \mid z_j)$ with one further triangle inequality splits it into the grounding and generation terms. □

Remark A.4 (Empirical diagnostics for the three terms)**.** *The quantities in Eq. A.8 are distributional total-variation terms and are not directly identified by the reported experiments. The top-label ECE in Appendix C.7 probes whether the largest estimated interpretation weight is calibrated with respect to the available interpretation reference, but it neither estimates nor upper-bounds* $\mathrm{TV}(\hat{w}, w)$ *for the full posterior. The coverage-reduction study of Table E.9 and the class-mean control of Table 3 probe sensitivity to the grounding resource associated with* $\varepsilon_{\mathrm{gr}}$*. The NRMSE and cosine metrics of Table F.4 probe feature-level completion fidelity before task fusion, but they are not direct estimates of* $\varepsilon_{\mathrm{gen}}$ *in total variation. These diagnostics are reported separately because no single downstream accuracy value identifies the three terms.*

A.6 Ideal Fixed-Point Criterion and Implemented Surrogate

Eq. 4 states an ideal consistency test: after inserting a generated completion, the branch's conditioning candidate should remain the most compatible member of the retained set $\mathcal{C}$. If every retained candidate were rescored on the completed state, then

$$\arg\max_{z' \in \mathcal{C}} p\big(z' \mid X^{\mathrm{obs}} \cup \hat{X}^{\mathrm{miss}}\big) = \arg\min_{z' \in \mathcal{C}} E_\omega\big(z', \cdot\big) \tag{A.11}$$

under the temperature-scaled softmax of Eq. 6. The implemented guidance does not compute this full comparison at each diffusion check. It evaluates only the current candidate j through $E_{\mathrm{cyc}}^{(s,j)}$ and follows its gradient. This is a lower-cost differentiable surrogate for candidate compatibility, not an exact enforcement of the argmax relation. In particular, lowering $E_{\mathrm{cyc}}^{(s,j)}$ does not establish that another retained candidate $z' \neq z_j$ would receive higher energy on the same completion. The paper therefore uses Eq. 4 as a motivating ideal and interprets the R3 ablations as evidence for the implemented surrogate rather than as verification of exact fixed-point consistency.

A.7 Relation to the Conditional ELBO

This subsection explains why the training objective takes the form it does. It is explanatory rather than load-bearing, and the implemented loss in Eq. 13 does not evaluate every divergence appearing below.

Factorization. HyDiff introduces the semantic anchor z_s and structural priors z_p. For multiple missing targets, z_p is shorthand for the collection of target-specific priors $\{z_{p,m} : m \in \mathcal{M}_{\mathrm{miss}}\}$. Two architectural modeling assumptions are made, namely that each structural prior is selected from the semantic anchor and that generation is conditioned on the corresponding pair $(z_s, z_{p,m})$. These are modeling choices and not an identified causal graph. The conditional joint is approximated as

$$\begin{aligned} p(X^{\mathrm{miss}}, z_s, z_p \mid X^{\mathrm{obs}}) \\ &= p(X^{\mathrm{miss}} \mid z_s, z_p, X^{\mathrm{obs}}) \\ &\quad \cdot p(z_p \mid z_s, X^{\mathrm{obs}})\, p(z_s \mid X^{\mathrm{obs}}) \\ &\approx p_\theta(X^{\mathrm{miss}} \mid z_s, z_p)\, p_\phi(z_p \mid z_s)\, p_\psi(z_s \mid X^{\mathrm{obs}}), \end{aligned} \tag{A.12}$$

where p_ψ is the reasoner-induced distribution over anchors, p_ϕ the retrieval-induced distribution over prototypes, and p_θ the conditional generator. In the implementation, the p_θ term is realized target-wise as $p_\theta(X_m \mid z_s, z_{p,m})$ and the resulting streams are collected; Eq. A.12 is bookkeeping notation rather than an independence assumption across missing targets. Dropping the remaining direct dependence on X^{obs} is the semantic–structural bottleneck used by the implemented denoiser. It is an architectural modeling choice rather than a consequence of a conditional-independence assumption, because observed information reaches generation only after being compressed into the anchor and target-specific retrieved prior.

Variational bound. Introducing a variational posterior $q(z_s, z_p \mid X^{\text{miss}}, X^{\text{obs}})$ and applying Jensen's inequality gives

$$\begin{aligned} &\log p(X^{\text{miss}} \mid X^{\text{obs}}) \\ &\quad = \log \iint p(X^{\text{miss}}, z_s, z_p \mid X^{\text{obs}})\, dz_s\, dz_p \\ &\quad \geq \mathbb{E}_q\Big[\log p(X^{\text{miss}}, z_s, z_p \mid X^{\text{obs}}) - \log q\Big], \end{aligned} \tag{A.13}$$

and substituting Eq. A.12 yields

$$\begin{aligned} \mathcal{L}_{\text{ELBO}} = \mathbb{E}_q\big[\log p_\theta(X^{\text{miss}} \mid z_s, z_p)\big] \\ - D_{\text{KL}}\Big(q \,\Big\|\, p_\psi(z_s \mid X^{\text{obs}})\, p_\phi(z_p \mid z_s)\Big). \end{aligned} \tag{A.14}$$

The first term has the form of a conditional reconstruction objective and is what the denoising loss surrogates. The divergence term would regularize a learned variational posterior, but HyDiff does not evaluate it. Verifier pre-training and contrastive retrieval alignment play analogous roles operationally without being unbiased estimators of it.

Hierarchical split. Adopting $q(z_s, z_p \mid \cdot) = q(z_s \mid X^{\text{obs}})\, q(z_p \mid z_s, X^{\text{miss}})$ and applying the chain rule for divergences gives

$$\begin{aligned} &D_{\text{KL}}\big(q \,\|\, p_\psi(z_s \mid X^{\text{obs}})\, p_\phi(z_p \mid z_s)\big) \\ &\quad = D_{\text{KL}}\big(q(z_s \mid X^{\text{obs}}) \,\|\, p_\psi(z_s \mid X^{\text{obs}})\big) \\ &\qquad + \mathbb{E}_{z_s \sim q}\left[D_{\text{KL}}\big(q(z_p \mid z_s, X^{\text{miss}}) \,\|\, p_\phi(z_p \mid z_s)\big)\right], \end{aligned} \tag{A.15}$$

which separates semantic from structural regularization and mirrors the split between R1 and R2. The correspondence is conceptual, since neither divergence is computed during training.

From reconstruction to denoising. Let z_0 be the latent encoding of X^{miss} under the forward process $z_t = \sqrt{\bar\alpha_t} z_0 + \sqrt{1-\bar\alpha_t}\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$. The true posterior $q(z_{t-1} \mid z_t, z_0)$ is Gaussian with mean

$$\tilde\mu_t(z_t, z_0) = \frac{\sqrt{\bar\alpha_{t-1}}\beta_t}{1-\bar\alpha_t} z_0 + \frac{\sqrt{\alpha_t}(1-\bar\alpha_{t-1})}{1-\bar\alpha_t} z_t, \tag{A.16}$$

and the reverse model is Gaussian with mean parameterized through a noise predictor,

$$\mu_\theta(z_t, t, C) = \frac{1}{\sqrt{\alpha_t}}\left(z_t - \frac{\beta_t}{\sqrt{1-\bar\alpha_t}}\epsilon_\theta(z_t, t, C)\right), \tag{A.17}$$

where C collects the two conditioning signals. Matching the two reverse transitions reduces, up to timestep-dependent weights, to

$$\mathcal{L}_{\text{simple}} \propto \mathbb{E}_{z_0, \epsilon, t}\big[\|\epsilon - \epsilon_\theta(z_t, t, C)\|_2^2\big], \tag{A.18}$$

which is Eq. 9 of the main text. This connection does not imply that the full bound in Eq. A.14 is optimized, because its latent divergence terms are never computed.

B Method Details

B.1 Self-Reflective Abductive Reasoner

The prompt contract, output schema, and deterministic parsing rules referenced by the main text are stated in full in Table C.2, which is placed with the reproducibility material in Appendix C.1. This subsection specifies the verifier, its pre-training objective, the construction of negatives, the pseudo-reference subset, and the candidate-set bookkeeping.

Verifier architecture. The verifier encodes the candidate record and each modality-wise evidence item,

$$h_s^{(\ell)} = \mathrm{Enc}_{\mathrm{hyp}}(z_s^{(\ell)}), \qquad e_m^{(\ell)} = \mathrm{Enc}_m(r_m^{(\ell)}), \tag{B.1}$$

after which a lightweight cross-attention transformer lets the candidate state attend to the evidence set,

$$\tilde{h}_s^{(\ell)} = \mathrm{CrossAttn}\Big(h_s^{(\ell)}, \{e_m^{(\ell)}\}_{m\in\mathcal{M}_{\mathrm{obs}}}\Big), \tag{B.2}$$

and the compatibility energy is made non-negative by a softplus head,

$$E_{\mathrm{abd}}^{(\ell)} = \mathrm{softplus}\Big(w_\omega^\top \tilde{h}_s^{(\ell)} + b_\omega\Big). \tag{B.3}$$

Verifier pre-training. For a fully observed sample we construct a positive candidate record z_s^+ from complete evidence and hard negatives $\{z_s^{-,j}\}$ by label flipping, attribute shuffling, and pairing evidence with candidate records drawn from semantically different samples. The margin Δ is selected on the validation split and then fixed for all test runs. The verifier minimizes the margin objective

$$\begin{aligned}\mathcal{L}_{\mathrm{abd}} = \sum_j \Big[\Delta &+ E_\omega(h_s^+, \{e_m\}) \\ &- E_\omega(h_s^{-,j}, \{e_m\})\Big]_+,\end{aligned} \tag{B.4}$$

which calibrates the ordering of energies but not their scale. Scale is handled separately by the temperature of Eq. 6, and reporting the two steps apart is deliberate, since a margin loss cannot by itself produce probabilities.

Full-observation pseudo-references. Temperature fitting and candidate-weight calibration require a reference candidate. For each fully observed validation example, we run the same frozen reasoner on all available modalities and retain the resulting schema record only when every field that can be checked against dataset annotations is consistent with those annotations. We denote this retained record by z_s^{ref}. The retained index set is fixed before fitting the calibration temperature and is reused for all calibrated and uncalibrated variants. This construction is a *pseudo-reference*, not an oracle: fields without annotations inherit reasoner error, and consistency filtering conditions the evaluation subset on reasoner success. Candidate-weight calibration is therefore treated as an internal diagnostic on this fixed subset rather than as calibration against a human-annotated interpretation distribution.

Candidate set and deduplication. The candidate set $\mathcal{C}$ is the union of valid records produced along the refinement trajectory. Two records are merged when their schema fields match exactly after normalization of casing, whitespace, and numeric formatting, and the merged record keeps the lower energy. Because candidates come from one SRAR execution with at most $L_{\max} = 3$ deterministic refinement rounds, retaining J candidates does not trigger an additional reasoner execution. If all rounds fail schema validation, we insert a deterministic null record $z_s^{\varnothing}$ whose categorical fields are `unknown`, whose numeric intensity is zero, and whose evidence lists contain only normalized observations. This record receives unit weight, bypasses memory retrieval by setting $\gamma = 0$ and $z_p = z_p^0$, and is assigned $u = 1$ so that selective prediction treats the sample as maximally uncertain. No sample is discarded.

Calibration, stopping, and pruning. Weights follow Eq. 6 with the validation-fitted calibration temperature $T_{\mathrm{cal}} = 1.35$ (the scalar denoted by T in Eq. 6). Qwen decoding is deterministic at temperature 0. From the second refinement round onward, candidate-weight stability is defined as $1 - \mathrm{TV}(w^{(\ell)}, \tilde{w}^{(\ell-1)})$, where $\tilde{w}^{(\ell-1)}$ is the previous weight vector aligned to the current candidate set by candidate identity and zero-padded for newly introduced candidates. Refinement stops when this stability reaches $\tau_{\mathrm{abd}} = 0.8$ or when $\ell = L_{\max}$. Candidates whose weight falls below 0.08 are discarded and the remaining weights are renormalized, yielding $J = 2.7$ retained candidates on average. For the single-anchor case, we define $u = 0$; otherwise $u = H(w)/\log J$. Table G.2 compares the default operating point with the single-anchor $J = 1$ control.

Feedback construction. Let $a_{m,u}^{(\ell)}$ be the attention assigned to evidence unit u in modality m. We flag units whose attention is weak or whose local gradient indicates high incompatibility,

$$\mathcal{U}^{(\ell)} = \big\{(m,u) : a_{m,u}^{(\ell)} < \rho_a \ \text{ or } \ \|\nabla_{e_{m,u}} E_{\text{abd}}^{(\ell)}\|_2 > \rho_g\big\}, \tag{B.5}$$

The thresholds ρ_a and ρ_g are selected on validation data before test evaluation and then held fixed for all test runs; no test example is used for threshold selection. We convert $\mathcal{U}^{(\ell)}$ into a short directive such as re-checking pitch, facial expression, or a conflicting phrase. The directive is appended to the next evidence-distillation prompt. Because flagged units differ across candidates, feedback broadens the candidate set instead of merely repairing one record, which is what makes the loop a search over interpretations. Since feedback is constructed from a completed verification round, it has no effect on the first proposal, so a run with refinement limit $L = 0$ is identical for the score-only and score+feedback variants of Table E.6.

B.2 Semantic-Anchored Retrieval

Alignment objective. For each potential target modality m, fully observed training samples define a target-specific memory $\mathcal{B}_m = \{(k_i, v_{i,m})\}_{i=1}^{M}$. For a training example with ground-truth target feature $x^{(m^\star)}$, positives in $\mathcal{B}_{m^\star}$ are prototypes whose standardized target features lie within the validation-selected radius δ_p (we suppress the target index on v_j below),

$$\mathcal{P} = \big\{j : \|v_j - x^{(m^\star)}\|_2 \le \delta_p\big\}, \tag{B.6}$$

and the query projection and keys are pre-trained with a multi-positive InfoNCE loss,

$$\mathcal{L}_{\text{ret}} = -\log \frac{\sum_{j\in\mathcal{P}} \exp(s(q,k_j)/\tau_r)}{\sum_{j=1}^{M} \exp(s(q,k_j)/\tau_r)}. \tag{B.7}$$

After this stage the memory keys are frozen and only the projection head continues to receive gradients, which prevents the bank from drifting with the generator.

Per-candidate, per-target queries. Retrieval is executed for every retained candidate–target pair. The gated anchor $\bar{z}_s^{(j)}$ produces the branch query, and that query is evaluated against the target-specific bank $\mathcal{B}_m$ for each $m \in \mathcal{M}_{\text{miss}}$. Thus competing candidates can retrieve different prototype sets while different missing modalities draw their structural values from the appropriate target bank. This provides branch- and target-specific grounding but does not guarantee that the retrieved neighborhoods, or the generated branches, are disjoint. It is the operational mechanism associated with the grounding term ε_{gr} in Eq. A.8. Breadth is set by the ambiguity score of Eq. 6 rather than by an unbounded energy, so no additional scale constant has to be calibrated for retrieval.

Base and adaptive breadth. Full HyDiff adapts both the number of retrieved neighbors and the retrieval temperature according to the ambiguity score, with $K_{\text{eff}} \in [K_{\min}, K_{\max}]$ and $\sigma_{\text{eff}} \in [\sigma_0, \sigma_0 + \Delta\sigma]$. The main ablation in Table 3 removes this ambiguity-adaptive breadth by setting $u = 0$ in the breadth controller, which reduces the retrieval configuration to $K_{\text{eff}} = K_{\min} = 4$ and $\sigma_{\text{eff}} = \sigma_0 = 0.1$.
This ablation is distinct from the fixed-hyperparameter sensitivity study in Table E.7. In that study, K is swept while $\sigma = 0.1$ is held fixed, and σ is swept while $K = 10$ is held fixed. Consequently, the $K = 10$ and $\sigma = 0.1$ rows correspond to the same fixed default retrieval configuration and therefore report identical results.

Coverage gate diagnostics. The gate of Eq. 8 reduces reliance on retrieved prototypes when their maximum similarity is low. We evaluate this mechanism through two complementary interventions: Table E.9 reduces the amount of available memory, and the paired no-gate control in Table 3 removes fallback interpolation. Together they test coverage dependence and gate utility without treating the raw gate value itself as a calibrated confidence score.

B.3 RETRIEVAL-INJECTED DIFFUSION

Dual conditioning operators. The gated anchor is a sequence of field embeddings $\bar{Z}_s \in \mathbb{R}^{L_s \times d}$ rather than a single vector, so the cross-attention softmax is non-degenerate,

$$\begin{aligned} &\mathrm{Attn}(H, \bar{Z}_s) \\ &\quad = \mathrm{softmax}\left(\frac{(HW_Q)(\bar{Z}_s W_K)^\top}{\sqrt{d}} \right) \bar{Z}_s W_V, \end{aligned} \tag{B.8}$$

while the structural prior modulates normalization statistics,

$$\mathrm{AdaGN}(H, z_p) = \gamma_p(z_p) \odot \mathrm{GN}(H) + \beta_p(z_p). \tag{B.9}$$

Here H is an intermediate U-Net hidden sequence and L_s is the number of serialized anchor fields.

The verifier bridge Φ_{cross}. The verifier of Eq. B.3 is trained on encoded evidence descriptions, whereas cycle scoring must accept a completed feature. Φ_{cross} bridges the two spaces. It is a two-layer cross-attention block that takes the clean-latent estimate together with an observed evidence embedding e_m and returns a vector in the evidence space E_ω consumes. It is fitted during verifier pre-training on ground-truth completions, using the same margin objective of Eq. B.4 with real features substituted for the evidence slot, and it is frozen thereafter. Fitting it on real completions is what keeps the cycle energy in distribution at sampling time.

Reference-relative cycle supervision. Training a generator to reduce a frozen score invites off-manifold solutions that satisfy the scorer without resembling data. We therefore use a reference-relative hinge rather than minimizing the raw verifier energy without limit,

$$\mathcal{L}_{\mathrm{cyc}} = \mathbb{E}_{s \in \mathcal{S}_{\mathrm{check}},\, j} \Big[\big(E^{(s,j)}_{\mathrm{cyc}} - E^{(j)}_{\mathrm{real}} + m \big)_+ \Big], \tag{B.10}$$

where $E^{(j)}_{\mathrm{real}}$ is computed once per training sample by scoring the ground-truth completion under candidate j, and $m > 0$ is the implemented margin. The expectation in Eq. 11 of the main text is over the same guided iterations $\mathcal{S}_{\mathrm{check}}$ rather than over all diffusion timesteps, since the cycle energy is only evaluated where guidance is applied. With the $+m$ sign in Eq. B.10, the zero-loss region is $E^{(s,j)}_{\mathrm{cyc}} \le E^{(j)}_{\mathrm{real}} - m$. Thus, the objective deliberately uses a stricter reference-relative target; it does not define the real-data energy itself as a floor. Once the stricter threshold is reached, the cycle term no longer directly rewards further energy reduction. Appendix F.1 checks whether this pressure is accompanied by abnormal verifier energies or degraded feature fidelity.

Guidance schedule and gradient. Let $N_{\mathrm{samp}} = 50$ and let $s \in \{1, \dots, N_{\mathrm{samp}}\}$ denote the reverse-sampling iteration. The associated diffusion indices satisfy $t_1 > \cdots > t_{N_{\mathrm{samp}}} > t_{N_{\mathrm{samp}}+1} = 0$, so larger s denotes a later, cleaner stage of the trajectory. Guidance is inactive before $S_{\mathrm{start}} = 30$ and is ramped up linearly during late refinement,

$$\eta_{\mathrm{guide}}(s) = \eta_0 \max\left(0, \frac{s - S_{\mathrm{start}}}{N_{\mathrm{samp}} - S_{\mathrm{start}}}\right), \qquad \eta_0 = 0.25, \tag{B.11}$$

with checks at $\mathcal{S}_{\mathrm{check}} = \{30, 35, 40, 45, 50\}$; the corresponding diffusion index at iteration s is t_s. At iteration s, the clean-latent estimate is

$$\hat{z}_0^{(s)} = \frac{z_{t_s} - \sqrt{1 - \bar{\alpha}_{t_s}}\, \epsilon_\theta(z_{t_s}, t_s, \bar{z}_s, z_p)}{\sqrt{\bar{\alpha}_{t_s}}}. \tag{B.12}$$

The guided update of Eq. 10 follows from viewing the target as an energy-tilted density,

$$\begin{aligned} \nabla_{z_{t_s}} \log p(z_{t_s} \mid \bar{z}_s, z_p, X^{\mathrm{obs}}) \approx{}& \nabla_{z_{t_s}} \log p_\theta(z_{t_s} \mid \bar{z}_s, z_p) \\ &- \eta_{\mathrm{guide}}(s) \nabla_{z_{t_s}} E^{(s)}_{\mathrm{cyc}}. \end{aligned} \tag{B.13}$$

Since $\nabla_{z_{t_s}} \log p_\theta(z_{t_s}) = -\epsilon_\theta(z_{t_s}) / \sqrt{1 - \bar{\alpha}_{t_s}}$ under noise prediction, the negative gradient in score space becomes a positive correction in noise space. Evaluating $\nabla_{z_{t_s}} E^{(s)}_{\mathrm{cyc}}$ exactly would require differentiating through the denoiser at every guided step, because $\hat{z}_0^{(s)}$ depends on $\epsilon_\theta(z_{t_s})$. We use the standard approximation of treating ϵ_θ as constant, under which

$$\frac{\partial \hat{z}_0^{(s)}}{\partial z_{t_s}} = \frac{1}{\sqrt{\bar{\alpha}_{t_s}}} I, \qquad \nabla_{z_{t_s}} E^{(s)}_{\mathrm{cyc}} = \frac{\nabla_{\hat{z}_0} E^{(s)}_{\mathrm{cyc}}}{\sqrt{\bar{\alpha}_{t_s}}}. \tag{B.14}$$

The resulting cost is one backward pass through Φ_{cross} and E_ω per guided step, both of which are small relative to the U-Net, which is why the latency in Table G.3 is dominated by the reasoner and by sampling rather than by guidance. With deterministic DDIM sampling the corrected prediction yields

$$z_{t_{s+1}} = \sqrt{\bar{\alpha}_{t_{s+1}}}\hat{z}_0^{(s)} + \sqrt{1-\bar{\alpha}_{t_{s+1}}}\,\tilde{\epsilon}_\theta(z_{t_s}, t_s). \tag{B.15}$$

B.4 Mixture Prediction, Task Loss, and Inference

Each retained candidate produces a task score $\hat{s}_j$ and class-probability vector $\hat{p}_j$. We aggregate them as $\hat{s} = \sum_j w_j \hat{s}_j$ and $\hat{p} = \sum_j w_j \hat{p}_j$. For sentiment benchmarks the task objective is

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{reg}}(\hat{s}, s) + \lambda_{\text{cls}}\mathcal{L}_{\text{ce}}(\hat{p}, y), \qquad \lambda_{\text{cls}} = 1, \tag{B.16}$$

evaluated on the mixture output. The reasoner, verifier, candidate set, and calibration temperature are frozen during the second training phase, so w_j is a stop-gradient sample-wise coefficient. Task gradients are distributed to the trainable completion and prediction branches in proportion to w_j rather than updating the candidate weights themselves. For multi-label non-sentiment tasks the task term is binary cross-entropy. The denoising and cycle objectives are candidate-weighted,

$$\mathcal{L}_{\text{diff}} = \sum_{j=1}^{J} w_j \mathcal{L}_{\text{diff}}^{(j)}, \qquad \mathcal{L}_{\text{cyc}} = \sum_{j=1}^{J} w_j \mathcal{L}_{\text{cyc}}^{(j)}, \tag{B.17}$$

which keeps their scale invariant to the retained candidate count. For each fully observed training item and each target modality being simulated as missing, $\mathcal{L}_{\text{diff}}^{(j)}$ uses the same factual target stream for every retained j; there is no candidate-specific ground-truth completion assignment. Thus, the denoising term is compatible with learning to ignore the candidate, and candidate dependence must be learned through the conditioning architecture, retrieval, task gradients, and cycle surrogate. This limitation is why the main text treats branch separation as an empirical question rather than a property guaranteed by the objective. At inference HyDiff executes one SRAR procedure containing at most three refinement rounds, then performs target-wise retrieval and sampling for every retained candidate and returns the mixture prediction together with ambiguity u.

B.5 Algorithm

Algorithms B.1 and B.2 state the complete flow. The first constructs a temperature-scaled candidate-weight distribution with a deterministic failure path. The second executes candidate- and target-specific retrieval and diffusion for each retained candidate, aggregates candidate-weighted branch losses, and returns the mixture prediction and ambiguity score. In the training branch, TRAINING-COMPLETION denotes the differentiable target-wise denoising path defined by Eqs. 9, B.12, and B.10; it returns the clean-estimate completion set used by the task head together with the branch losses. Observed evidence is available to the cycle verifier bridge, but it is not concatenated to or otherwise supplied as a direct input to the diffusion denoiser.

C Reproducibility Details

This appendix specifies the implementation and evaluation protocol used to produce the main-paper and appendix results, including matched-input controls, multi-sample controls with approximately matched trajectory counts, anchor-access controls, candidate-weight calibration, and selective prediction.

C.1 Reasoner and Prompt Interface

We instantiate the omni-modal reasoner with QWEN2.5-OMNI-7B (Xu et al., 2025), frozen throughout training and inference. Unless stated otherwise we use deterministic decoding with temperature 0, maximum generation length 1024, and $L_{\max} = 3$. Inference uses bfloat16 with FlashAttention (Dao et al., 2022) where available. The reasoner emits a structured record rather than free-form text, which gives retrieval and diffusion a stable interface and makes schema-equality deduplication well defined. Table C.2 states the system instruction, the initial and refinement templates, the output schema, and the deterministic parsing, weighting, and failure-accounting rules referenced by Appendix B.1.

C.2 Datasets and Feature Preprocessing

We use the standard public splits and preserve the original sequence-level train, validation, and test partitions for all datasets. CMU-MOSI contains 1,284 training, 229 validation, and 686 test samples;

Algorithm B.1: HyDiff Stage 1: iterative candidate construction and weighting

Input: Observed raw streams R^{obs}, frozen verifier E_ω, T_{cal}, $L_{\max} = 3$, and $\tau_{\mathrm{abd}} = 0.8$.
Output: Candidate set $\mathcal{C}$, weights w, ambiguity u, and fallback flag $b_\varnothing$.

$fb^{(0)} \leftarrow \emptyset, \mathcal{C}^{(0)} \leftarrow \emptyset, w^{(0)} \leftarrow \emptyset$ **for** $\ell = 1$ **to** $L_{\max}$ **do**
  $\mathcal{E}^{(\ell)} \leftarrow \textsc{ExtractEvidence}(R^{\mathrm{obs}}, fb^{(\ell-1)})$ $z_s^{(\ell)} \leftarrow \textsc{Propose}(\mathcal{E}^{(\ell)})$
  $(E_{\mathrm{abd}}^{(\ell)}, fb^{(\ell)}) \leftarrow \textsc{Verify}(E_\omega, \mathcal{E}^{(\ell)}, z_s^{(\ell)})$ $\mathcal{C}^{(\ell)} \leftarrow \textsc{Merge}(\mathcal{C}^{(\ell-1)}, z_s^{(\ell)}, E_{\mathrm{abd}}^{(\ell)})$
  $w^{(\ell)} \leftarrow \textsc{Softmax}(-E_{\mathrm{abd}}/T_{\mathrm{cal}})$ over $\mathcal{C}^{(\ell)}$ **if** $\ell > 1$ **then**
    $\tilde{w}^{(\ell-1)} \leftarrow \textsc{AlignAndZeroPad}(w^{(\ell-1)}, \mathcal{C}^{(\ell-1)}, \mathcal{C}^{(\ell)})$ **if** $1 - \mathrm{TV}(w^{(\ell)}, \tilde{w}^{(\ell-1)}) \geq \tau_{\mathrm{abd}}$ **then**
      break
    end
  end
end
$\mathcal{C} \leftarrow \textsc{Prune}(\mathcal{C}^{(\ell)}, w^{(\ell)})$; renormalize w **if** $|\mathcal{C}| = 0$ **then**
  $\mathcal{C} \leftarrow \{z_s^{\varnothing}\}, w_1 \leftarrow 1, u \leftarrow 1, b_\varnothing \leftarrow 1$
else
  $J \leftarrow |\mathcal{C}|$; $u \leftarrow 0$ if $J = 1$, else $H(w)/\log J$; $b_\varnothing \leftarrow 0$
end
return $\mathcal{C}, w, u, b_\varnothing$

CMU-MOSEI contains 16,326 training, 1,871 validation, and 4,659 test samples; CH-SIMS v2.0 contains 2,722 training, 647 validation, and 1,034 test samples; and MM-IMDb contains 15,552 training, 2,608 validation, and 7,799 test samples. The completion backbone receives word-aligned BERT (Devlin et al., 2019), FACET (Stöckli et al., 2018), and COVAREP (Degottex et al., 2014) features projected to 128 dimensions, while the reasoner separately receives the available raw text, video, and audio streams. Dynamic masking is applied to training samples, and to evaluation samples only as declared by the test protocol.

For the non-sentiment transfer experiment we use MM-IMDb (Arevalo et al., 2017), a multi-label genre benchmark with plot text and poster images. The reasoner schema is adapted to task-aligned genre candidates and multimodal evidence cues, while the retrieval and diffusion mechanisms are unchanged. Because these semantic candidates overlap the task vocabulary, this experiment should be interpreted as task transfer with task-aligned semantic assistance rather than as label-agnostic generalization.

C.3 Baseline Reproduction Protocol

All locally trained methods use the same pre-extracted feature tensors, public train/validation/test partitions, missingness masks, seed set, and evaluation scripts used for HyDiff. Hyperparameters and early stopping are selected on validation data only, and the test set is evaluated once for each selected seed. Reported local results are means over the five seeds $\{13, 17, 23, 41, 89\}$. Any prototype or memory resource is constructed from training samples only; validation and test examples are never inserted. Input adapters may reconcile feature dimensions, but they may not expose an unavailable modality or change the realized mask assigned to a sample.

For a source-paper number to be compared directly, its dataset version, split, feature representation, missingness protocol, and metric definition must match the present evaluation. Otherwise the method must be rerun under the common protocol. Anchor-conditioned IMDer and P-RMF use the same masks and training budget as their unconditioned counterparts, and the additional hidden widths are restricted to the 4.8% parameter increase stated in Appendix C.6.

To avoid mixing incomparable settings, a baseline result is included only when its dataset version, public split, feature tensors, realized missingness masks, and metric implementation match the HyDiff evaluation. Trainable baselines start from their author-recommended configurations, use validation-only hyperparameter selection and early stopping, and are evaluated once on the test set for each selected seed. The same validation budget and stopping criterion are used for the anchor-conditioned controls.

Two aggregate conventions are used throughout and are defined once in Appendix D. All Mod-DW and Mod-PM values reported in this appendix are computed from the realized-pattern results of

Algorithm B.2: HyDiff Stages 2–3: target-wise retrieval, completion, and prediction

Input: $\mathcal{C}, w, u, b_\varnothing$, observed features X^{obs}, missing-target set $\mathcal{M}_{\text{miss}}$, target-indexed training memories $\{\mathcal{B}_m\}$, and $N_{\text{samp}} = 50$.

Output: Candidate-conditioned completion sets, prediction $\hat{y}$, and ambiguity u.

$\mathcal{L}_{\text{diff}} \leftarrow 0$, $\mathcal{L}_{\text{cyc}} \leftarrow 0$ **foreach** $z_s^{(j)} \in \mathcal{C}$ **do**
  $\bar{z}_s^{(j)} \leftarrow \textsc{GatedProject}(z_s^{(j)})$ $(K_{\text{eff}}, \sigma_{\text{eff}}) \leftarrow \textsc{AdaptiveBreadth}(u)$ $q^{(j)} \leftarrow \Phi_{\text{proj}}(\bar{z}_s^{(j)})$
  foreach $m \in \mathcal{M}_{\text{miss}}$ **do**
    if $b_\varnothing = 1$ **then**
      $z_{p,m}^{(j)} \leftarrow z_p^0$, $\gamma_m^{(j)} \leftarrow 0$
    else
      $\mathcal{N}_m \leftarrow \textsc{TopK}(\mathcal{B}_m, q^{(j)}, K_{\text{eff}})$ $\tilde{z}_{p,m}^{(j)} \leftarrow \textsc{Aggregate}(\mathcal{N}_m, q^{(j)}, \sigma_{\text{eff}})$
      $z_{p,m}^{(j)} \leftarrow \textsc{CoverageFallback}(\bar{z}_s^{(j)}, \tilde{z}_{p,m}^{(j)}, z_p^0)$
    end
  end
  if *Training* **then**
    $(\hat{X}^{\text{miss},(j)}, \ell_{\text{diff}}^{(j)}, \ell_{\text{cyc}}^{(j)})$
    $\leftarrow \textsc{TrainingCompletion}(\{X_m\}_{m \in \mathcal{M}_{\text{miss}}}, \bar{z}_s^{(j)}, \{z_{p,m}^{(j)}\}_{m \in \mathcal{M}_{\text{miss}}}; X^{\text{obs}})$
    $\mathcal{L}_{\text{diff}} \mathrel{+}= \textsc{StopGrad}(w_j)\ell_{\text{diff}}^{(j)}$ $\mathcal{L}_{\text{cyc}} \mathrel{+}= \textsc{StopGrad}(w_j)\ell_{\text{cyc}}^{(j)}$
  else
    foreach $m \in \mathcal{M}_{\text{miss}}$ **do**
      sample $z_{t_1}^{(m)} \sim \mathcal{N}(0, I)$ **for** $s = 1$ **to** N_{samp} **do**
        predict noise from $(z_{t_s}^{(m)}, t_s, \bar{z}_s^{(j)}, z_{p,m}^{(j)})$; apply Eq. 10 when $s \in \mathcal{S}_{\text{check}}$
        $z_{t_{s+1}}^{(m)} \leftarrow \text{DDIMStep}(z_{t_s}^{(m)}, t_s, \epsilon)$
      end
      $\hat{X}_m^{(j)} \leftarrow \mathcal{D}_m(z_0^{(m)})$
    end
    $\hat{X}^{\text{miss},(j)} \leftarrow \{\hat{X}_m^{(j)} : m \in \mathcal{M}_{\text{miss}}\}$
  end
end
$\hat{y} \leftarrow \sum_j w_j \textsc{TaskHead}(X^{\text{obs}}, \hat{X}^{\text{miss},(j)})$ **if** *Training* **then**
  optimize Eq. 13 with the frozen components specified above
end
return $\{\hat{X}^{\text{miss},(j)}\}, \hat{y}, u$

main-paper Table 1 using Eqs. D.3 and D.4, so they agree with the aggregate gains reported in the main text by construction.

C.4 Matched-Input Feature-Corruption Control

Main-paper Table 2 separates the raw-assisted Intra-RA diagnostic (abbreviated I-RA in all tables) from a strictly matched-input comparison. In HyDiff-Feat, the raw text, waveform, and video frames are disabled. The model receives only the same masked BERT, FACET, and COVAREP tensors, sequence masks, and binary availability indicators supplied to P-RMF. The masked tensors are processed by the same modality encoders used by the completion backbone; pooled feature-derived evidence states replace raw-stream evidence in candidate construction. No unmasked coordinate, raw-stream token, or target-modality feature is exposed to the reasoner or verifier.

The retrieval memory, diffusion architecture, task head, dynamic training masks, seed set, optimization schedule, early-stopping rule, and validation budget are unchanged. Thus, P-RMF versus HyDiff-Feat isolates the interpretation–retrieval–completion pipeline under matched observations. HyDiff-RA is reported only as a raw-assisted reference and is not used as the input-matched comparison. Tables that report an I-RA column for HyDiff without further qualification refer to HyDiff-RA, since the raw-assisted configuration is the default model; the matched-input variant is always named HyDiff-Feat explicitly.

Control	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	I-RA	M-PM	I-RA	M-PM	I-RA	M-PM
(a) Reasoner as predictor	58.43	62.34	62.05	65.22	57.97	61.22
(b) Anchor with linear head	64.83	68.31	68.25	70.72	63.44	67.13
(c) IMDer with matched anchor	72.15	77.22	77.98	78.69	72.73	75.80
(d) P-RMF with matched anchor	72.77	77.62	78.47	78.94	73.94	76.06
(e) HyDiff, explicit target fields removed	72.06	77.23	78.37	78.74	73.37	76.06
(f) P-RMF, reduced matched anchor	71.12	76.24	77.35	77.58	72.31	74.87
(g) HyDiff, same reduced anchor	**71.68**	**76.92**	**78.14**	**78.03**	**72.86**	**75.57**
HyDiff (full)	**73.68**	**78.43**	**79.55**	**79.40**	**74.62**	**76.58**

Table C.1: Matched-access and semantic-anchor controls (Acc-2, %). I-RA is the raw-assisted feature-corruption diagnostic at $\eta = 0.9$; M-PM is the equal-weight macro-average over the seven non-empty modality-availability patterns at pre-rejection $\mu = 0.7$. Rows (c)–(d) use the full projected semantic anchor; row (e) removes `polarity`/`intensity` only. Rows (f)–(g) use the same stricter reduced-anchor protocol with `polarity`, `intensity`, and `dominant_affect` removed. The HyDiff (full) row reproduces the corresponding entries of the main paper.

C.5 Multi-Sample Controls with Approximately Matched Trajectory Counts

Main-paper Table 2 separates generic stochastic sampling, repeated top-1 candidate conditioning, and maintaining distinct candidate-conditioned branches. The controls use the same observed inputs, projected feature dimensions, 50-step DDIM sampler, task head, missingness masks, seed set, and validation budget. The multi-sample and repeated-top-1 controls each use three diffusion trajectories, which approximately matches HyDiff's average retained candidate count of $J = 2.7$. We therefore describe the comparison as approximately matched in trajectory count rather than compute-matched; it is not an exact wall-clock or per-sample trajectory match, because HyDiff retains a variable number of candidates and SRAR plus candidate- and target-specific retrieval introduce additional overhead quantified in Appendix G.

Single-sample diffusion. The semantic candidate set and candidate- and target-specific retrieval are removed. The resulting non-semantic diffusion control produces one completion for each missing target, and the recovered stream(s) are combined with the observed streams by the shared task head.

Multi-sample diffusion. The same non-semantic diffusion control generates three independent recovery trajectories from different Gaussian initializations. Each trajectory yields one task prediction, and the three predictions are averaged with equal weight.

Repeated top-1 anchor. SRAR is run once, the highest-weight candidate is selected, and the same semantic anchor together with its target-specific retrieved prior(s) is reused for three independent recovery trajectories. Their task predictions are averaged uniformly. This preserves top-1 candidate conditioning and the number of trajectories while removing variation in semantic conditioning across trajectories.

C.6 Anchor-Control Protocols

Table C.1 is the paper's principal control, so we state exactly how each row is produced.

Row (a), reasoner as predictor. The reasoner runs on the same masked inputs given to HyDiff and its anchor field is mapped to a task prediction by the fixed rule described in the released prompt, with no retrieval, no diffusion, and no trained head. Records that fail schema parsing are counted as errors rather than discarded, since discarding them would inflate the control.

Row (b), anchor with linear head. A linear head is trained on $\mathrm{Enc}_{\mathrm{hyp}}(z_s)$ alone, using the same seeds and budget as the full model. This bounds what the schema carries after the reasoner, independent of the schema-to-label rule used in row (a).

Rows (c) and (d), baselines with anchor. IMDer and P-RMF receive $\bar{z}_s$ as an additional conditioning vector, concatenated at the point each method already accepts conditioning, with hidden widths adjusted so the anchor-conditioned variants increase parameter count by 4.8%. This control removes the conditioning-capacity asymmetry while preserving the same validation budget as the corresponding unconditioned baseline.

Item	Specification
System instruction	"You are a multimodal evidence analyst. Infer one task-relevant interpretation from the observed modalities. Use only the supplied evidence, identify conflicts explicitly, and return only a JSON object matching the required schema. Do not report self-confidence."
Initial user template	"Task: [task description]. Observed modalities: [list]. Missing modalities: [list]. Text evidence: [serialized text cues]. Visual evidence: [serialized visual cues]. Acoustic evidence: [serialized acoustic/prosodic cues]. Produce one interpretation using the required JSON schema."
Refinement template	The initial template is followed by "Previous candidates: [normalized records]. Verifier feedback: [flagged evidence and conflict directive]. Recheck the cited evidence, propose a distinct interpretation when supported, and return one JSON object only."
Sentiment JSON schema	{`polarity: string, intensity: number, dominant_affect: string, salient_event: string, modality_evidence: {text: [string], visual: [string], audio: [string]}, uncertainty_notes: [string]`}. The explicit-target-field control drops `polarity`/`intensity` but retains `dominant_affect`; the stricter reduced-anchor control additionally removes `dominant_affect`, retaining only event/evidence/uncertainty fields. For MM-IMDb, the schema instead contains candidate genre cues together with plot, visual, and cross-modal evidence fields; these candidates are generated from the observed modalities and are not supplied from ground-truth test annotations.
Parsing and normalization	A record is accepted only when all required keys and value types are present. Strings are stripped, lower-cased for equality tests, and whitespace is collapsed; numeric fields are converted to a canonical decimal representation. Extra explanatory text outside the JSON object invalidates the record. Candidate equality and deduplication use these normalized fields.
Verification and feedback	The learned scorer E_ω evaluates candidate–evidence compatibility. Weakly attended or high-gradient evidence units form the next-round directive. Self-reported model confidence is ignored.
Weighting and stopping	Energies are converted to weights by Eq. 6. From round two onward, align the previous candidate-weight vector by candidate identity and zero-padding; stop when $1 - \mathrm{TV}(w^{(\ell)}, \tilde{w}^{(\ell-1)}) \geq \tau_{\mathrm{abd}} = 0.8$ or $\ell = L_{\max}$, then prune and renormalize.
Failure accounting	An invalid record contributes no candidate. In the reasoner-only control, a parsing failure is counted as an incorrect prediction rather than discarded. If all full-model rounds fail, SRAR inserts $z_s^{\varnothing}$, sets $w_1 = 1$ and $u = 1$, bypasses candidate- and target-specific retrieval by setting $z_{p,m} = z_p^0$ for every missing target, and continues the standard completion path.

Table C.2: SRAR prompt contract and deterministic parsing rules. Dataset-specific prompts instantiate the bracketed fields without changing the output schema.

Method	MOSI	MOSEI	SIMS	Avg.
Single-sample diffusion	74.05	76.70	72.95	74.57
Multi-sample diffusion ($J = 3$)	76.18	77.68	74.62	76.16
Repeated top-1 anchor ($J = 3$)	77.15	78.52	75.52	77.06
HyDiff	**78.43**	**79.40**	**76.58**	**78.14**

Table C.3: Per-dataset Mod-PM Acc-2 (%) for the candidate-mechanism controls summarized in main-paper Table 2.

Row (e), explicit-target-field removal. The schema drops `polarity` and `intensity` but retains `dominant_affect`, `salient_event`, and `modality_evidence`. Thus it removes the two explicit sentiment-target fields, not all affective semantics: `dominant_affect` may still correlate with sentiment. Everything else is unchanged, including re-training the verifier for the modified anchor input. Row (e) therefore isolates the contribution of those two fields.

Rows (f) and (g), matched reduced-anchor comparison. The stricter reduced schema removes `polarity`, `intensity`, and `dominant_affect`, retaining `salient_event`, `modality_evidence`, and `uncertainty_notes`. P-RMF and HyDiff use the same reduced-anchor generation protocol, reasoner, observed inputs, missing masks, seed set, training budget, and validation protocol; the recovery mechanism is the intended difference. This is a target-field-reduced comparison rather than a claim of label-free semantics, because the retained evidence can still carry sentiment-relevant information.

Full HyDiff exceeds the reasoner-only control by 15.84 points on average over the six columns. Relative to full-anchor P-RMF, the average margin is 0.74 points. Removing only `polarity`/`intensity` lowers HyDiff by 1.07 points, confirming that explicit task-facing fields materially contribute to performance.

The stricter matched comparison yields a six-setting mean of 74.91 for P-RMF and 75.53 for HyDiff. HyDiff is higher in every column by 0.45–0.79 points (0.62 on average) and in $26/30$ paired seed runs. We therefore interpret the architecture-specific effect as smaller than the total system gain but directionally consistent after removing the three most target-facing fields. Because the frozen reasoner's pretraining corpus is not fully auditable and MOSI/MOSEI are long-standing public benchmarks, prior exposure to related public content cannot be ruled out; the matched-input and reduced-anchor controls reduce, but do not eliminate, this threat.

C.7 Calibration and Selective-Prediction Protocol

We distinguish candidate-weight calibration from task-level confidence calibration (Guo et al., 2017) and report them separately because they define different events. For HyDiff, we use a quantile-binned ECE variant with 15 equal-mass bins over $\max_j w_j$ on the fixed pseudo-reference subset defined in Appendix B.1; a prediction is correct when the normalized arg-max candidate equals z_s^{ref}. The uncalibrated variant sets $T_{\mathrm{cal}} = 1$ and is otherwise identical. Table C.5 therefore supports only the within-HyDiff conclusion that validation scaling improves candidate-weight calibration against these pseudo-references.

A single-completion baseline does not expose a candidate-weight distribution over alternative interpretations. For P-RMF, $q_{\text{P-RMF}} = \max_c p_c$ defines confidence in the final task class, and Table C.4 reports the corresponding task-level ECE. This value is a separate baseline diagnostic and is never ranked against HyDiff's candidate-weight ECE.

For selective prediction, HyDiff task confidence is defined after the candidate-weighted aggregation used for prediction,

$$q_{\text{HyDiff}} = \max_c \left(\sum_j w_j \hat{p}_{j,c} \right), \tag{C.1}$$

where $\hat{p}_{j,c}$ is branch j's probability for task class c. Thus, the P-RMF and HyDiff task-confidence curves rank examples by the maximum class probability of their respective final predictive distributions. Multi-sample disagreement and candidate ambiguity u provide alternative HyDiff rankings. Ties are broken by a stable sample identifier. For seed q, let π_q denote a reliability ordering and define the prefix risk

$$R_q(k) = \frac{1}{k} \sum_{n=1}^{k} \mathbf{1}\left[\hat{y}_{\pi_q(n)} \neq y_{\pi_q(n)}\right], \qquad k = 1, \ldots, N. \tag{C.2}$$

The discrete area under the risk–coverage curve follows the empirical AURC construction (Zhou et al., 2025):

$$\mathrm{AURC}_q = \frac{1}{N} \sum_{k=1}^{N} R_q(k), \qquad \mathrm{AURC} = \frac{1}{5} \sum_{q=1}^{5} \mathrm{AURC}_q \,. \tag{C.3}$$

AURC is computed on the complete stochastic Mod test set at pre-rejection $\mu = 0.7$. We do not macro-average separately computed pattern-level AURCs and do not interpolate from displayed coverage points. At coverage c, each method retains its own most reliable fraction and reports accuracy against the same downstream task outcome.

Task-confidence diagnostic	MOSI	MOSEI	SIMS	Avg.
P-RMF task-level ECE	0.089	0.080	0.089	0.086

Table C.4: Task-level calibration of P-RMF under sample-level Mod-DW at pre-rejection $\mu = 0.7$ (lower is better). Correctness refers to the downstream task class. These values are not comparable to candidate-weight ECE.

HyDiff candidate weights	MOSI	MOSEI	SIMS	Avg.
$T_{\text{cal}} = 1$	0.071	0.075	0.074	0.073
$T_{\text{cal}} = 1.35$	**0.043**	**0.043**	**0.058**	**0.048**

Table C.5: Candidate-weight ECE for HyDiff under sample-level Mod-DW at pre-rejection $\mu = 0.7$ (lower is better). Correctness is defined against the fixed full-observation pseudo-reference subset, not a human-annotated interpretation distribution. The comparison is strictly within HyDiff.

Selective-prediction ranking signal	MOSI	MOSEI	SIMS	Avg.
P-RMF task confidence	0.216	0.204	0.221	0.214
HyDiff task confidence	0.187	0.166	0.199	0.184
Multi-sample disag.	0.169	0.154	0.173	0.165
HyDiff ambiguity u	**0.152**	**0.142**	**0.163**	**0.152**

Table C.6: AURC under sample-level Mod-DW at pre-rejection $\mu = 0.7$ (lower is better). Every ranking signal is evaluated against the same downstream task error; only the ordering of retained examples differs.

C.8 Implementation Settings

Tables C.7 and C.8 summarize the architecture, inference, optimization, and masking settings used in the main experiments. Validation-selected thresholds are fixed for all test runs.

Component	Setting
Feature projection	BERT (Devlin et al., 2019) 768-d, FACET 35-d, and COVAREP (Degottex et al., 2014) 74-d projected to 128-d; standardized missing entries are zero-filled with a binary availability mask
Predictor	2-layer ReLU MLP, hidden size 128, dropout 0.1; $\lambda_{\text{cls}} = 1$
Reasoner	Qwen2.5-Omni-7B (Xu et al., 2025), $L_{\max} = 3$, $\tau_{\text{abd}} = 0.8$, temperature 0, maximum 1024 generated tokens
Verifier	Lightweight cross-attention transformer over 128-d candidate and evidence states; margin Δ and feedback thresholds ρ_a, ρ_g are selected on validation data and fixed for testing
Candidate weighting	$T_{\text{cal}} = 1.35$; pruning threshold 0.08; mean $J = 2.7$; branch losses use stop-gradient candidate weights
Retrieval	Base (fixed-breadth) $K = 10$, $\sigma_0 = 0.1$; adaptive range $K_{\text{eff}} \in [4, 12]$ and $\sigma_{\text{eff}} \in [0.1, 0.16]$ with $\Delta\sigma = 0.06$; δ_p and τ_r selected on validation data

Table C.7: Core architecture and inference settings used in the main experiments.

Component	Setting
Bridge	2-layer cross-attention mapping completed features and observed evidence to the verifier evidence space
Diffusion	Sequence-aware 1D U-Net with kernel-size-3 residual blocks, group normalization, dual semantic/structural conditioning, $N_{\text{diff}} = 1000$, and 50-step DDIM (Song et al., 2021) sampling
Cycle	Margin $m = 0.12$; guidance $\eta_0 = 0.25$, $S_{\text{start}} = 30$; $\mathcal{S}_{\text{check}} = \{30, 35, 40, 45, 50\}$
Optimization	Adam, learning rate 2×10^{-3}, weight decay 10^{-4}; $\lambda_1 = 0.1$, $\lambda_2 = 0.05$; halve the learning rate after 5 stagnant validation epochs; seeds $\{13, 17, 23, 41, 89\}$
Masking	50% of training samples remain complete; the corrupted half is balanced between Intra and Mod. Intra uses $\eta \sim \mathcal{U}(0, 1)$ and Mod uses independent drop-and-redraw
Batch size	32 for CMU-MOSI and CH-SIMS; 128 for CMU-MOSEI

Table C.8: Generation, optimization, and masking settings used in the main experiments.

Hyperparameter selection rationale. All validation-selected values are chosen before test evaluation and then fixed across the reported test conditions. The temperature $T_{\text{cal}} = 1.35$ minimizes candidate-weight ECE on the fixed pseudo-reference validation subset rather than maximizing test accuracy. The refinement limit $L_{\max} = 3$ is selected at the accuracy–cost knee observed in Table E.6: feedback-assisted refinement improves through three rounds and then saturates or declines. The base retrieval values $K = 10$ and $\sigma_0 = 0.1$ are the maxima of the explicit sweeps in Table E.7; the adaptive range $K \in [4, 12]$ and $\sigma \in [0.1, 0.16]$ is centered on that operating region so that ambiguity changes breadth without allowing unbounded neighborhoods. We set $\lambda_1 = 0.1$ from the loss-weight sweep in Table E.10. The cycle term is kept auxiliary with $\lambda_2 = 0.05$, while $m = 0.12$ defines the verifier-energy tolerance and $\eta_0 = 0.25$ limits guidance to a small late-sampling correction; these three values are selected on validation performance and then fixed. The 50-step DDIM schedule is used as the accuracy-oriented default, while Appendix G separately reports the latency effect of a 10-step sampler without treating it as an accuracy-equivalent replacement.

D Backbone, Missingness Protocols, and Optimization

D.1 Sequence-Aware Dual-Conditioned 1D U-Net

For a missing target modality m, HyDiff implements $p_\theta(X_m \mid z_s, z_{p,m})$ with a Sequence-Aware Dual-Conditioned 1D U-Net designed for word-aligned multimodal sequences; the target index is suppressed in the architectural details below. When multiple modalities are missing, the same target-wise procedure is applied to each one and the recovered streams are collected before task fusion. Compared with generic diffusion backbones it targets two properties, namely temporal dependency in sentiment dynamics and support for two heterogeneous conditioning signals produced by reasoning and retrieval.

Encoder and decoder with temporal residual blocks. The backbone follows a temporal U-Net layout with downsampling and upsampling paths joined by skip connections. Each resolution contains kernel-size-3 temporal convolutions, residual connections, and group normalization. The anchor sequence enters through multi-head cross-attention at the bottleneck and intermediate resolutions, while the structural prior modulates residual blocks through AdaGN.

Dual conditioning. The two conditions enter at different functional locations to reflect their roles.

1. **Semantic conditioning for global coherence.** The anchor represents an abstract hypothesis about the missing modality. It is injected through cross-attention at the bottleneck and optionally at multiple scales, so the denoising trajectory follows sequence-level semantics such as affective direction and discourse cues.
2. **Structural conditioning for local pattern.** The retrieved prior supplies prototype-like, instance-grounded statistics through adaptive group normalization,

$$\text{AdaGN}(f, z_p) = \gamma(z_p) \cdot \frac{f - \mu}{\sigma} + \beta(z_p), \tag{D.1}$$

Availability pattern	Probability (%)
ooo	4.11
Each of xoo, oxo, oox	9.59
Each of xxo, xox, oxx	22.37

Table D.1: Conditional Mod availability distribution at pre-rejection $\mu = 0.7$ after rejecting the all-missing pattern.

	MOSI		MOSEI		CH-SIMS	
Method	DW	PM	DW	PM	DW	PM
MissRAG	74.26	77.07	75.93	77.93	73.24	75.28
HyDiff	**75.54**	**78.43**	**77.24**	**79.40**	**74.48**	**76.58**

Table D.2: Mod-DW and Mod-PM Acc-2 (%) at pre-rejection $\mu = 0.7$, computed from the realized-pattern results of Table 1 with Eqs. D.2–D.4. These values reproduce the main-text aggregate results and are repeated here so the aggregation formulas and the numbers they produce appear together.

where μ and σ are computed per group and channel and γ, β are learned affine projections of z_p. This recovers fine-grained attributes such as intensity, rhythm, and local facial dynamics while remaining consistent with the higher-level constraint the anchor carries.

D.2 Missingness Protocols and Optimization

Intra-modal missingness. For feature-level corruption, a Bernoulli mask is sampled independently for each feature coordinate with pre-mask probability η. Masked coordinates are replaced by the same missing-value representation used during training. Evaluation reports $\eta \in [0, 0.9]$, with $\eta = 0.9$ as the severe Intra setting.

Modality-level missingness. We use one stochastic modality-drop sampler throughout the Mod experiments. For each sample and each modality, an independent Bernoulli variable drops that modality with pre-rejection probability μ. If all modalities are dropped, the draw is rejected and resampled. Hence, μ is a nominal pre-rejection probability rather than the realized fraction of missing modalities. Figure 3 sweeps this sampler over $\mu \in [0.1, 0.7]$. Table 1 uses the same sampler at $\mu = 0.7$ and groups test samples by their realized availability pattern. Its pattern columns are therefore conditional reporting bins, not a separate fixed-mask experiment.
For the three-modality sentiment datasets, these probabilities follow

$$P(r \mid r \neq \mathrm{xxx}) = \frac{P(r)}{1 - \mu^3}. \tag{D.2}$$

For the two-modality MM-IMDb setting, the same rule rejects xx; at $\mu = 0.7$, oo has conditional probability 17.65%, and each one-observed pattern has conditional probability 41.18%.
Let $\mathcal{R}$ denote the seven non-empty sentiment availability patterns. We report two aggregates rather than using the label "Mod" for two different quantities. The sample-level distribution-weighted metric is

$$\mathrm{Acc}_{\mathrm{DW}} = \sum_{r \in \mathcal{R}} P(r \mid r \neq \mathrm{xxx})\, \mathrm{Acc}(r), \tag{D.3}$$

whereas the pattern-macro metric is

$$\mathrm{Acc}_{\mathrm{PM}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \mathrm{Acc}(r), \qquad |\mathcal{R}| = 7. \tag{D.4}$$

Mod-DW estimates expected sample-level performance under the declared stochastic sampler; Mod-PM gives equal importance to frequent and rare availability patterns. Throughout the appendix, M-PM is used as the abbreviation for Mod-PM and I-RA for the raw-assisted Intra diagnostic.

Dynamic missingness during training. For each epoch, 50% of the training samples are left fully observed and 50% are dynamically corrupted. The corrupted half is split equally between Intra and Mod. Under Intra corruption, we draw $\eta \sim \mathcal{U}(0, 1)$ and sample the feature mask independently. Under Mod corruption, we draw a pre-rejection rate $\mu \sim \mathcal{U}(0, 1)$ and apply the same independent-drop-and-redraw sampler used at evaluation. After per-dimension standardization, missing entries

Distillation Strategy	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	I-RA	M-PM	I-RA	M-PM	I-RA	M-PM
No-Focus	66.43	71.41	72.45	73.12	67.15	69.82
Single-Focus	70.12	75.32	76.12	76.85	70.85	73.42
Dual-Focus	72.15	77.15	78.35	78.42	73.12	75.55
Feedback-Driven	**73.68**	**78.43**	**79.55**	**79.40**	**74.62**	**76.58**

Table E.1: Evidence-distillation ablation (Acc-2, %). I-RA uses raw-assisted feature corruption at $\eta = 0.9$ and M-PM uses the pattern-macro modality-level protocol at pre-rejection $\mu = 0.7$. Bold marks the highest reported mean.

are set to zero and a binary availability mask is concatenated before feature projection. This exposes the model to a continuum of corruption severities rather than matching only the severe test settings $\eta = 0.9$ and $\mu = 0.7$.

Optimization and inference. We optimize with Adam (Kingma & Ba, 2015) at learning rate 2×10^{-3} and weight decay 10^{-4}, halving the rate after five validation epochs without improvement. Batch sizes are 32 for CMU-MOSI and CH-SIMS and 128 for CMU-MOSEI. Diffusion uses $N_{\text{diff}} = 1000$ training timesteps and 50-step DDIM inference. In Eq. 13, $\lambda_1 = 0.1$ weights denoising and $\lambda_2 = 0.05$ weights cycle consistency; Eq. B.16 uses $\lambda_{\text{cls}} = 1$. Experiments use one NVIDIA A100 with 80GB.

E Additional Ablations and Sensitivity Analyses

This section reports additional ablations and sensitivity analyses covering evidence distillation, extended availability-pattern metrics, retrieval query and aggregation, refinement depth, retrieval hyperparameters, memory coverage, dual conditioning, loss weights, and anchor projection.

E.1 Evidence Distillation

The four variants keep the reasoner backbone, verifier, retrieval module, diffusion model, and optimization budget fixed, and vary only how observed evidence is condensed before candidate generation.

- **No-Focus** uses one-shot schema generation from all observed evidence, without an explicit evidence-selection stage or verifier feedback.
- **Single-Focus** first extracts one global list of salient evidence units shared by all schema fields, then generates the candidate set in one pass without iterative feedback.
- **Dual-Focus** separately distills task-semantic cues and modality-specific observable cues before one-pass candidate generation, but does not feed verifier diagnostics back to the reasoner.
- **Feedback-Driven** is the full SRAR procedure: it starts from the dual-focus record and appends candidate-specific weak-attention or high-gradient directives from Eq. B.5 in subsequent refinement rounds.

All variants use deterministic decoding and the same downstream candidate-weighting, retrieval, and completion pipeline, so the comparison targets evidence selection and iterative refocusing rather than model capacity.

Feedback-driven distillation gives the highest reported mean on every dataset and protocol. Under our formulation, feedback can flag different evidence for different candidates and thereby support a broader candidate set rather than only sharpening one record. The contribution of retaining multiple candidates is evaluated by the single-anchor R1 control in Table 3 and summarized against the default operating point in Table G.2.

E.2 Pseudo-Reference Candidate Coverage

We directly diagnose the upstream failure mode in which the reasoner does not retain a useful interpretation. For each incomplete input in the fixed pseudo-reference subset of Appendix B.1, we test whether the complete-observation pseudo-reference matches the first proposal, the final highest-weight candidate, or any candidate retained after pruning. The same normalized schema-equality rule used by candidate-weight calibration is used here. This diagnostic does not measure recall of

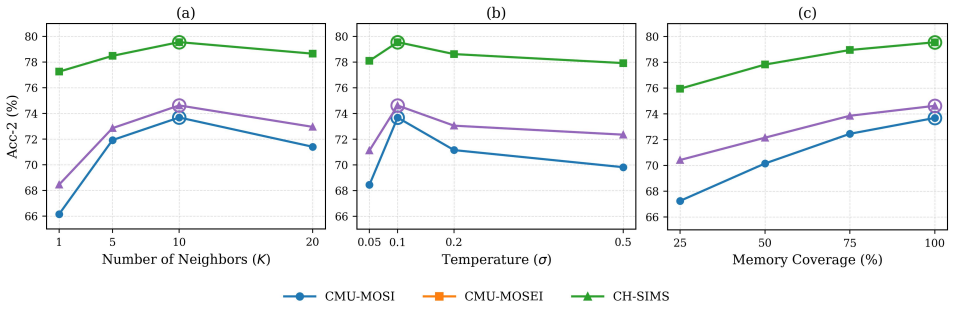

Figure E.1: Sensitivity to neighbor count K, retrieval temperature σ, and memory coverage $\mathcal{B}$ under Intra corruption at $\eta = 0.9$. Circle markers identify the highest reported means.

Dataset	First proposal	Final top-1	Set coverage	Null fallback
CMU-MOSI	61.3	71.8	86.4	1.18
CMU-MOSEI	65.2	76.6	90.3	0.74
CH-SIMS	58.7	68.4	84.9	1.42
Average	61.7	72.3	87.2	1.11

Table E.2: Pseudo-reference candidate coverage (%). First proposal denotes the initial structured candidate before feedback refinement; final top-1 is the highest-weight retained candidate; set coverage records whether any retained candidate matches the full-observation pseudo-reference. Null fallback denotes all-round schema-validation failure. The pseudo-reference is diagnostic rather than human-annotated ground truth.

all human-plausible interpretations: z_s^{ref} is produced by the same frozen reasoner under complete observations and is therefore only a controlled internal reference.

Feedback-assisted refinement raises top-1 pseudo-reference matching over the first proposal by 10.5, 11.4, and 9.7 percentage points on CMU-MOSI, CMU-MOSEI, and CH-SIMS, respectively. More importantly, retained-set coverage exceeds final top-1 matching by 14.6, 13.7, and 16.5 points, reaching an average coverage of 87.2%. This is consistent with the role of set-valued reasoning: a pseudo-reference can remain available downstream even when it is not assigned the largest candidate weight. Nevertheless, 9.7–15.1% of samples remain uncovered, confirming candidate proposal as an upstream limitation that subsequent retrieval and diffusion cannot repair. Null fallback is rare (0.74–1.42%), so most uncovered cases reflect semantic candidate omission rather than schema-parsing failure.

E.3 Extended Results under Modality Availability

All rows in Tables E.3 and E.4 follow the common splits, feature tensors, realized masks, seed set, metric implementation, and reproduction rules of Appendix C.3. Beyond Acc-2, Table E.3 reports Acc-5 for CMU-MOSI and CMU-MOSEI. For CH-SIMS, we retain the standard Acc-3 metric in Table E.4 and do not introduce an Acc-5 based on manually selected dataset-specific thresholds. Availability patterns follow [L, V, A], where ○ is observed and × is missing. As in Table 1, columns condition on the realized pattern produced by the Mod sampler at pre-rejection $\mu = 0.7$.

The baseline set here is a subset of the one used in main-paper Table 1. We evaluated MPLMM, T^{2}DR, and MissRAG under the common protocol for Acc-2 only and did not run additional Acc-5, Acc-7, or Acc-3 variants for these methods. We therefore restrict Tables E.3 and E.4 to methods for which the corresponding finer-granularity results were actually obtained, rather than extrapolating or reusing incompatible source-paper numbers. The protocol-matched Acc-2 comparison against MissRAG, the most structurally related comparator, is reported in main-paper Tables 1 and 1 and in Table F.2 below.

The reported means are highest in every displayed column, including language-missing patterns such as ××○ and ×○○. These results show that the completed representation remains useful to the task head under finer label granularity. They do not by themselves attribute the gain to any module, which is what Table 3 is for, and they do not measure feature fidelity, which is what Table F.4 is for.

(a) CMU-MOSI, Acc-5							
Model	oxx	xox	xxo	oox	oxo	xoo	ooo
DMD	33.05	26.28	27.15	34.52	34.85	28.12	35.40
CorrKD	35.12	29.35	30.12	36.12	36.45	32.15	37.12
LNLN	35.85	30.12	30.85	37.15	37.52	33.25	38.25
MMIN	31.62	25.12	24.85	32.82	33.15	26.15	34.62
DiCMoR	33.15	28.25	27.15	34.25	34.52	30.12	35.80
IMDer	33.42	27.85	28.15	34.62	34.85	30.25	36.15
SMCMSA	32.55	26.12	26.12	33.85	34.12	29.15	35.05
P-RMF	<u>36.12</u>	<u>30.45</u>	<u>31.12</u>	<u>37.42</u>	<u>37.85</u>	<u>33.52</u>	<u>38.45</u>
HyDiff	**38.12**	**33.15**	**34.05**	**39.42**	**39.75**	**36.12**	**40.24**

(b) CMU-MOSEI, Acc-5							
Model	oxx	xox	xxo	oox	oxo	xoo	ooo
DMD	44.05	<u>39.52</u>	<u>40.55</u>	45.12	45.45	42.12	46.05
CorrKD	43.15	38.45	39.22	44.25	44.62	41.15	45.38
LNLN	<u>44.25</u>	<u>39.52</u>	40.25	<u>45.32</u>	<u>45.85</u>	<u>42.15</u>	<u>46.12</u>
MMIN	40.65	37.15	36.85	41.25	41.52	38.95	43.12
DiCMoR	42.15	39.25	38.15	43.15	43.52	40.12	44.15
IMDer	42.85	<u>39.52</u>	39.52	43.85	44.12	40.55	44.82
SMCMSA	42.12	38.65	38.45	43.15	43.42	39.85	44.88
P-RMF	44.02	39.25	40.12	45.12	45.65	41.85	45.85
HyDiff	**46.25**	**42.85**	**44.15**	**47.25**	**47.52**	**45.85**	**47.95**

Table E.3: Acc-5 (%) by realized Mod availability pattern at pre-rejection $\mu = 0.7$. Patterns follow [L,V,A]. Bold and underlining mark the best and second-best reported means among the listed methods.

E.4 Retrieval Query and Aggregation

The query-source block tests whether retrieval should be driven by the partially observed feature vector or by the structured interpretation anchor. The aggregation block keeps anchor queries fixed and varies how the retrieved neighbors are combined.

Anchor queries improve over corrupted-feature queries in every setting, showing that the retrieval neighborhood is determined more effectively by the hypothesized meaning of the missing modality than by a partial feature vector. Attention pooling remains close to softmax weighting, so the main effect is not tied to one aggregation rule.

E.5 Refinement Depth and Retrieval Hyperparameters

Refinement depth. Table E.6 reports refinement results under I-RA at $\eta = 0.9$ and under M-PM at pre-rejection $\mu = 0.7$ across all three benchmarks. At $L = 0$ the reasoner emits a single proposal and no verifier directive has yet been produced, so the score-only and score+feedback variants coincide by construction and share one reported value. All six score+feedback curves peak at $L = 3$. Four score-only curves also peak at $L = 3$, while I-RA MOSI and M-PM MOSEI peak at $L = 2$. Score+feedback exceeds score-only at every matched refinement limit $L \geq 1$, and the gains diminish beyond three rounds, supporting $L_{\max} = 3$ as the default accuracy–cost operating point. Table E.6 reports means only; the five-seed analysis in Table F.2 concerns final HyDiff-versus-baseline comparisons rather than individual depth contrasts.

Retrieval hyperparameters. Table E.7 reports one-factor-at-a-time sensitivity to the fixed retrieval hyperparameters. When sweeping K, we keep $\sigma = 0.1$ fixed; when sweeping σ, we keep $K = 10$ fixed. Consequently, the $K = 10$ and $\sigma = 0.1$ rows correspond to the same default fixed retrieval configuration and produce identical results.

E.6 Memory-Bank Coverage

Table E.9 reports Acc-2 as the training-only memory bank is uniformly subsampled. Performance improves monotonically with increasing memory coverage in all six dataset–protocol settings;

(a) CMU-MOSI, Acc-7

Model	oxx	xox	xxo	oox	oxo	xoo	ooo
DMD	28.45	21.35	22.05	29.35	29.62	22.45	30.70
CorrKD	31.12	25.42	26.15	32.15	32.45	27.25	33.88
LNLN	<u>31.85</u>	<u>26.15</u>	<u>26.85</u>	<u>32.85</u>	<u>33.15</u>	<u>27.92</u>	<u>34.31</u>
MMIN	29.15	22.45	21.95	30.12	30.45	24.15	31.65
DiCMoR	30.25	24.12	23.15	31.15	31.42	26.12	32.38
IMDer	30.85	24.15	24.85	31.65	32.05	26.85	32.95
SMCMSA	29.85	23.42	23.42	30.45	30.85	25.12	31.92
P-RMF	31.62	25.85	26.55	32.55	32.85	27.65	34.08
HyDiff	**33.15**	**28.55**	**29.25**	**34.25**	**34.62**	**30.85**	**35.75**

(b) CMU-MOSEI, Acc-7

Model	oxx	xox	xxo	oox	oxo	xoo	ooo
DMD	<u>43.05</u>	<u>38.45</u>	<u>39.52</u>	<u>44.12</u>	<u>44.35</u>	<u>40.85</u>	45.42
CorrKD	41.45	36.85	37.45	42.15	42.35	38.42	44.20
LNLN	42.45	37.52	38.15	43.25	43.52	39.12	<u>45.55</u>
MMIN	39.12	35.15	34.85	39.52	39.85	36.42	41.20
DiCMoR	40.12	36.15	35.15	41.15	41.35	38.42	42.15
IMDer	41.05	37.25	37.25	41.85	42.15	39.05	42.85
SMCMSA	41.12	38.15	38.15	42.12	42.45	39.12	43.12
P-RMF	41.85	37.05	37.65	42.45	42.85	38.75	44.58
HyDiff	**45.12**	**41.25**	**42.55**	**46.52**	**46.85**	**44.15**	**47.82**

(c) CH-SIMS, Acc-3

Model	oxx	xox	xxo	oox	oxo	xoo	ooo
DMD	54.15	46.25	47.25	55.05	55.42	48.25	<u>57.38</u>
CorrKD	54.15	48.45	49.35	55.25	55.65	50.42	56.85
LNLN	<u>55.12</u>	<u>49.45</u>	<u>50.25</u>	<u>56.12</u>	<u>56.45</u>	<u>51.45</u>	57.10
MMIN	51.25	43.05	42.15	52.25	52.42	45.15	54.08
DiCMoR	52.45	45.25	44.15	53.42	53.85	47.25	55.42
IMDer	53.05	46.12	46.12	54.12	54.35	48.05	56.15
SMCMSA	52.12	45.12	44.25	52.85	53.12	47.15	54.88
P-RMF	52.12	48.85	49.75	53.25	53.55	50.85	54.72
HyDiff	**56.42**	**52.85**	**53.55**	**57.65**	**58.02**	**55.12**	**59.28**

Table E.4: Acc-7 (%) for CMU-MOSI and CMU-MOSEI and Acc-3 (%) for CH-SIMS, grouped by realized Mod availability at pre-rejection $\mu = 0.7$. Bold and underlining mark the best and second-best reported means among the listed methods.

reducing coverage from 100% to 25% lowers performance by 3.55–9.28 points. The effect is strongest for CMU-MOSI under M-PM, where the full memory bank improves Acc-2 by 9.28 points over the 25% setting. CMU-MOSI and CH-SIMS are also more sensitive to coverage under M-PM than under I-RA, suggesting that whole-modality recovery places greater demands on representative retrieval support. This analysis measures dependence on memory coverage, while the no-gate row of Table 3 separately evaluates coverage-aware fallback.

E.7 Dual Injection, Loss Weight, and Anchor Projection

Dual injection. We compare semantic-only cross-attention, prior-only adaptive normalization, simple concatenation of the two signals into a single injection path, and the full dual pathway. The first block of Table E.10 reports lower task means for all three reduced variants, and the concatenation row is identical to the single-injection-path row of main-paper Table 3. Column order is I-RA then M-PM, matching the main ablation table. The result supports the utility of both conditioning paths for

Strategy	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	I-RA	M-PM	I-RA	M-PM	I-RA	M-PM
Query source						
Raw feature X^{obs}	64.52	72.18	71.35	73.82	66.42	70.55
Anchor $\bar{z}_s$	**73.68**	**78.43**	**79.55**	**79.40**	**74.62**	**76.58**
Neighbor aggregation with anchor query						
Top-1	71.20	76.85	76.85	78.12	71.45	74.32
Uniform	72.15	77.32	77.52	78.55	72.25	74.90
Attention pooling	73.42	78.20	79.25	79.15	74.40	76.45
Softmax weighting	**73.68**	**78.43**	**79.55**	**79.40**	**74.62**	**76.58**

Table E.5: Semantic-retrieval ablations (Acc-2, %). I-RA is raw-assisted feature corruption at $\eta = 0.9$; M-PM is pattern-macro modality missingness at pre-rejection $\mu = 0.7$.

Protocol	**Model / Refinement limit (L)**	$L=0$	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$
I-RA	Score-only (MOSI)	64.12	68.55	**70.85**	70.42	70.15	69.42	68.80
	Score+feedback (MOSI)	64.12	71.42	73.10	**73.68**	73.15	72.35	71.50
	Score-only (MOSEI)	67.45	72.80	74.95	**75.20**	74.85	74.12	73.55
	Score+feedback (MOSEI)	67.45	75.12	78.45	**79.55**	78.90	78.10	77.25
	Score-only (SIMS)	63.85	69.12	71.35	**71.60**	71.05	70.25	69.60
	Score+feedback (SIMS)	63.85	71.25	73.85	**74.62**	74.05	73.20	72.45
M-PM	Score-only (MOSI)	71.55	75.32	76.45	**76.80**	76.25	75.85	75.12
	Score+feedback (MOSI)	71.55	77.50	78.25	**78.43**	78.10	77.65	77.05
	Score-only (MOSEI)	72.85	76.10	**78.51**	78.25	77.95	77.40	76.85
	Score+feedback (MOSEI)	72.85	77.85	79.15	**79.40**	79.05	78.60	77.90
	Score-only (SIMS)	69.80	73.45	75.10	**75.35**	74.90	74.25	73.80
	Score+feedback (SIMS)	69.80	75.15	76.45	**76.58**	76.35	75.90	75.25

Table E.6: Refinement-depth sensitivity (Acc-2, %). I-RA uses raw-assisted feature corruption at $\eta = 0.9$ and M-PM uses the pattern-macro modality-level protocol at pre-rejection $\mu = 0.7$. The two variants share the $L = 0$ entry, since no feedback has been generated before the first verification round. Bold marks the highest reported mean in each row.

prediction, while the claim that one path carries semantics and the other carries structure is assessed separately by the completion measurements in Table F.4.

Diffusion-loss weight. We sweep λ_1 in $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{diff}} + \lambda_2 \mathcal{L}_{\text{cyc}}$ with λ_2 fixed at the implemented value. The $\lambda_1 = 0$ row removes only the denoising objective; the generator and task pipeline remain trainable through the task and cycle terms, so this setting is not a diffusion-free architecture. The second block of Table E.10 peaks in all six settings at $\lambda_1 = 0.1$. Removing denoising produces the largest degradation, whereas larger weights gradually reduce accuracy, indicating that the denoising objective is necessary but should remain subordinate to the task loss.

Anchor projection and dimensionality. The final two blocks of Table E.10 compare projection architectures and anchor dimensionality d_a. The gated projection improves every setting over a plain MLP. Increasing d_a from 32 to 128 yields substantial gains, whereas doubling the width to 256 changes the six scores by only $+0.04$ to $+0.06$ points, a 0.05-point average gain that is well within seed variation. We therefore select $d_a = 128$ as the lower-cost point on the observed plateau; bold marks the numerical maximum in each column and underlining marks the selected default.

F Additional Confirmatory Analyses

This section consolidates confirmatory diagnostics that do not fit in the main narrative, including five-seed variation, the energy-exploitation test, and feature-space completion fidelity. All repeated runs use seeds $\{13, 17, 23, 41, 89\}$ and reuse the same evaluation masks for the compared methods within each seed.

Variant / Value	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	I-RA	M-PM	I-RA	M-PM	I-RA	M-PM
Number of neighbors K (adaptivity disabled)						
$K = 1$	66.15	70.02	77.25	77.20	68.45	70.12
$K = 5$	71.92	76.45	78.48	78.42	72.85	73.85
$K = 10$	**72.61**	**76.79**	**78.83**	**78.72**	**73.53**	**74.65**
$K = 20$	71.40	75.80	78.65	78.51	72.95	73.35
Temperature coefficient σ (adaptivity disabled)						
$\sigma = 0.05$	68.45	71.40	78.10	77.85	71.12	72.55
$\sigma = 0.1$	**72.61**	**76.79**	**78.83**	**78.72**	**73.53**	**74.65**
$\sigma = 0.2$	71.15	72.15	78.62	78.60	73.05	73.12
$\sigma = 0.5$	69.82	71.05	77.92	77.95	72.35	72.45

Table E.7: Retrieval-hyperparameter sensitivity (Acc-2, %). I-RA uses raw-assisted feature corruption at $\eta = 0.9$ and M-PM uses the pattern-macro modality-level protocol at pre-rejection $\mu = 0.7$. When sweeping K, $\sigma = 0.1$ is held fixed; when sweeping σ, $K = 10$ is held fixed. Bold marks the highest reported mean in each block.

Variant	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	I-RA	M-PM	I-RA	M-PM	I-RA	M-PM
Base, direct completion	61.23	66.17	73.14	72.81	64.86	67.92
R1, candidate set and weighting						
Single-pass anchor, $J = 1$	70.18	71.94	78.12	76.63	73.27	72.65
Argmax instead of mixture	69.60	74.33	75.09	75.93	70.95	73.17
Uncalibrated weights, $T = 1$	68.63	75.25	75.46	75.25	70.37	73.82
R2, grounding						
No retrieval, $z_p = \mathbf{0}$	68.49	75.46	76.88	78.57	71.53	75.19
No coverage gate	70.50	74.84	74.57	76.55	70.29	73.54
No ambiguity-adaptive breadth ($u = 0$)	71.95	76.47	78.11	77.82	72.86	74.93
Class-mean prior	67.83	74.44	75.64	77.11	70.99	74.47
R3, candidate-consistency surrogate						
Single injection path	72.31	76.68	79.04	78.72	73.66	75.74
No cycle training loss	69.72	74.91	75.22	75.34	71.50	74.48
No inference-time guidance	70.41	74.12	76.18	76.79	71.25	74.40
Raw energy without hinge	69.37	74.13	75.18	75.48	71.27	74.34
HyDiff (Ours)	**73.68**	**78.43**	**79.55**	**79.40**	**74.62**	**76.58**

Table E.8: Full per-setting requirement ablations corresponding to the six-setting means in main-paper Table 3.

F.1 Energy-Exploitation Diagnostic

The reference-relative hinge is intended to limit unconstrained optimization of the frozen verifier; it is not a formal guarantee against off-manifold score exploitation. The raw-energy and hinged models use the same test split, stochastic Mod masks, reasoner JSONL cache, candidate identities and ordering, temperature scaling and pruning, frozen verifier E_ω, frozen bridge Φ_{cross}, retrieval memory, DDIM initial noise, and seeds $\{13, 17, 23, 41, 89\}$. The only changed objective is

$$\mathcal{L}_{\text{cyc}}^{\text{raw}} = \text{mean}(E_{\text{cyc}}), \qquad \mathcal{L}_{\text{cyc}}^{\text{hinge}} = \left[E_{\text{cyc}} - E_{\text{real}} + m\right]_+. \tag{F.1}$$

For compact notation, define the candidate-specific scoring map

$$\mathcal{E}_i^{(j)}(x) = E_\omega\left(\bar{z}_{s,i}^{(j)}, \Phi_{\text{cross}}(x, e_i^{\text{obs}})\right). \tag{F.2}$$

For sample i, retained candidate j, missing modality m, and check iteration $s \in \mathcal{S}_{\text{check}}$, generated and real energies are

$$E_{i,j,m}^{(s,v)} = \mathcal{E}_i^{(j)}\left(\hat{z}_{0,i,j,m}^{(s,v)}\right), \quad v \in \{\text{raw}, \text{hinge}\}, \tag{F.3}$$

$$E_{i,j,m}^{(\text{real})} = \mathcal{E}_i^{(j)}\left(z_{0,i,m}^{\text{real}}\right). \tag{F.4}$$

Coverage	MOSI		MOSEI		SIMS	
	I-RA	M-PM	I-RA	M-PM	I-RA	M-PM
25%	67.25	69.15	75.95	75.85	70.42	71.12
50%	70.15	72.26	77.82	77.62	72.15	72.65
75%	72.45	75.27	78.95	78.75	73.85	74.63
100%	**73.68**	**78.43**	**79.55**	**79.40**	**74.62**	**76.58**

Table E.9: Memory-bank coverage (Acc-2, %). I-RA and M-PM denote raw-assisted feature corruption at $\eta = 0.9$ and pattern-macro modality missingness at pre-rejection $\mu = 0.7$, respectively. The bank contains training prototypes only.

Variant	MOSI		MOSEI		SIMS	
	I-RA	M-PM	I-RA	M-PM	I-RA	M-PM
Dual-conditioning pathway						
Semantic-only (z_s)	71.35	75.86	78.37	78.42	73.12	75.55
Prior-only (z_p)	68.42	73.55	76.55	76.82	70.85	73.12
Concatenation (single injection path)	72.31	76.68	79.04	78.72	73.66	75.74
HyDiff	**73.68**	**78.43**	**79.55**	**79.40**	**74.62**	**76.58**
Denoising-loss weight λ_1 ($\lambda_2 = 0.05$)						
0 (no denoising loss)	69.78	74.08	75.33	75.25	69.78	71.59
0.01	71.54	76.73	78.07	77.64	72.41	75.34
0.1	**73.68**	**78.43**	**79.55**	**79.40**	**74.62**	**76.58**
0.5	73.01	77.98	79.09	79.10	74.49	76.07
1.0	72.48	77.42	78.10	78.27	73.58	75.22
Projection architecture						
MLP	72.15	76.82	78.45	78.22	73.12	75.05
Gated MLP	**73.68**	**78.43**	**79.55**	**79.40**	**74.62**	**76.58**
Anchor dimensionality d_a						
32	69.45	74.12	75.82	75.55	70.35	72.18
64	71.82	76.95	77.95	77.62	72.85	74.92
128	73.68	78.43	79.55	79.40	74.62	76.58
256	**73.72**	**78.48**	**79.60**	**79.45**	**74.68**	**76.64**

Table E.10: Conditioning, loss-weight, and anchor-projection ablations (Acc-2, %). I-RA and M-PM denote raw-assisted feature corruption at $\eta = 0.9$ and pattern-macro modality missingness at pre-rejection $\mu = 0.7$, respectively. The single-injection row reproduces the corresponding row of main-paper Table 3. In the dimensionality block, bold marks the numerical maximum and underlining marks the selected 128-dimensional default.

The real completion is scored separately under the same candidate j; it is not replaced by a candidate-independent sample score. Generated energies are first averaged over the five check iterations,

$$\bar{E}^{(v)}_{i,j,m} = \frac{1}{|\mathcal{S}_{\text{check}}|} \sum_{s \in \mathcal{S}_{\text{check}}} E^{(s,v)}_{i,j,m}, \qquad \mathcal{S}_{\text{check}} = \{30, 35, 40, 45, 50\}. \tag{F.5}$$

while $\bar{E}^{(\text{real})}_{i,j,m} = E^{(\text{real})}_{i,j,m}$. This checkpoint average matches the locations at which cycle supervision is applied rather than scoring only the final DDIM output.

Candidates are aggregated with the same stop-gradient candidate weights used by training and prediction,

$$\bar{E}^{(v)}_{i,m} = \sum_{j=1}^{J_i} \text{sg}(w_{i,j}) \bar{E}^{(v)}_{i,j,m}, \qquad \sum_{j=1}^{J_i} w_{i,j} = 1. \tag{F.6}$$

Dataset	E_{real}	E_{raw}	E_{hinge}
CMU-MOSI	1.314 ± 0.021	1.188 ± 0.028	$\mathbf{1.311 \pm 0.019}$
CMU-MOSEI	1.281 ± 0.018	1.146 ± 0.025	$\mathbf{1.269 \pm 0.022}$
CH-SIMS	1.409 ± 0.024	1.249 ± 0.031	$\mathbf{1.400 \pm 0.027}$

	I-RA fidelity		**M-PM task accuracy**	
Dataset	**NRMSE raw**	**NRMSE hinge**	**Acc-2 raw**	**Acc-2 hinge**
CMU-MOSI	0.702	**0.487**	74.13	**78.43**
CMU-MOSEI	0.658	**0.452**	75.48	**79.40**
CH-SIMS	0.721	**0.521**	74.34	**76.58**

Table F.1: Energy-exploitation diagnostic. The upper panel reports mean verifier energy $\pm$ SD across five seeds after missing-pattern macro-averaging. E_{real} uses the held-out real completion under the same candidate, while E_{raw} and E_{hinge} use generators trained with the unconstrained and reference-relative objectives. The lower panel reports NRMSE on held-out I-RA entries at $\eta = 0.9$ and Acc-2 under M-PM at pre-rejection $\mu = 0.7$.

If sample i has multiple missing modalities, they are averaged within the sample so that every test example contributes one value,

$$\bar{E}_i^{(v)} = \frac{1}{|\mathcal{M}_i^{\text{miss}}|} \sum_{m \in \mathcal{M}_i^{\text{miss}}} \bar{E}_{i,m}^{(v)}. \tag{F.7}$$

Energy excludes ○○○, which requires no completion, and xxx, which is rejected by the sampler. For the six incomplete patterns $\mathcal{R}_{\text{miss}} = \{\text{○xx}, \text{x○x}, \text{xx○}, \text{○○x}, \text{○x○}, \text{x○○}\}$, we compute a pattern mean and then an equal-weight missing-pattern macro-average for each seed q,

$$\bar{E}_{q,r}^{(v)} = \frac{1}{N_{q,r}} \sum_{i \in \mathcal{I}_{q,r}} \bar{E}_i^{(v)}, \tag{F.8}$$

$$\bar{E}_q^{(v)} = \frac{1}{|\mathcal{R}_{\text{miss}}|} \sum_{r \in \mathcal{R}_{\text{miss}}} \bar{E}_{q,r}^{(v)}. \tag{F.9}$$

This quantity is termed *missing-pattern-macro energy*; unlike Mod-PM Acc-2, it averages six rather than seven patterns. Finally,

$$\mu_E^{(v)} = \frac{1}{5} \sum_{q=1}^{5} \bar{E}_q^{(v)}, \qquad \sigma_E^{(v)} = \sqrt{\frac{1}{4} \sum_{q=1}^{5} \left(\bar{E}_q^{(v)} - \mu_E^{(v)} \right)^2}. \tag{F.10}$$

Because each dataset has its own verifier, absolute energy scales are not averaged across datasets. Within each dataset, both generated-energy variants are compared with the held-out real-completion reference E_{real}. The quantity $E_{\text{real}} - m$ is the train-time zero-loss boundary of the hinge in Eq. F.1; it is not an expected value that the held-out mean E_{hinge} should numerically match.

The diagnostic should be read jointly with the "Raw energy without hinge" row of Table 3. A systematic energy shift substantially below the held-out real-data reference is consistent with direct optimization of the frozen verifier. The hinged model is instead evaluated by whether it avoids such a downward shift while preserving feature fidelity and task performance. This test targets the specific failure mode of optimizing a frozen score; it is not a standalone proof of perceptual realism, semantic diversity, or candidate coverage.

Table F.1 shows raw-objective energies 0.126–0.160 below the corresponding real-completion references, while the hinged means remain within 0.003–0.012 of them. This proximity should not be interpreted as the hinge converging to its train-time boundary $E_{\text{real}} - m$; rather, the held-out diagnostic tests whether the learned generator avoids the systematic downward energy shift produced by unconstrained score minimization. Across-seed SDs range from 0.018 to 0.031. Relative to the raw objective, the hinge reduces NRMSE by 0.215, 0.206, and 0.200 and raises M-PM Acc-2 by 4.30, 3.92, and 2.24 points on CMU-MOSI, CMU-MOSEI, and CH-SIMS, respectively. This joint pattern is consistent with the intended guard against direct energy minimization, but it does not prove that every form of verifier exploitation is absent; the result is therefore treated as a bounded diagnostic.

F.2 Five-Seed Variation and Paired Tests

For each dataset–protocol comparison, a two-sided paired randomization test independently swaps the two methods' correctness outcomes within each matched test-item–seed pair and recomputes the mean Acc-2 difference over $10{,}000$ permutations. The nine resulting p values are adjusted with the Holm procedure, controlling the family-wise error rate at $\alpha = 0.05$.

Two feature-corruption comparisons must be distinguished, following Appendix C.4. The raw-assisted comparison places P-RMF against HyDiff-RA, which additionally observes the available raw streams; this is a diagnostic rather than an input-matched test. The strictly matched comparison places P-RMF against HyDiff-Feat, which receives exactly the same masked tensors and availability masks. Table F.2 reports both and labels each explicitly, so that the significance statement attached to the matched-input claim is the one computed on the matched pair.

Dataset	Pair	Comp.	HyDiff	Gain	Holm p
I-RA, $\eta = 0.9$, strictly matched input					
MOSI	P-RMF vs. HyDiff-Feat	72.35 ± 0.58	73.24	$+0.89$	**0.036**
MOSEI	P-RMF vs. HyDiff-Feat	78.28 ± 0.61	79.02	$+0.74$	**0.041**
CH-SIMS	P-RMF vs. HyDiff-Feat	73.41 ± 0.50	74.05	$+0.64$	**0.048**
I-RA, $\eta = 0.9$, raw-assisted reference					
MOSI	P-RMF vs. HyDiff-RA	72.35 ± 0.58	73.68 ± 0.46	$+1.33$	**0.018**
MOSEI	P-RMF vs. HyDiff-RA	78.28 ± 0.61	79.55 ± 0.68	$+1.27$	**0.024**
CH-SIMS	P-RMF vs. HyDiff-RA	73.41 ± 0.50	74.62 ± 0.43	$+1.21$	**0.031**
M-PM, pre-rejection $\mu = 0.7$					
MOSI	MissRAG vs. HyDiff	77.07 ± 0.47	78.43 ± 0.47	$+1.36$	**0.008**
MOSEI	MissRAG vs. HyDiff	77.93 ± 0.51	79.40 ± 0.41	$+1.47$	**0.005**
CH-SIMS	MissRAG vs. HyDiff	75.28 ± 0.48	76.58 ± 0.61	$+1.30$	**0.004**

Table F.2: Acc-2 variation (%). Rows with reported seed variation show mean $\pm$ SD using identical seeds and masks. The matched-input rows show the supplied means for P-RMF versus HyDiff-Feat; their gain and Holm-corrected p values are reported without inventing unprovided SDs. Holm p is adjusted across all nine dataset–protocol comparisons.

HyDiff has a positive mean gap in all nine settings. The six repeated-run comparisons have standard deviations below 0.7 points for both methods, and all nine paired differences remain significant after Holm correction ($p_{\mathrm{Holm}} \leq 0.048$), controlling the family-wise error rate at 0.05.

The reduced-anchor comparison is treated as a separate six-test family because it addresses a distinct semantic-shortcut hypothesis. Table F.3 reports the five-seed variation and raw paired-randomization p values. Standard Holm correction across these six raw p values gives adjusted values $0.055, 0.055, 0.036, 0.055, 0.055, 0.055$ in table order; only MOSEI I-RA remains below 0.05 after family-wise correction. The direction is nevertheless stable: all six mean gains are positive and HyDiff wins 26 of the 30 paired seed runs.

Dataset	Protocol	P-RMF reduced	HyDiff reduced	Gain	Pos. seeds	Raw p / Holm p
MOSI	I-RA	71.12 ± 0.38	71.68 ± 0.44	$+0.56$	4/5	0.022 / 0.055
MOSI	M-PM	76.24 ± 0.40	76.92 ± 0.57	$+0.68$	4/5	0.014 / 0.055
MOSEI	I-RA	77.35 ± 0.29	78.14 ± 0.33	$+0.79$	5/5	**0.006 / 0.036**
MOSEI	M-PM	77.58 ± 0.29	78.03 ± 0.36	$+0.45$	5/5	0.011 / 0.055
CH-SIMS	I-RA	72.31 ± 0.37	72.86 ± 0.46	$+0.55$	4/5	0.025 / 0.055
CH-SIMS	M-PM	74.87 ± 0.36	75.57 ± 0.68	$+0.70$	4/5	0.013 / 0.055

Table F.3: Matched reduced-anchor control over five seeds. The reduced schema removes `polarity`, `intensity`, and `dominant_affect`. "Pos. seeds" counts seeds for which HyDiff exceeds P-RMF. Raw paired-randomization tests use the same matched-item–seed procedure as Table F.2; Holm adjustment is applied across these six comparisons as a separate family.

F.3 Feature-Space Completion Fidelity

Let $\Omega_{s,m}$ denote the held-out entries of modality m for sample s, and let $\sigma^{\mathrm{tr}}_{m,k}$ be the training-set standard deviation of feature dimension k. We evaluate only the entries hidden by the Intra mask.

Dataset	Method	NRMSE ↓	Cosine similarity ↑
CMU-MOSI	P-RMF	0.612	0.712
	HyDiff	**0.487**	**0.821**
CMU-MOSEI	P-RMF	0.578	0.734
	HyDiff	**0.452**	**0.842**
CH-SIMS	P-RMF	0.645	0.695
	HyDiff	**0.521**	**0.798**

Table F.4: Feature-space completion fidelity under I-RA corruption at $\eta = 0.9$, evaluated only on held-out entries before task fusion.

The normalized root mean-square error is

$$\mathrm{NRMSE} = \left[\frac{\sum_{s,m} \sum_{k \in \Omega_{s,m}} \left(\frac{\hat{x}_{s,m,k} - x_{s,m,k}}{\sigma^{\mathrm{tr}}_{m,k} + 10^{-8}} \right)^2}{\sum_{s,m} |\Omega_{s,m}|} \right]^{1/2}. \tag{F.11}$$

and cosine similarity is computed per sample between the vectorized ground-truth and recovered held-out entries, then averaged over samples and missing modalities. For HyDiff, $\hat{x}$ is the candidate-weighted mean completion $\sum_j w_j \hat{x}^{(j)}$ before task fusion. These feature-space metrics probe completion fidelity; they are not direct estimates of the total-variation terms in Eq. A.8.

The candidate-weighted mean completion reduces NRMSE by 0.125, 0.126, and 0.124 on CMU-MOSI, CMU-MOSEI, and CH-SIMS, respectively, while increasing cosine similarity by 0.109, 0.108, and 0.103. These results show that the aggregate completion is useful under pointwise feature-fidelity metrics. Because squared-error metrics favor conditional means, they do not establish that individual branches are realistic, diverse, or cover distinct semantic modes.

G Task Transfer and Efficiency

G.1 Task Transfer to MM-IMDb

MM-IMDb (Arevalo et al., 2017) tests whether the interpretation–retrieval–completion pipeline can be transferred to a non-sentiment, multi-label task. Plot text is encoded by a frozen BERT-base-uncased encoder using the final-layer `[CLS]` representation, and poster images are encoded by a frozen ImageNet-pretrained ResNet-50 (He et al., 2016) using the global-average-pooled feature. The 768-dimensional text feature and 2048-dimensional image feature are independently projected to 128 dimensions. The reasoner receives only the observed raw plot and/or poster together with the task description and produces

$$\{\texttt{candidate_genres}, \texttt{plot_cues}, \texttt{visual_cues}, \texttt{cross_modal_evidence}, \texttt{uncertainty_notes}\}.$$

Here, `candidate_genres` denotes reasoner-generated task-aligned genre candidates inferred from the observed evidence; ground-truth validation/test genre annotations are not supplied to the reasoner. These candidates are used as semantic conditioning for recovery rather than as the task prediction itself. Because their vocabulary is aligned with the downstream genre task, this setting does not establish label-agnostic transfer and should instead be read as a task-transfer stress test under task-aligned semantic conditioning. Retrieval memory contains training samples only. For the missing text or image target, the diffusion module is conditioned on the semantic anchor and the retrieved target-modality prior; the observed projected feature influences these conditions upstream and remains available to the task head, but is not supplied directly to the denoiser. A two-layer task head with sigmoid outputs is trained with binary cross-entropy over the genre labels. Micro-F1 and macro-F1 use a fixed threshold of 0.5 for every label.

Text and image are dropped independently with pre-rejection probability $\mu = 0.7$, and an all-missing `xx` draw is rejected and redrawn. Consequently, `oo` occurs with probability 17.65%, and each one-observed pattern occurs with probability 41.18%. P-RMF uses the same frozen features, 128-dimensional projections, task head, missingness masks, train/validation/test split, and optimization budget; only the recovery mechanism differs.

HyDiff improves micro-F1 by 3.43 points and macro-F1 by 3.77 points over the matched P-RMF adaptation. The larger macro-F1 gain is consistent with improved performance on less frequent labels, although per-class analysis would be required to attribute that gain specifically to interpretation

Method	Micro-F1 ↑	Macro-F1 ↑
P-RMF	56.42	41.15
HyDiff	**59.85**	**44.92**
Gain	**+3.43**	**+3.77**

Table G.1: MM-IMDb task transfer under stochastic modality missingness at pre-rejection $\mu = 0.7$; the all-missing draw is rejected and redrawn. The HyDiff reasoner uses task-aligned genre candidates inferred from the observed modalities, not ground-truth test annotations.

diversity. Because the semantic schema contains task-aligned genre candidates, this single-dataset result supports portability of the recovery architecture when paired with task-specific semantic conditioning; it is not evidence that the method transfers without task semantics or that the reasoner is label-agnostic.

G.2 Candidate-Set Operating Points

Table G.2 compares the single-anchor R1 control with the default candidate set. Averaged over the three M-PM settings, retaining the temperature-scaled candidate set raises Acc-2 from 73.74 to 78.14, a 4.40-point gain. However, the stronger repeated-top-1 control already reaches 77.06 (Table C.3), leaving a 1.08-point residual gain for distinct weighted candidate branches. We therefore treat the latter as a modest architecture-specific gain rather than attributing the full 4.40 points to candidate diversity. The $J = 1$ latency is measured end to end. For the default mean $J = 2.7$, we additionally report a transparent serial-cost estimate obtained from the measured stage means; it is an engineering estimate rather than an independent benchmark.

Configuration	Acc-2	Latency
Single anchor ($J = 1.0$)	73.74	689 ± 15 ms
Default candidate set ($J = 2.7$)	78.14	1116 ms (est.)

Table G.2: Candidate-set operating points under Mod at pre-rejection $\mu = 0.7$. Acc-2 is the M-PM value averaged over the three sentiment datasets.

G.3 Stage-wise Inference Efficiency

Under the measured $J = 1$ diagnostic, HyDiff takes 689 ± 15 ms per sample, compared with 431 ± 9 ms for P-RMF and 742 ± 17 ms for DiCMoR. Thus, the $J = 1$ configuration is slower than P-RMF and faster than DiCMoR; the default $J = 2.7$ serial estimate is about 1.12 s and should not be described as faster than either baseline. The 438 ms reasoner entry includes one complete SRAR procedure with up to three deterministic refinement rounds. Retrieval contributes 48 ms, while each retained candidate adds retrieval and recovery work. Reducing DDIM from 50 to 10 steps lowers the measured $J = 1$ total from 689 ± 15 to 610 ± 14 ms (11.5%), but we report this only as a latency reduction because the corresponding task accuracy is not included.

G.4 Latency Measurement Protocol

All methods are timed on the same NVIDIA A100 with batch size 1, 10 warm-up runs, 100 measured runs, and explicit GPU synchronization before and after each timed region. Timing excludes offline BERT, FACET, and COVAREP feature extraction, raw-stream decoding, and disk input. It includes the complete SRAR procedure over already loaded inputs, query projection, top-K search, prototype aggregation, diffusion recovery, scheduled cycle guidance, and the task head. The reported means are arithmetic averages over the 100 runs, and the standard deviations describe run-to-run latency variation. HyDiff candidates are executed serially in this diagnostic; consequently, the default-J value in Table G.2 is obtained by linear composition, $438 + 2.7 \times (48 + 203) \approx 1116$ ms, and is not presented as a measured wall-clock result. The 10-step variant changes only the DDIM sampling length, so no accuracy–latency trade-off is inferred from its latency alone. Anchor-conditioned baseline timings include the same SRAR procedure used by the corresponding anchor-access controls in Table C.1.

Method	Component	Latency (ms)
P-RMF	Total	431 ± 9
DiCMoR	Total	742 ± 17
HyDiff, 50-step	Reasoner	438 ± 12
	Retrieval	48 ± 3
	Recovery	203 ± 7
	Total	689 ± 15
HyDiff, 10-step	Reasoner	438 ± 12
	Retrieval	48 ± 3
	Recovery	124 ± 5
	Total	610 ± 14

Table G.3: Stage-wise latency on one NVIDIA A100, reported over 100 synchronized single-sample runs after 10 warm-up runs. HyDiff uses $J = 1$.