

Plan the Semantics, Then Render the Signal: A Hybrid State-Space Model for Incomplete Multimodal Learning

Ruiting Dai, Zhihao Qin, Yilin Chen, Xianhua Yan, Lisi Mo, Tao He *

University of Electronic Science and Technology of China

{rtdai,morris}@uestc.edu.cn, {zhihao.qin,xianhua.yan}@std.uestc.edu.cn
chenyilin2005@gmail.com, tao.he01@hotmail.com

Abstract

Incomplete multimodal learning seeks robust perception when modalities are absent at training or test time, using available streams to recover the missing ones. Most methods cast this as direct regression in a fixed continuous latent, which (i) collapses many valid solutions into an averaged reconstruction when observations underdetermine the target, and (ii) suffers semantic drift when cross-modal relationships shift with context. We argue that completion should first resolve *what* is missing and only then render *how* it appears. We realize this as **Semantic-first Completion**, instantiated by a **Hybrid State-Space Model (HySSM)** that decouples semantic intent from signal realization. A frozen omni-modal LLM plans over observation-only evidence, generating an evidence-grounded Markovian trajectory of discrete semantic states that act as a compact information bottleneck for stable cross-modal reasoning. The plan is projected into a coordinate-wise control field that drift-guides a conditional diffusion model, synthesizing missing signals or feature streams as diverse and semantically consistent completions. Across MNIST, Hateful Memes, and CMU-MOSEI, HySSM surpasses recent generative baselines and, under agnostic masking shift, remains robust up to 90% missingness.

1 Introduction

Multimodal learning (Baltrušaitis et al., 2018) builds unified representations by jointly modeling complementary streams such as vision, audio, and language, improving perception and reasoning. In the real world, however, modalities are frequently incomplete: sensors fail, streams desynchronize, and privacy constraints withhold data, leaving fragmented observations whose missing patterns vary across time and deployments. Incomplete multimodal learning (Ma et al., 2021) targets this regime

by reconstructing the missing modalities from the available ones, enabling robust perception under partial inputs (Zhang et al., 2025a).

Existing completion methods fall into two paradigms. Discriminative approaches, such as matrix completion (Thung et al., 2018) or regression-style predictors (Xu et al., 2025), are efficient but produce a single point estimate, so they cannot represent ambiguity when the missing modality is not uniquely determined (Ma et al., 2021; Baltrušaitis et al., 2018; Zhao et al., 2021). Generative approaches instead model conditional distributions: VAEs tend to over-smooth (Wu and Goodman, 2018), GANs can be unstable and mode collapse (Li et al., 2016), while diffusion models deliver high-fidelity, flexibly conditioned generation (Ho et al., 2020; Zhang et al., 2023; Dai et al., 2026b). Yet across both paradigms, completion is still cast as a *static mapping* from observed features to missing targets inside a fixed continuous latent.

We argue that this static view is misaligned with natural missingness for two reasons. **(1) Semantics are better determined than signals.** Observations constrain high-level meaning (who and what) far more reliably than the low-level signal (exact pixels, fine-grained timing): text may fix an event yet leave many plausible appearances, and a partial video may fix the action yet leave the audio underdetermined. Mapping such inputs directly into signal space forces a one-to-many problem into a single prediction, encouraging averaged realizations that surface as blur or attenuated semantic evidence. **(2) Cross-modal dependencies are time-varying.** The relationships for completion are not fixed: audio-visual correspondence shifts with speaker turns, scene changes, and action phases (Dong et al., 2026). Such instability is also reflected in anchor drift under incomplete observations (Dai et al., 2026a). A static fusion rule cannot track these evolving alignments, so when the observed subset changes across samples or timesteps

*Corresponding author.

it amplifies mismatch and produces semantically inconsistent completions.

These observations suggest that completion should not begin with signal synthesis. It should first resolve *what* is happening and only then render *how* it manifests in the missing modality. We turn this principle into **Semantic-first Completion (SFC)**, which reframes recovery as a two-stage generative process: semantic planning (Huang et al., 2025; Zhu et al., 2017) followed by modality completion. SFC first infers a semantic state from the observed modalities and then generates the missing signals conditioned on it. By decoupling semantic inference from signal generation through an explicit bottleneck, it reduces ambiguity in under-determined settings and keeps completions semantically aligned even as missing patterns shift.

We instantiate SFC as the **Hybrid State-Space Model (HySSM)**, which couples a discrete symbolic space with a continuous signal space through three modules. *Markovian Semantic State Transition* (MST) uses a frozen omni-modal LLM to plan: the observed modalities are turned into observation-only evidence, namely verbalized descriptors and projected soft tokens, from which the LLM samples evidence-grounded trajectories of discrete semantic states. Here discrete tokens act as an information bottleneck that suppresses low-level noise while stabilizing semantic content, and a grounding score together with a subset-consistency constraint keeps the plan faithful to the evidence and invariant to which modalities are observed. *Cross-Space Manifold Projection* (CMP) then densifies the plan into a continuous control state and injects it via cross-attention, yielding a coordinate-wise semantic field that specifies where and when each concept applies rather than conditioning the generator globally. Finally, *Drift-Guided Stochastic Recovery* (DSR) treats this field as a semantic drift in a conditional latent diffusion process, steering a reverse-time SDE so that, from different noise seeds, it synthesizes diverse yet semantically consistent, temporally coherent completions.

Across MNIST, Hateful Memes, and CMU-MOSEI, HySSM consistently surpasses strong discriminative and generative baselines under diverse missing patterns. It remains robust under agnostic masking shift, where training uses a single missing rate and evaluation spans the full range, and degrades gracefully up to 90% missingness, with further gains from a randomized masking strategy. **Contributions.** We make three contributions.

(1) we propose **SFC**, a two-stage paradigm that separates semantic planning from modality completion, addressing the one-to-many ambiguity and the time-varying dependencies that defeat static-mapping completion; (2) we instantiate it as **HySSM**, which links discrete semantic trajectories to diffusion-based synthesis through an information-bottleneck design, including an explicit interface that lets a frozen LLM plan over observation-only multimodal evidence; and (3) we develop **cross-space manifold projection** for localized, fine-grained control and **drift-guided diffusion** for diverse, high-fidelity recovery, and show consistent gains and strong robustness to extreme and shifted missingness across three benchmarks.

2 Related Work

Incomplete multimodal learning. Real-world multimodal datasets are often partially observed (e.g., sensor failures, cost, privacy), so models must handle arbitrary modality subsets (Wu et al., 2024a; Ma et al., 2021; Wu et al., 2024b). Prior work mainly (i) learns subset-robust representations via subset-invariant fusion, shared/specific factorization, prompt-based adaptation (Dai et al., 2025b), or hierarchical prompting and distillation (Mo et al., 2026), and (ii) reconstructs missing modalities. Recent single-model-for-all-subsets designs improve scalability using shared-specific modeling (Wang et al., 2023a), modality dropout (Alfasly et al., 2022) and modality-masked transformers (Bachmann et al., 2022; Guo et al., 2024). Yet, implicit handling of missingness can underestimate uncertainty when multiple modalities are absent (Zeng et al., 2022; Sun et al., 2024).

Generative multimodal completion. Completion methods explicitly model $p(\text{missing} \mid \text{observed})$ and generate diverse samples. VAEs/GANs enable cross-modal synthesis but can be blurry or unstable (Wu and Goodman, 2018; Li et al., 2016; Kebaili et al., 2025). Diffusion models are increasingly used for severe missingness due to flexible conditioning and strong sample quality (Ho et al., 2020; Zhang et al., 2025a), with masked conditioning and unified training across missing patterns (Kebaili et al., 2025; Meng et al., 2024). Recent variants add modality-aware encoding/fusion (Zhong et al., 2025; Zhang et al., 2025b), bidirectional recovery to reduce bias (Dai et al., 2025a), retrieval-augmented diffusion for robust incomplete multimodal learning (Dai et al., 2025c), and modality-

masked diffusion for random subsets (Meng et al., 2024). However, most approaches still learn a largely static feature-to-target mapping, which can break when cross-modal relations shift with semantic context, yielding over-smoothed or mismatched completions.

3 Method

3.1 Motivation and Overview

The static-mapping trap. Most completion methods learn a fixed map from observed features to missing targets in a single continuous latent. Under natural missingness this is a poor fit for two reasons. (i) *Semantics are better determined than signals.* Observations usually pin down *what* is happening far more reliably than *how* it is realized: a silent clip of a breaking wave implies loudness but not pitch; a partial video fixes the action but leaves the audio underdetermined. Regressing directly into signal space collapses these many valid realizations into one averaged output, e.g., blur, or attenuated semantic evidence. (ii) *Cross-modal dependencies drift.* Audio–visual correspondence shifts with speaker turns and scene changes, so a fixed fusion rule amplifies mismatch as the observed subset varies across samples and timesteps.

Our principle: resolve *what*, then render *how*. We therefore decouple completion into *semantic planning* followed by *signal rendering*, and instantiate it as the Hybrid State-Space Model (HySSM). HySSM evolves a coupled latent state across two manifolds: a *symbolic* manifold $\mathcal{M}_{\text{sym}}$, where a frozen omni-modal LLM reasons over discrete tokens, and a *signal* manifold $\mathcal{M}_{\text{sig}}$, where conditional diffusion synthesizes the target signal or feature representation. Discrete tokens act as an *information bottleneck*: they retain compositional, causal structure while discarding incidental signal-level variability, which stabilizes reasoning under partial supervision. Given observations $X^{\mathcal{O}}$, the process reads

$$\begin{aligned} \mathbf{e} &= \mathcal{E}_{\text{enc}}(X^{\mathcal{O}}), \quad s_0^{\text{sym}} = \text{Proj}(\mathbf{e}), \\ s_k^{\text{sym}} &\sim p_{\phi}(s_k^{\text{sym}} | s_{k-1}^{\text{sym}}, \mathcal{I}, \mathbf{e}), \quad k = 1, \dots, K, \\ H_{\text{ctl}} &= \text{Enc}_{\psi}(s_K^{\text{sym}}) \in \mathbb{R}^{L \times d}, \\ \hat{x}^{(m)} &\sim p_{\theta_m}(x^{(m)} | H_{\text{ctl}}), \quad \forall m \in \mathcal{T}_i. \end{aligned}$$

As illustrated in Fig. 1, three modules realize this pipeline and hand off cleanly to one another: (1) Markovian Semantic State Transition (MST, section 3.2) plans *what* to generate; (2) Cross-Space

Manifold Projection (CMP, section 3.3) bridges the discrete plan to the continuous generator and decides *where* each concept applies; and (3) Drift-Guided Stochastic Recovery (DSR, section 3.4) renders the signal stochastically, preserving diversity around the inferred semantics.

Problem setup. Let $\mathcal{D} = \{X_i\}$ be defined over a modality universe $\mathcal{M} = \{\text{img}, \text{aud}, \text{txt}\}$ (the formulation is agnostic to $|\mathcal{M}|$). Each sample exposes an *available* subset $\mathcal{A}_i \subseteq \mathcal{M}$; modalities in $\mathcal{M} \setminus \mathcal{A}_i$ are intrinsically unobserved and never supervised. At each step we split $\mathcal{A}_i$ into observations $\mathcal{O}_i$ and withheld-but-available targets $\mathcal{T}_i = \mathcal{A}_i \setminus \mathcal{O}_i$, and learn to sample $\{\hat{x}_i^m\}_{m \in \mathcal{T}_i}$ from $X_i^{\mathcal{O}} = \{x_i^m | m \in \mathcal{O}_i\}$. Losses are defined only for $m \in \mathcal{A}_i$.

A note on terminology. “State space” here specifically refers to the coupled latent state (s_k^{sym}, x_t) with Markov transitions $p(\mathcal{S}^{\text{sym}}) = p(s_0^{\text{sym}}) \prod_k p(s_k^{\text{sym}} | s_{k-1}^{\text{sym}})$ and terminal-state emission $p(x | s_K^{\text{sym}})$. This is a latent-variable state-space process; it is *not* a structured SSM in the S4/Mamba sense (Gu et al., 2021; Gu and Dao, 2023) (transitions come from an autoregressive LLM, not learned linear recurrences), and we make no efficiency claims of that kind.

3.2 Markovian Semantic State Transition

Why plan in a discrete space. Sparse observations constrain meaning but not realization. Updating a continuous latent under this ambiguity entangles semantic inference with low-level synthesis and tends to drift under partial supervision. MST instead carries the reasoning in natural language, whose discrete topology filters aleatoric noise while preserving the causal skeleton that later conditions generation. The same mechanism degrades gracefully to the unimodal case ($|\mathcal{A}_i|=1$, e.g., patch-masked MNIST): the symbolic state then encodes within-modality structure (digit identity, coarse stroke layout) read from the visible region.

Feeding raw modalities to the LLM. A natural objection is that a text-pretrained LLM cannot read pixels or waveforms. We make the interface explicit. A *frozen* perception encoder $\mathcal{E}_{\text{enc}}$ turns each observed modality into two complementary, *observation-only* views: (i) *verbalized descriptors* $\mathbf{e} = \{e^m\}_{m \in \mathcal{O}_i}$, short strings from off-the-shelf recognizers (ASR transcripts, object/attribute tags, prosody/affect tags) that the LLM reads as text; and (ii) *soft evidence tokens*, continuous features projected by a light per-modality adapter g_m into

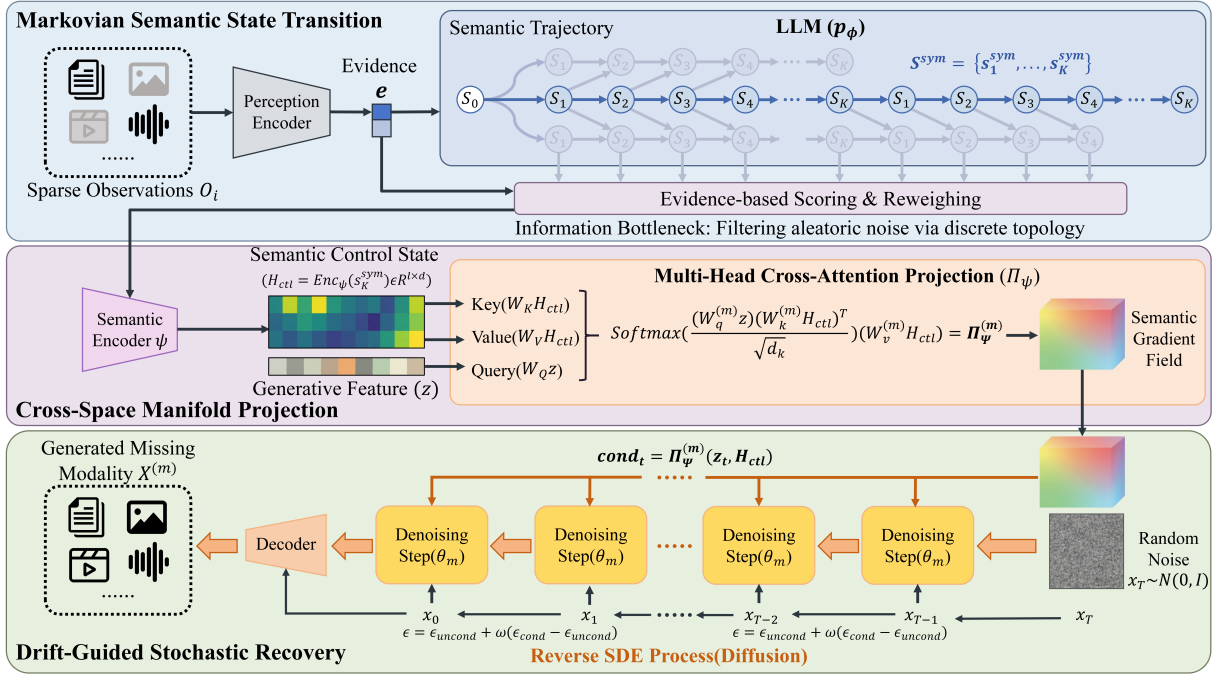


Figure 1: **Overview of HySSM.** HySSM resolves *what* is missing before rendering *how* it appears. From sparse observations, a frozen encoder extracts observation-only evidence. (1) **MST** samples S evidence-grounded trajectories of discrete semantic states via a frozen omni-modal LLM and reweights them into a control state H_{ctl} . (2) **CMP** injects H_{ctl} through cross-attention, yielding a coordinate-wise semantic field that localizes each concept. (3) **DSR** uses this field to drift-guide a conditional diffusion model, synthesizing the missing modalities.

the LLM embedding space, $\tilde{e}^m = g_m(\text{feat}(x^m))$, prepended as a soft prompt. With the fixed system instruction $\mathcal{I}$ (App. A.4), the LLM is “omni-modal” in *input*, conditioning only on $\mathcal{O}_i$, never on the label, so leakage is precluded by construction.

What a semantic state is. A state s_k^{sym} is a block of discrete tokens, a short proposition about the missing content (e.g., “a medium-pitched dog bark, single animal”). The plan is the chain of such blocks, generated autoregressively,

$$p_\phi(\mathcal{S}^{\text{sym}} | s_0^{\text{sym}}, \mathcal{I}, \mathbf{e}) = \prod_{k=1}^K p_\phi(s_k^{\text{sym}} | s_{k-1}^{\text{sym}}, \mathcal{I}, \mathbf{e}),$$

so sampling S trajectories approximates the posterior over plausible explanations of $X^{\mathcal{O}}$.

Evidence grounding. Free-form reasoning can hallucinate. We score each state by its agreement with the observed evidence using a calibrated similarity,

$$r(s_k^{\text{sym}}, \mathbf{e}) = \sigma\left(\frac{1}{\kappa} \langle \text{enc}(s_k^{\text{sym}}), \text{enc}(\mathbf{e}) \rangle\right) \in (0, 1],$$

with a frozen encoder $\text{enc}(\cdot)$ and temperature κ . Since $\mathbf{e}$ depends on $\mathcal{O}_i$, r cannot reference the target. We penalize ungrounded transitions, $\mathcal{L}_{\text{ground}} = -\sum_k \log r(s_k^{\text{sym}}, \mathbf{e})$, confining the plan to the evidence-compatible region of $\mathcal{M}_{\text{sym}}$.

Invariance across observed subsets. The same event may appear as $\{\text{img}, \text{txt}\}$ or $\{\text{aud}\}$. To stop the planner latching onto set-specific cues, we draw two subsets $\mathcal{O}_a, \mathcal{O}_b \subseteq \mathcal{A}_i$, pool each control into a single summary $\bar{H}(X^{\mathcal{O}})$ (mean over the L control tokens), and match them:

$$\mathcal{L}_{\text{cons}} = \mathbb{E}_{\mathcal{O}_a, \mathcal{O}_b \subseteq \mathcal{A}_i} [\|\bar{H}(X^{\mathcal{O}_a}) - \bar{H}(X^{\mathcal{O}_b})\|_2^2]. \quad (1)$$

This constrains only the *pooled, shared* intent; the per-token field H_{ctl} keeps modality- and location-specific detail through the coordinate-wise injection of Sec. 3.3. Collapse to a constant summary is prevented because the dominant reconstruction loss demands fidelity, the planner is frozen, and λ_2 is small (we confirm low sensitivity to λ_2 in Sec. 4).

Connecting discrete plans to differentiable training. Token sampling is non-differentiable, so we keep the LLM *frozen* and link planning to generation by a grounding-weighted aggregate. With cumulative reward $R_n = \sum_k \log r(s_{k,n}^{\text{sym}}, \mathbf{e}) \leq 0$,

$$H_{\text{ctl}} = \sum_{n=1}^S \omega_n \text{Enc}_\psi(s_{K,n}^{\text{sym}}), \quad (2)$$

$$\omega_n = \frac{\exp(R_n/\tau)}{\sum_j \exp(R_j/\tau)}.$$

Gradients reach only the differentiable encoder Enc_ψ and the emitters; the planner stays a stable bottleneck. *On planner adaptation:* any parameter-efficient tuning of the LLM (e.g., a QLoRA warm-up of the adapters g_m for input alignment) is done once as a separate stage and then frozen, i.e., it aligns the soft evidence tokens but neither makes the discrete bottleneck differentiable nor alters the main training dynamics.

3.3 Cross-Space Manifold Projection

Global conditioning is not enough. The plan s_K^{sym} lives on $\mathcal{M}_{\text{sym}}$ while the emitter generates on $\mathcal{M}_{\text{sig}}$. Simply concatenating token embeddings with continuous latents yields coarse, *global* conditioning: the generator knows *what* to produce but cannot localize *where/when* to apply it. CMP closes this gap. We first densify the plan into a differentiable, modality-agnostic control sequence, $H_{\text{ctl}} = \text{Enc}_\psi(s_K^{\text{sym}}) \in \mathbb{R}^{L \times d}$ (under reweighting, the aggregate of Eq. (2)). We then inject it through multi-head cross-attention from an emitter feature map $z \in \mathbb{R}^{L_z \times d_z}$:

$$\Pi_\psi^{(m)}(z, H_{\text{ctl}}) = \tau\left(\frac{(W_Q z)(W_K H_{\text{ctl}})^\top}{\sqrt{d}}\right)(W_V H_{\text{ctl}}).$$

where $\tau(\cdot)$ is the softmax function.

Reading the attention. Queries index the generator’s spatiotemporal coordinates (*where* guidance is needed); keys select *which* semantic attributes are relevant; values carry the content. The resulting map is a coordinate-wise semantic field over $\mathcal{M}_{\text{sig}}$, e.g., binding “barking” to the temporal frames where it should sound. This *localized* binding, rather than a single global vector, is what lets a discrete plan exert fine-grained control, and is the operative novelty distinguishing HySSM from uniform classifier-free or posterior guidance.

3.4 Drift-Guided Stochastic Recovery

Why stochastic, not regression. A fixed semantics still admits many signals: “a dog barking” maps to countless waveforms. Deterministic emitters average them; we instead model emission as conditional latent diffusion (Rombach et al., 2022; Ho et al., 2020), which keeps fidelity and diversity. Starting from $x_T \sim \mathcal{N}(0, I)$, we integrate the reverse-time SDE (Song et al., 2020)

$$dx = [f(x, t) - g^2(t) \nabla_x \log p_t(x | H_{\text{ctl}})] dt + g(t) d\bar{w}.$$

Semantic drift, localized. Building on posterior/classifier guidance (Ho and Salimans, 2022;

Chung et al., 2023), we approximate the conditional score with an ϵ -network, but route the condition through the localized projection $\Pi_\psi^{(m)}$ of section 3.3 rather than as a global vector:

$$\nabla_x \log p_t(x | H_{\text{ctl}}) \approx -\frac{1}{\sigma_t} \epsilon_{\theta_m}\left(x_t, t, \Pi_\psi^{(m)}(x_t, H_{\text{ctl}})\right).$$

Conditioning thus induces a coordinate-wise *semantic drift* that biases trajectories toward plan-consistent signals, while the diffusion term retains the stochasticity needed for diverse completions.

Tunable adherence. We balance realism against semantic fidelity with classifier-free guidance (Ho and Salimans, 2022) and a learnable null $\emptyset$:

$$\tilde{\epsilon} = \epsilon_{\theta_m}(x_t, t, \emptyset) + \omega \left(\epsilon_{\theta_m}(x_t, t, H_{\text{ctl}}) - \epsilon_{\theta_m}(x_t, t, \emptyset) \right). \quad (3)$$

where larger ω tightens adherence at the cost of diversity (analyzed in section 4). Each emitter is trained with the standard denoising objective

$$\mathcal{L}_{\text{inst}}^{(m)} = \mathbb{E}_{x_0^{(m)}, \epsilon, t} [\|\epsilon - \epsilon_{\theta_m}(x_t^{(m)}, t, H_{\text{ctl}})\|_2^2]. \quad (4)$$

3.5 Training and Inference

Randomized masking. For each sample we split $\mathcal{A}_i$ into $(\mathcal{O}_i, \mathcal{T}_i)$ at a fixed missing rate r but a randomized pattern, forcing inter-modal dependencies rather than a fixed observed $\rightarrow$ target map, the basis for robustness to agnostic masking shift (Sec. 4). The training loss is:

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{i, \mathcal{O}_i \subseteq \mathcal{A}_i} \left[\sum_{m \in \mathcal{T}_i} \mathcal{L}_{\text{inst}}^{(m)} + \lambda_1 \mathcal{L}_{\text{ground}} + \lambda_2 \mathcal{L}_{\text{cons}} \right]. \quad (5)$$

We update the emitters $\{\theta_m\}$ and encoder ψ (and adapters g_m if not pre-frozen); planner stays frozen.

Inference. Given $X^\mathcal{O}$, HySSM (i) samples S semantic plans via the frozen LLM, (ii) aggregates them into a deterministic H_{ctl} by grounding-based reweighting, and (iii) renders each missing modality by integrating the guided reverse SDE from Gaussian noise. The control is deterministic while the signal is stochastic, so multiple noise seeds yield *diverse* completions that share the inferred semantics, a property we quantify in section 4.

Algorithm 1: HySSM training

Require: Dataset $\mathcal{D}$; frozen Omni-LLM p_ϕ ; denoisers $\{\epsilon_{\theta_m}\}$; encoder Enc_ψ .

- 1: **while** not converged **do**
- 2: Sample batch $\mathcal{B} \subseteq \mathcal{D}$.
- 3: **for** each $i \in \mathcal{B}$ **do**
- 4: **Mask:** split $\mathcal{A}_i \rightarrow (\mathcal{O}_i, \mathcal{T}_i)$ at rate r , random pattern.
- 5: **Evidence:** $\mathbf{e} = \mathcal{E}_{\text{enc}}(X^{\mathcal{O}_i})$ (verbalized + soft tokens).
- 6: **Plan:** sample S trajectories $\{\mathcal{S}_n^{\text{sym}}\}$ from $p_\phi(\cdot \mid \mathbf{e}, \mathcal{I})$.
- 7: **Control:** $R_n = \sum_k \log r(s_{k,n}^{\text{sym}}, \mathbf{e})$; $H_{\text{ctl}} = \sum_n \omega_n \text{Enc}_\psi(s_{K,n}^{\text{sym}})$, $\omega_n \propto e^{R_n/\tau}$.
- 8: **Consistency:** if $|\mathcal{A}_i| > 1$, draw $\mathcal{O}_a, \mathcal{O}_b$ and accumulate $\mathcal{L}_{\text{cons}}$ (Eq. (1)).
- 9: **for** $m \in \mathcal{T}_i$ **do**
- 10: $t \sim \mathcal{U}[0, 1]$, $\epsilon \sim \mathcal{N}(0, I)$; accumulate $\mathcal{L}_{\text{inst}}^{(m)}$ (Eq. (4)) via $\Pi_\psi^{(m)}$.
- 11: **end for**
- 12: **end for**
- 13: Update $\{\theta_m\}, \psi$ by $\nabla \mathcal{L}_{\text{total}}$; planner frozen.
- 14: **end while**

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate HySSM on MNIST (LeCun et al., 1998), Hateful Memes (Kielbaso et al., 2020), and CMU-MOSEI (Zadeh et al., 2018) under missingness patterns. MNIST benchmarks image reconstruction. Hateful Memes and CMU-MOSEI evaluate feature-level cross-modal completion for classification, using CLIP/Roberta and BERT/COVAREP/Facet streams, respectively. We report RMSE/Accuracy on MNIST, AUROC/Accuracy on Hateful Memes, and F1/Acc-3 on CMU-MOSEI, treating CMU-MOSEI as a sentiment task.

Missingness scenarios. We consider settings with missing rate $r \in [0.1, 0.9]$. (1) Intra-modal (MNIST): randomized window masking corrupts image signals (Dai et al., 2025a). (2) Inter-modal (Hateful Memes, CMU-MOSEI): we drop entire modalities to simulate sensor loss, while enforcing $|\mathcal{O}_i| \geq 1$ for evidence grounding. To evaluate agnostic masking shift (AMS), all methods are trained only at $r = 0.5$ and tested across the full range.

Implementation details. We encode each modality with off-the-shelf features and project them into a shared 512-D latent via linear layers. Our model comprises a frozen omni-modal LLM planner, a 6-layer Transformer semantic encoder ($d=512$), and a HySSM evolution module with state size $N=64$, followed by modality-specific latent diffusion models. We reconstruct signals with 50-step DDIM sampling. Training uses AdamW (lr= 10^{-4} , batch size 64) for 100 epochs at $r=0.5$, with $\lambda_1=0.5$ and $\lambda_2=0.1$. Experiments run on NVIDIA A100 (80GB) GPUs. More implementation details, please refer to Appendix B.

Baselines. We compare HySSM with 10 methods spanning two paradigms. *Non-generative representation* approaches improve robustness, including Robult (Nguyen et al., 2025), Memory-Driven Prompt (Zhao et al., 2025), HME (Zhuang et al., 2025), and StableMiss (Zhu et al., 2023). *Generative/completion* approaches reconstruct missing signals or align manifolds, including MD²N (Dai et al., 2025a), Knowledge Bridger (Ke et al., 2025), CAPIMAC (Ma et al., 2025), SMCMSA (Sun et al., 2024), IMDer (Wang et al., 2023c), DiC-MoR (Wang et al., 2023b). Adaptation details for three datasets are deferred to Appendix B.2.

4.2 Results and Analysis

Tables 1–3 compare HySSM with recent baselines under agnostic masking shift.

MNIST. HySSM achieves the best RMSE/accuracy at all missing rates, with the advantage becoming more evident under severe missingness. At $r = 0.9$, HySSM still obtains 0.215/69.4%, outperforming the strongest baseline and indicating that semantic planning helps preserve digit-level structure when most pixels are unobserved.

Hateful Memes. HySSM also performs best under whole-modality missingness. Compared with Knowledge Bridger, it improves from 51.3/47.1 to 54.6/48.2 at $r = 0.9$, suggesting that evidence-grounded semantic states are useful when either visual or textual cues are largely absent.

CMU-MOSEI. HySSM attains 54.1/59.2 at $r = 0.9$, surpassing StableMiss (42.1/48.1) and SMCMSA (52.1/58.4). The gain on this temporally aligned multimodal dataset further shows that semantic-first planning can mitigate semantic drift under extreme missingness and masking shifts.

Across reconstruction (MNIST) and downstream classification (Hateful Memes, CMU-MOSEI), HySSM consistently ranks first under AMS and

Table 1: MNIST performance under Agnostic Masking Shift (AMS) (RMSE ↓ / Acc ↑). Models are trained at $r = 0.5^*$ and evaluated across all $r \in [0.1, 0.9]$. **Bold** and underline denote the best and second-best results.

Category	Method	0.1	0.2	0.3	0.4	0.5*	0.6	0.7	0.8	0.9
Non-Gen	StableMiss	0.129/90.1	0.125/89.4	0.128/88.5	0.118/87.2	0.114/87.1	0.158/84.4	0.202/76.1	0.268/68.5	0.325/53.1
	HME	0.118/92.2	0.112/93.1	0.114/92.5	0.108/94.2	0.102/94.1	0.135/87.8	0.181/79.2	0.241/70.5	0.288/61.2
	Robult	0.102/96.2	0.098/95.1	0.101/95.5	<u>0.092/94.1</u>	0.091/93.2	0.125/90.8	0.158/85.4	<u>0.198/77.1</u>	0.259/64.5
	Memory-P	0.104/95.5	<u>0.095/94.2</u>	0.098/95.1	0.095/94.4	<u>0.092/94.8</u>	<u>0.116/92.5</u>	0.152/86.1	0.202/78.4	0.248/66.8
Gen	CAPIMAC	0.168/84.1	0.164/82.5	0.158/84.8	0.151/83.4	0.144/84.5	0.195/78.1	0.245/70.2	0.312/59.4	0.388/44.5
	DiC-MoR	0.124/91.2	0.121/90.4	0.118/91.8	0.115/92.4	0.112/93.5	0.152/86.4	0.195/78.5	0.251/70.1	0.312/55.1
	IMDer	0.101/95.8	0.104/94.8	0.098/95.4	<u>0.092/94.8</u>	0.095/93.6	0.128/89.8	0.165/83.2	0.218/76.8	0.272/62.5
	SMCMSA	0.098/96.1	0.096/ <u>95.5</u>	0.095/95.2	0.094/94.1	0.085/93.5	0.122/91.2	<u>0.148/85.8</u>	0.208/ <u>78.7</u>	<u>0.242/68.5</u>
	KnowledgeB	0.104/95.2	0.099/93.8	0.102/91.5	0.094/90.2	0.091/91.1	0.128/88.1	0.168/82.5	0.215/75.4	0.271/61.8
	MD ² N	<u>0.094/96.8</u>	0.098/95.2	<u>0.094/95.6</u>	<u>0.092/94.8</u>	<u>0.088/93.3</u>	0.124/91.8	0.152/ <u>87.2</u>	0.204/77.8	0.248/66.1
Ours	HySSM	0.082/97.3	0.085/96.4	0.087/95.7	0.089/94.9	0.092/95.1	0.105/92.9	0.130/88.2	0.170/79.0	0.215/69.4

Table 2: Hateful Memes performance under Agnostic Masking Shift (AMS) (AUROC / Acc ↑). Models are trained at $r = 0.5^*$ and evaluated across all $r \in [0.1, 0.9]$. **Bold** and underline denote the best and second-best results.

Category	Method	0.1	0.2	0.3	0.4	0.5*	0.6	0.7	0.8	0.9
Non-Gen	StableMiss	66.1/61.2	65.4/60.5	64.8/59.8	63.2/58.5	63.8/59.1	60.4/55.2	56.1/51.1	51.4/46.5	45.1/39.2
	HME	68.2/63.1	67.5/62.4	66.8/61.8	65.5/61.1	66.1/61.5	63.2/58.1	59.4/54.2	54.2/49.1	48.4/42.1
	Robult	<u>72.1/66.4</u>	<u>71.4/65.8</u>	<u>70.5/65.1</u>	69.1/64.2	69.8/64.8	67.4/62.1	<u>64.1/59.1</u>	58.5/53.1	51.1/46.2
	Memory-P	70.5/64.8	69.8/64.1	68.5/63.1	67.1/62.1	67.8/62.5	64.5/59.4	61.2/56.1	56.4/51.4	50.1/44.1
Gen	CAPIMAC	62.5/57.1	61.1/56.2	60.2/55.4	58.5/54.1	59.1/54.5	55.4/50.1	51.2/46.1	46.1/40.1	40.2/34.1
	DiC-MoR	67.1/62.4	66.2/61.5	65.5/60.8	64.1/59.1	64.8/59.8	61.2/56.2	57.1/52.1	52.1/47.2	45.5/39.2
	IMDer	69.4/64.8	68.8/64.1	68.1/63.2	66.5/62.1	67.2/62.8	64.1/59.4	60.2/55.4	55.1/50.5	49.2/43.1
	SMCMSA	<u>71.2/66.5</u>	70.5/65.4	69.8/64.5	<u>69.5/64.5</u>	<u>70.2/65.1</u>	<u>68.1/63.1</u>	65.0/60.1	56.1/51.1	49.1/44.1
	KnowledgeB	70.8/65.4	70.1/64.4	69.4/63.5	68.1/62.1	68.8/62.8	65.1/60.2	61.2/56.4	59.1/54.4	51.3/47.1
	MD ² N	71.4/66.2	70.8/65.1	69.8/64.4	69.1/64.1	69.8/64.8	66.5/61.1	62.8/57.2	57.5/52.1	51.0/46.5
Ours	HySSM	74.5/69.4	74.3/69.0	73.8/68.3	71.5/67.4	71.0/67.6	68.5/64.5	65.0/60.5	59.2/54.5	54.6/48.2

degrades gracefully as missingness increases. This pattern supports our key design choice: semantic-first planning with a stable symbolic bottleneck, followed by conditioned generation, improves robustness to both extreme missing rates and masking distribution shifts.

4.3 Ablation Study

We quantify the contribution of each component in HySSM across architectural choices, bottleneck design, projection operators, and training objectives. Unless stated otherwise, results are averaged over missing rates $r \in \{0.1, \dots, 0.9\}$ and five runs (additional ablations in the supplementary).

HySSM architecture. Tab. 4 validates the core design of semantic planning $\rightarrow$ signal rendering. Removing the hierarchy (ii) consistently degrades performance (e.g., CMU-MOSEI F1: 69.5 $\rightarrow$ 66.2), indicating that explicit semantic evolution mitigates the one-to-many ambiguity of incomplete observations. Eliminating hybrid coupling (iii) causes the largest drop (e.g., Hateful Memes AUROC: 68.0 $\rightarrow$ 61.2), showing that aligning discrete se-

mantics with continuous synthesis is essential. Replacing Markov transitions with global pooling (iv) is worse than the Markovian variant, suggesting completion benefits from step-wise state transitions rather than static mapping. A plain MLP (v) performs worst, confirming that structured dynamics and cross-space coupling are both necessary.

Discrete symbolic bottleneck. Fig. 2 shows that discrete tokens provide an effective information bottleneck. Replacing tokens with continuous embeddings (ii) reduces robustness, consistent with increased sensitivity to mask-induced variability. Soft-Gumbel (iii) partially recovers performance but remains below hard tokens, while removing planning entirely (iv) causes the largest degradation, confirming that explicit semantic trajectories are critical under agnostic mask shifts.

Manifold projection. Fig. 3 demonstrates that cross-attention is the most effective projection from symbolic controls to signal-space generation. Simple feature concatenation (ii) is notably weaker, indicating that coarse global conditioning fails to localize semantic cues in the continuous manifold.

Table 3: CMU-MOSEI performance under Agnostic Masking Shift (AMS) (F1 / Acc-3 $\uparrow$). Models are trained at $r = 0.5^*$ and evaluated across all $r \in [0.1, 0.9]$. **Bold** and underline denote the best and second-best results.

Category	Method	0.1	0.2	0.3	0.4	0.5*	0.6	0.7	0.8	0.9
Non-Gen	StableMiss	68.2/71.4	67.4/70.5	66.1/69.8	65.1/68.4	65.5/68.5	62.1/65.1	57.1/61.1	50.2/55.4	42.1/48.1
	HME	70.1/72.8	69.4/72.1	69.8/72.4	67.8/70.2	68.1/70.8	64.2/67.5	60.1/63.4	54.1/58.1	46.5/51.2
	Robult	73.2/75.1	71.5/74.2	<u>72.1/74.8</u>	70.5/73.1	71.1/73.8	68.2/71.1	64.2/67.4	58.1/63.5	50.2/56.4
	Memory-P	71.4/74.1	70.8/73.4	69.4/72.5	68.1/71.5	69.4/72.1	69.4/72.8	61.2/65.8	55.1/60.1	47.5/54.1
Gen	CAPIMAC	64.1/67.1	62.1/65.4	62.5/65.8	60.4/63.2	61.1/63.8	57.1/60.4	51.2/55.1	44.1/48.1	36.2/41.5
	DiC-MoR	68.8/71.1	67.5/70.4	66.8/69.8	65.4/68.2	66.1/68.8	62.1/65.4	57.2/61.1	51.2/56.1	44.5/50.1
	IMDer	72.1/74.8	70.8/73.5	71.1/73.8	<u>71.8/74.1</u>	70.1/72.8	67.4/70.1	63.1/66.2	57.1/61.4	48.2/55.1
	SMCMSA	74.1/76.5	73.2/75.1	71.8/74.2	71.2/73.8	<u>72.4/75.1</u>	68.5/71.8	<u>65.1/68.5</u>	58.2/62.4	<u>52.1/58.4</u>
	KnowledgeB	72.1/74.5	71.1/73.8	71.1/73.8	69.1/72.1	69.8/72.8	67.1/70.1	62.4/66.8	57.1/61.8	49.1/55.1
	MD ² N	72.4/75.1	<u>73.5/75.8</u>	71.2/74.1	70.8/73.4	71.5/74.1	68.1/70.8	64.2/67.8	<u>59.1/63.1</u>	50.4/57.2
Ours	HySSM	76.1/78.4	75.8/78.2	75.0/77.5	74.1/75.7	73.4/75.2	70.8/72.0	66.5/68.8	60.0/63.9	54.1/59.2

Table 4: Ablation study on HySSM architecture. Values averaged across $r \in [0.1, 0.9]$. **Bold** denotes best.

Variants	MNIST (RMSE / Acc)	Hateful Memes (AUROC / Acc)	CMU-MOSEI (F1 / Acc)
(i) Full HySSM (Ours)	0.117 / 89.0	68.0 / 63.2	69.5 / 72.1
(ii) w/o Hierarchy	0.136 / 86.4	64.9 / 60.1	66.2 / 68.8
(iii) w/o HySSM	0.154 / 82.5	61.2 / 56.9	63.1 / 65.4
(iv) Non-Markovian	0.125 / 87.8	66.7 / 62.4	68.1 / 70.8
(v) Baseline MLP	0.185 / 76.2	55.4 / 49.8	56.8 / 61.2

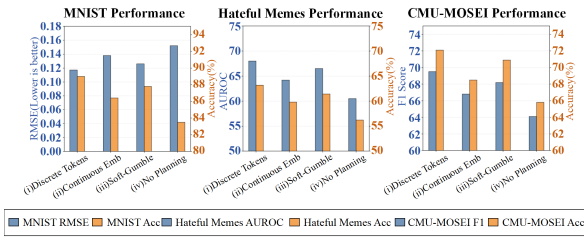


Figure 2: Ablation of Discrete Symbolic Bottleneck on MNIST, Hateful Memes, and CMU-MOSEI. All values are averaged across missing rates $r \in [0.1, 0.9]$.

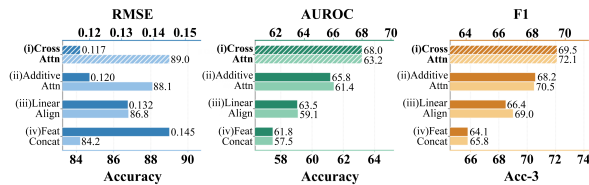


Figure 3: Ablation on Manifold Projection operator Π_ψ . Values are averaged across missing rates $r \in [0.1, 0.9]$.

Linear alignment (iii) and additive attention (iv) improve over concatenation but still lag behind cross-attention, supporting our manifold projection for fine-grained semantic-to-signal binding.

Regularization terms. Tab. 5 confirms the importance of grounding and subset-consistency regularization. Removing $\mathcal{L}_{\text{ground}}$ sharply reduces performance (e.g., Hateful Memes AUROC: 68.0 $\rightarrow$ 61.5), consistent with ungrounded semantic transitions. Removing $\mathcal{L}_{\text{cons}}$ yields a smaller but consis-

Table 5: Ablation study on regularization terms $\mathcal{L}_{\text{ground}}$ and $\mathcal{L}_{\text{cons}}$. Values averaged across $r \in [0.1, 0.9]$. **Bold** denotes best.

Variants	MNIST (RMSE / Acc)	Hateful Memes (AUROC / Acc)	CMU-MOSEI (F1 / Acc)
(i) Full Objective	0.117 / 89.0	68.0 / 63.2	69.5 / 72.1
(ii) w/o $\mathcal{L}_{\text{ground}}$	0.142 / 85.3	61.5 / 57.4	65.4 / 67.8
(iii) w/o $\mathcal{L}_{\text{cons}}$	0.125 / 87.5	65.2 / 60.9	67.1 / 69.8
(iv) Baseline ($\mathcal{L}_{\text{inst}}$ only)	0.170 / 78.5	58.6 / 53.4	60.2 / 62.5

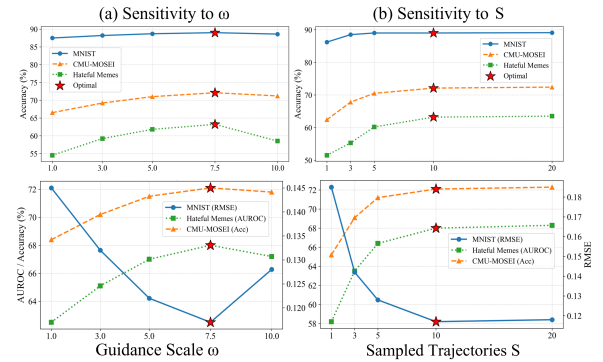


Figure 4: Sensitivity of guidance scale ω (a) and sampled semantic trajectories S (b). All values are averaged across missing rates $r \in [0.1, 0.9]$. Red stars denote peak performance.

tent drop (e.g., CMU-MOSEI F1: 69.5 $\rightarrow$ 67.1), indicating that enforcing invariance across observation subsets improves robustness to missing-pattern shifts. Using only $\mathcal{L}_{\text{inst}}$ performs worst, showing that structured regularization is critical beyond instance-level supervision.

Hyper-parameter sensitivity. Fig. 4 analyzes the guidance scale ω and the number of sampled semantic trajectories S on CMU-MOSEI. Increasing ω initially improves conditional adherence, reaching an optimal range around $\omega = 7.5$ (marked by stars). However, overly large guidance scales lead

to a drop in fidelity or performance plateau, reflecting the standard guidance trade-off. Regarding trajectory count, performance improves significantly with S but saturates near $S \approx 10$. The marginal gains observed when increasing S to 20 are negligible compared to the linear growth in runtime, motivating $S = 10$ as an efficient practical default.

5 Conclusion

We presented Semantic-first Completion (SFC) for incomplete multimodal learning, which first resolves evidence-consistent semantics and then stochastically realizes missing signals under natural missingness. Instantiated as HySSM, our method combines Markovian semantic state transitions with a discrete token bottleneck, cross-space manifold projection, and drift-guided conditional diffusion for missing-modality recovery. Experiments across multiple datasets show consistent gains under diverse missing patterns and strong robustness to agnostic masking shift, including extreme incompleteness up to 90%. Ablations further verify the complementary roles of semantic planning, grounding, and consistency regularization.

Limitations

Although HySSM demonstrates strong robustness under diverse missingness patterns, HySSM still has limitations. **First**, its semantic planning stage depends on the quality of the frozen planner and extracted evidence, so weak or misleading observations may lead to imperfect semantic trajectories. **Second**, sampling multiple plans and diffusion-based recovery introduce higher computational cost than single-pass completion methods. **Finally**, our experiments cover several representative benchmarks, but broader validation on larger-scale and more open-world multimodal missingness scenarios remains future work.

Acknowledgments

This research was partially supported by the National Natural Science Foundation of China (NSFC) (granted No. 62306064) and ZYGX2025XJ049. We appreciate all the authors for their fruitful discussions. In addition, thanks are extended to anonymous reviewers for their insightful comments.

References

- Saghir Alfasly, Jian Lu, Chen Xu, and Yuru Zou. 2022. Learnable irrelevant modality dropout for multimodal action recognition on modality-specific annotated videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20208–20217.
- Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. 2022. Multima: Multi-modal multi-task masked autoencoders. In *European conference on computer vision*, pages 348–367. Springer.
- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.
- Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. 2023. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. The International Conference on Learning Representations.
- Ruiting Dai, Zesen Cai, Lisi Mo, Guiduo Duan, Keren Shi, and Tao He. 2026a. Anchor drift no more: Hierarchical consistency-guided prompt distillation for incomplete multimodal learning. In *Proceedings of the ACM Web Conference 2026*, pages 7330–7341.
- Ruiting Dai, Aoqiang Jian, Ruoxuan Zhang, Yuang Jia, Lisi Mo, Siyu Zhan, and Tao He. 2026b. Multimodal learning with missing modalities: Can progressive diffusion achieve distribution-consistent learning from incomplete multimodal data? *Knowledge-Based Systems*, page 116272.
- Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. 2025a. Unbiased missing-modality multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24507–24517.
- Ruiting Dai, Yuqiao Tan, Lisi Mo, Tao He, Ke Qin, and Shuang Liang. 2025b. Robustpt: Dynamic disentanglement prompt tuning in vision-language models with missing modalities. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 164–172.
- Ruiting Dai, Wenwei Zhu, Zheyu Wang, Haoran Meng, Zhengdao Yuan, Yandong Yan, and Lisi Mo. 2025c. γ -crd: Gamma-cooperative retrieval diffusion model for robust incomplete multimodal learning. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 173–182.
- Qian Dong, Jiayu An, Tianqi Lu, Tao He, and Ruiting Dai. 2026. Causality matters: Temporal causal-aware reasoning for multimodal learning under missing modalities. In *Proceedings of the 2026 9th International Conference on Software Engineering and Information Management*, pages 55–59.

- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.
- Yuxin Guo, Siyang Sun, Shuailei Ma, Kecheng Zheng, Xiaoyi Bao, Shijie Ma, Wei Zou, and Yun Zheng. 2024. Crossmae: Cross-modality masked autoencoders for region-aware audio-visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26721–26731.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- Lun Huang, You Xie, Hongyi Xu, Tianpei Gu, Chenxu Zhang, Guoxian Song, Zenan Li, Xiaochen Zhao, Linjie Luo, and Guillermo Sapiro. 2025. Plan-x: Instruct video generation via semantic planning. *arXiv preprint arXiv:2511.17986*.
- Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. 2025. Knowledge bridge: Towards training-free missing modality completion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25864–25873.
- Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, Pierre Vera, and Su Ruan. 2025. Amm-diff: Adaptive multi-modality diffusion network for missing modality imputation. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Chongxuan Li, Jun Zhu, and Bo Zhang. 2016. Learning to generate with memory. In *International conference on machine learning*, pages 1177–1186. PMLR.
- Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2302–2310.
- Shubin Ma, Liang Zhao, Mingdong Lu, Yifan Guo, and Bo Xu. 2025. Consistency-aware padding for incomplete multi-modal alignment clustering based on self-repellent greedy anchor search. *arXiv preprint arXiv:2507.03917*.
- Xiangxi Meng, Kaicong Sun, Jun Xu, Xuming He, and Dinggang Shen. 2024. Multi-modal modality-masked diffusion network for brain mri synthesis with random modality missing. *IEEE Transactions on Medical Imaging*, 43(7):2587–2598.
- Lisi Mo, Aoqiang Jian, Ruofan Zhou, Letian Xu, and Ruiting Dai. 2026. Hipmd: Hierarchical prompting meets multi-layer distillation for incomplete multi-modal learning. In *ICC 2026-IEEE International Conference on Communications*, pages 1–6. IEEE.
- Duy A Nguyen, Abhi Kamboj, and Minh N Do. 2025. Robult: Leveraging redundancy and modality specific features for robust multimodal learning. *arXiv preprint arXiv:2509.03477*.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.
- Yuhang Sun, Zhizhong Liu, Quan Z Sheng, Dianhui Chu, Jian Yu, and Hongxiang Sun. 2024. Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 110:102454.
- Kim-Han Thung, Pew-Thian Yap, Ehsan Adeli, Seong-Wan Lee, Dinggang Shen, Alzheimer’s Disease Neuroimaging Initiative, and 1 others. 2018. Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion. *Medical image analysis*, 45:68–82.
- Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. 2023a. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15878–15887.
- Yuanzhi Wang, Zhen Cui, and Yong Li. 2023b. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision*, pages 22025–22034.
- Yuanzhi Wang, Yong Li, and Zhen Cui. 2023c. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36:17117–17128.
- Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31.
- Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. 2024a. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*.
- Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Michael Nalls, Jimeng Sun, and Faraz Faghri. 2024b. Multimodal patient representation learning with missing modalities and labels. In *International Conference on Learning Representations*, volume 2024, pages 55538–55557.
- Yingxue Xu, Fengtao Zhou, Chenyu Zhao, Yihui Wang, Can Yang, and Hao Chen. 2025. Distilled prompt learning for incomplete multimodal survival prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5102–5111.
- AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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)*, pages 2236–2246.
- Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. 2022. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1545–1554.
- Hengrui Zhang, Liancheng Fang, Qitian Wu, and Philip Yu. 2025a. Diffputer: Empowering diffusion models for missing data imputation. In *International Conference on Learning Representations*, volume 2025, pages 63164–63185.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847.
- Wei Zhang, Juan Chen, Yanbo J Wang, En Zhu, Xuan Yang, and Yiduo Wang. 2025b. Admc: Attention-based diffusion model for missing modalities feature completion. *arXiv preprint arXiv:2507.05624*.
- Jinming Zhao, Ruichen Li, and Qin Jin. 2021. 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)*, pages 2608–2618.
- Yihan Zhao, Wei Xi, Xiao Fu, and Jizhong Zhao. 2025. Enhancing multimodal model robustness under missing modalities via memory-driven prompt learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence, IJCAI 2025*, pages 2458–2466. International Joint Conferences on Artificial Intelligence.
- Wenjin Zhong, Cong Cong, Zihan Wang, Zeya Yan, Antonio Di Ieva, and Sidong Liu. 2025. Fmm-diff: A feature mapping and merging diffusion model for mri generation with missing modality. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 227–236. Springer.
- Yichen Zhu, Jian Yuan, Bo Jiang, Tao Lin, Haiming Jin, Xinbing Wang, and Chenghu Zhou. 2023. Prediction with incomplete data under agnostic mask distribution shift. *arXiv preprint arXiv:2305.11197*.
- Yuke Zhu, Daniel Gordon, Eric Kolve, Dieter Fox, Li Fei-Fei, Abhinav Gupta, Roozbeh Mottaghi, and Ali Farhadi. 2017. Visual semantic planning using deep successor representations. In *Proceedings of the IEEE international conference on computer vision*, pages 483–492.
- Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Wei Li, Jiawen Deng, and Fuji Ren. 2025. Hypermodality enhancement for multimodal sentiment analysis with missing modalities. *Advances in Neural Information Processing Systems*, 38:142100–142129.

A Theoretical Derivations and Model Details

A.1 Formal Dynamical Evolution of the HySSM

The core of the HySSM lies in coupling the discrete symbolic manifold $\mathcal{M}_s$ and the continuous signal manifold $\mathcal{M}_x$ via a joint score function. This section provides a step-by-step derivation from probabilistic factorization to the correction of the stochastic differential equation (SDE).

Chain Decomposition of the Joint Score Function. Consider the joint probability distribution $p_t(x_t, s_t)$ at diffusion time t . By applying the product rule of probability, the evolution of the total score function for signal state x_t is derived as:

$$\begin{aligned}\nabla_{x_t} \log p_t(x_t, s_t) &= \nabla_{x_t} \log [p_t(x_t) \cdot p_t(s_t|x_t)] \\ &= \nabla_{x_t} [\log p_t(x_t) + \log p_t(s_t|x_t)] \\ &= \nabla_{x_t} \log p_t(x_t) \\ &\quad + \nabla_{x_t} \log p_t(s_t|x_t)\end{aligned}\tag{6}$$

Utilizing Bayes' rule $p_t(s_t|x_t) = \frac{p_t(x_t|s_t)p_t(s_t)}{p_t(x_t)}$, and noting that $\nabla_{x_t} \log p_t(s_t) = 0$:

$$\begin{aligned}\nabla_{x_t} \log p_t(x_t, s_t) &= \nabla_{x_t} \log p_t(x_t) \\ &\quad + \nabla_{x_t} \log p_t(x_t|s_t) \\ &\quad - \nabla_{x_t} \log p_t(x_t) \\ &= \nabla_{x_t} \log p_t(x_t|s_t)\end{aligned}\tag{7}$$

This chain of equalities demonstrates that the signal state dynamics are governed by conditioning on the symbolic state s_t .

Coupled Reverse SDE Expansion. Substituting the score function into the reverse-time SDE, let $\mathbf{f}(x, t)$ be the drift term and $g(t)$ the diffusion term. The trajectory dx is expanded as:

$$\begin{aligned}dx &= [\mathbf{f}(x, t) - g^2(t) \nabla_x \log p_t(x, s)] dt + g(t) d\bar{w} \\ &= \left[\mathbf{f}(x, t) - g^2(t) (\nabla_x \log p_t(x) \right. \\ &\quad \left. + \nabla_x \log p_t(s|x)) \right] dt + g(t) d\bar{w} \\ &= \underbrace{[\mathbf{f}(x, t) - g^2(t) \nabla_x \log p_t(x)] dt}_{\text{Physical Diffusion}} \\ &\quad - \underbrace{g^2(t) \nabla_x \log p_t(s|x) dt}_{\text{Semantic Correction}} + g(t) d\bar{w}\end{aligned}\tag{8}$$

Trajectory Correction under AMS. Under Agnostic Masking Shift (AMS), where original observations are uncertain, we utilize the semantic trajectory $\mathcal{T}$ generated by the MST module as an equivalent parameterization of s_t . The guidance term $\nabla_x \log p_t(s|x)$ is instantiated as:

$$\begin{aligned}\nabla_x \log p_t(s|x) &\approx \nabla_x \log p_t(\mathcal{T}|x) \\ &\propto -\nabla_x \left[\frac{1}{2\sigma^2} \|\phi(x_t) - \mathcal{T}\|_2^2 \right] \\ &= -\frac{1}{\sigma^2} (\phi(x_t) - \mathcal{T}) \frac{\partial \phi(x_t)}{\partial x_t}\end{aligned}\tag{9}$$

where $\phi(\cdot)$ is the semantic mapping operator. This derivation shows how HySSM directs the signal manifold toward the semantic trajectory via the residual $(\phi(x_t) - \mathcal{T})$, ensuring logical consistency even with severe input deficits.

A.2 Mathematical Proof of Drift-Guided Stochastic Recovery (DSR)

DSR aims to transform the semantic trajectory $\mathcal{T}$ inferred from the discrete symbolic space into a dynamic force field within the reverse diffusion process. This section provides the complete derivation from the conditional score function to the instantiation of the drift term.

Bayesian Derivation Chain of the Conditional Score Function. Given observations $\mathcal{O}$, sampling on the signal manifold $\mathcal{M}_x$ is driven by the conditional score function $\nabla_{x_t} \log p_t(x_t|\mathcal{O})$. The decomposition follows:

$$\begin{aligned}\nabla_{x_t} \log p_t(x_t|\mathcal{O}) &= \nabla_{x_t} \log \left[\frac{p_t(\mathcal{O}|x_t)p_t(x_t)}{p(\mathcal{O})} \right] \\ &= \nabla_{x_t} [\log p_t(\mathcal{O}|x_t) \\ &\quad + \log p_t(x_t) - \log p(\mathcal{O})] \\ &= \nabla_{x_t} \log p_t(x_t) \\ &\quad + \nabla_{x_t} \log p_t(\mathcal{O}|x_t) - 0 \\ &= \underbrace{\nabla_{x_t} \log p_t(x_t)}_{\text{Prior Score}} \\ &\quad + \underbrace{\nabla_{x_t} \log p_t(\mathcal{O}|x_t)}_{\text{Likelihood Guidance}}.\end{aligned}\tag{10}$$

Parameterization of Likelihood via Semantic Trajectory $\mathcal{T}$. Under AMS, directly solving $p_t(\mathcal{O}|x_t)$ is ill-posed. We model the likelihood as an isotropic Gaussian centered on the semantic trajectory $\mathcal{T}$. The derivation of the log-likelihood

gradient is:

$$\begin{aligned}
\nabla_{x_t} \log p_t(\mathcal{O}|x_t) &\approx \nabla_{x_t} \log \mathcal{N}(\phi(x_t); \mathcal{T}, \sigma^2 \mathbf{I}) \\
&= \nabla_{x_t} \left[-\frac{1}{2\sigma^2} \|\phi(x_t) - \mathcal{T}\|_2^2 \right. \\
&\quad \left. - \log(\sqrt{2\pi}\sigma)^d \right] \\
&= -\frac{1}{2\sigma^2} \nabla_{x_t} \|\phi(x_t) - \mathcal{T}\|_2^2 \\
&= -\frac{1}{\sigma^2} (\phi(x_t) - \mathcal{T}) \mathbf{J}_\phi(x_t)
\end{aligned} \tag{11}$$

where $\mathbf{J}_\phi(x_t) = \frac{\partial \phi(x_t)}{\partial x_t}$ denotes the Jacobian matrix of the perception operator.

Instantiation of the SDE Drift Term. Combining the guidance scale ω and the diffusion coefficient $g(t)$, the total drift adjustment $\Delta \mathbf{f}$ in the reverse SDE is instantiated as:

$$\begin{aligned}
\Delta \mathbf{f} &= \omega \cdot g^2(t) \nabla_{x_t} \log p_t(\mathcal{O}|x_t) \\
&= \omega g^2(t) \left[-\frac{1}{\sigma^2} (\phi(x_t) - \mathcal{T}) \mathbf{J}_\phi(x_t) \right] \\
&= -\frac{\omega g^2(t)}{\sigma^2} (\phi(x_t) - \mathcal{T}) \frac{\partial \phi(x_t)}{\partial x_t}
\end{aligned} \tag{12}$$

This chain of equalities proves how DSR dynamically regulates the direction of stochastic evolution by calculating the residual between the current diffusion state and the preset semantic trajectory.

A.3 Weighted Balance Analysis of Loss Functions

The full gradient operator $\nabla_\theta \mathcal{L}_{total}$ dictates the convergence of model parameters θ . This section provides a rigorous derivation of the gradient components and analyzes the regularization mechanism of the consistency constraint.

Linear Expansion of the Total Gradient. Following Eq. (5) in the main text, the total loss $\mathcal{L}_{total}$ is a weighted combination of task-specific components. The full gradient $\nabla_\theta \mathcal{L}_{total}$ is derived as:

$$\begin{aligned}
\nabla_\theta \mathcal{L}_{total} &= \nabla_\theta [\mathcal{L}_{inst}(\theta) + \lambda_1 \mathcal{L}_{ground}(\theta) \\
&\quad + \lambda_2 \mathcal{L}_{cons}(\theta)] \\
&= \frac{\partial \mathcal{L}_{inst}}{\partial \theta} + \lambda_1 \frac{\partial \mathcal{L}_{ground}}{\partial \theta} + \lambda_2 \frac{\partial \mathcal{L}_{cons}}{\partial \theta} \\
&= \mathbf{g}_{inst} + \lambda_1 \mathbf{g}_{ground} + \lambda_2 \mathbf{g}_{cons}
\end{aligned} \tag{13}$$

where $\mathbf{g}_{inst}$ ensures physical generation quality, while $\mathbf{g}_{ground}$ and $\mathbf{g}_{cons}$ regulate semantic consistency.

Manifold Regularization Derivation for $\mathcal{L}_{cons}$.

Under the Agnostic Masking Shift (AMS), let $\mathbf{R} = \mathcal{T}(\mathcal{O}_a; \theta) - \mathcal{T}(\mathcal{O}_b; \theta)$ denote the residual between semantic trajectories inferred from random subsets $\mathcal{O}_a$ and $\mathcal{O}_b$. The gradient chain is expanded as:

$$\begin{aligned}
\nabla_\theta \mathcal{L}_{cons} &= \nabla_\theta \|\mathcal{T}(\mathcal{O}_a; \theta) - \mathcal{T}(\mathcal{O}_b; \theta)\|_2^2 \\
&= \nabla_\theta \left[(\mathcal{T}_a - \mathcal{T}_b)^\top (\mathcal{T}_a - \mathcal{T}_b) \right] \\
&= 2 (\mathcal{T}_a - \mathcal{T}_b)^\top \nabla_\theta (\mathcal{T}_a - \mathcal{T}_b) \\
&= 2 \mathbf{R}^\top (\mathbf{J}_a(\theta) - \mathbf{J}_b(\theta))
\end{aligned} \tag{14}$$

where $\mathbf{J}(\theta) = \partial \mathcal{T} / \partial \theta$ is the Jacobian matrix with respect to model parameters. This derivation confirms that λ_2 forces the capture of modality-invariant intrinsic representations by penalizing parameter sensitivity to specific masking patterns via the residual $\mathbf{R}$.

A.4 Prompt Mapping and Reasoning Logic of MST

The MST module leverages the intrinsic logical priors of LLMs to infer modality-invariant representations. This section details the formal chain of equalities for transforming discrete perception symbols $\mathbf{s}$ into the continuous semantic trajectory $\mathcal{T}$.

Complete Chain of Prompt Construction. Let $\mathbf{s} = \{s_1, \dots, s_K\}$ be discrete symbols and $\mathbf{m} = \{m_1, \dots, m_K\}$ be the mask vector. The prompt generation follows:

$$\begin{aligned}
\text{Prompt} &= \mathcal{F}_{prompt}(\mathbf{s}, \mathbf{m}) \\
&= I_{sys} \oplus \text{Concat}(\{\text{Lookup}(s_i, m_i)\}_{i=1}^K) \\
&= I_{sys} + \sum_{i=1}^K \text{Sem}(s_i) \cdot m_i \\
&\quad + \text{Unk} \cdot \sum_{i=1}^K (1 - m_i)
\end{aligned} \tag{15}$$

where $\text{Sem}(\cdot)$ maps indices to textual semantics. This chain illustrates how MST preserves signal-level evidence while identifying segments for cross-modal inference.

Mathematical Instantiation of Trajectory. The final semantic trajectory $\mathcal{T}$ is derived via the fol-

lowing projection:

$$\begin{aligned}
\mathcal{T} &= \mathcal{G}(\text{LLM}(\text{Prompt})) \\
&= \mathcal{G}(\text{Softmax}(\mathbf{W}_h \mathbf{H}_{last} + \mathbf{b})) \\
&= \{\text{Proj}(h_1), \dots, \text{Proj}(h_K)\} \\
&= \{\tau_1, \dots, \tau_K\}
\end{aligned} \tag{16}$$

where h_j are hidden states of the LLM and $\text{Proj}(\cdot)$ is the operator projecting features into the semantic space. The sequence $\{\tau_j\}$ serves as the reference for drift correction in Eq. (12), ensuring inherent logical alignment.

Prompt Details. The *System Prompt* is defined as: “You are a multimodal logical expert. Task: Infer the intrinsic trajectory $\mathcal{T}$ based on incomplete tokens, ensuring cross-modal dependency consistency.”

B Extended Experiment Setup

B.1 Datasets Details

We validate our method on a set of visual and multimodal benchmarks. These datasets include simple image classification tasks and complex multimodal comprehension tasks. The specific details are as follows:

- **MNIST**(LeCun et al., 1998): A benchmark dataset which contains handwritten digit images from 0-9, including a total of 60,000 train examples and 10,000 test examples. It has been widely used in the evaluation of image classification as well as generative modeling approaches due to its simplicity and well-defined recognition task.
- **Hateful Memes**(Kiela et al., 2020): A multimodal dataset released by Facebook to detect hate speech on the Internet memes. Each document consists of an image, corresponding text and labels describing whether the sample is a hateful meme. More than 10,000 examples are included and require the model to have vision-language reasoning ability. It is a standard benchmark for models that need to learn from both visual and textual information in order to discriminate malicious content.
- **CMU-MOSEI**(Zadeh et al., 2018): A multimodal dataset used in sentiment analysis and emotion recognition tasks. It includes over 23,000 video clips, which are annotated with

sentiment analysis (positive, neutral, or negative) and emotion intensity scores. The dataset aligns three modalities: textual transcripts of speech, visual signals (facial expressions and gestures), and acoustic features (vocal tone and prosody). It is a comprehensive benchmark for evaluating models’ ability to fuse linguistic, visual, and audio information to interpret human sentiment and emotion.

B.2 Baselines and Adaptation

B.2.1 Baselines Overview

- **Robult**(Nguyen et al., 2025): A multimodal learning approach that leverages both redundant cross-modal information and modality-specific features to improve performance under missing-modality conditions. By explicitly separating shared information from modality-specific cues, Robult can maintain stable performance when input modalities are absent.
- **Memory-Driven Prompt**(Zhao et al., 2025): Memory-Driven Prompt Learning proposes a prompt-based method to enhance model’s robustness against the absence of some modalities by storing multimodal representation in an external memory module, then injecting learned prompts to guide the model with missing modalities. This helps the model to recall missing information at inference time, thus improving robustness against incomplete multimodal input.
- **StableMiss**(Zhu et al., 2023): Tailored for prediction with incomplete data under unknown missingness patterns. It mitigates distribution shifts in missing-data masks by guiding the model to leverage stable features that are informative even when parts of the input are absent. Using a specialized training objective, it achieves robust performance, ensuring reliable predictions across diverse missingness scenarios.
- **MD²N**(Dai et al., 2025a): Unbiased generative multimodal completion approach based on diffusion model which impute missing modalities. It adopts a two-stage diffusion process: the initial stage generates a grainy estimate of missing modality content, while the second stage refines this output to iteratively improve its fidelity. By modeling the

missing data distribution through diffusion, MD²N provides unbiased and diverse reconstructions that are consistent with the observed modalities.

- **Knowledge Bridger**(Ke et al., 2025): A training-free multimodal completion method that infers missing modal information using external knowledge. Instead of training task-specific models for each missing-modality scenario, it leverages pre-trained model or external knowledge base as "bridge" for data imputation, using the prior knowledge to complete multimodal input without additional specific training.
- **SMCMSA**(Sun et al., 2024): A similarity-based sentiment analysis method tailored for uncertain missing-modality scenarios. Its core idea lies on exploiting one modality to simulate another one. By completing the missing modality with correlated information from available ones, SMCMSA exhibits strong performance on sentiment prediction.
- **IMDer**(Wang et al., 2023c): An emotion recognition framework that integrates a diffusion mechanism to handle missing modalities. It employs a diffusion model in the latent feature space to propagate cross-modal information, effectively inferring missing modal content based on available ones. This diffusion-based imputation allows the model to capture the underlying emotional context from partial inputs and improves emotion recognition accuracy when modalities are absent.
- **DiC-MoR**(Wang et al., 2023b): A multimodal missing modality recovery method which makes the recovered data are statistically consistent with the true data distribution. It is trained to reconstruct missing modality features whose distributions are similar as observed ones. By imposing distributional consistency on the imputed modalities, DiC-MoR mitigates biases induced by naive imputation, leading to more reliable multimodal prediction when some inputs are missing.
- **HME**(Zhuang et al., 2025): A hyper-modality enhancement framework designed to capture high-order correlations among modalities for missing-modality scenarios. Unlike traditional graph-based methods that rely on pair-

wise interactions, HME constructs a dynamic hypergraph to model complex, non-linear relationships between audio, visual, and textual features. This structure allows the model to effectively infer and enhance missing modality representations by propagating information through high-order connections from available data, ensuring robust sentiment analysis performance.

- **CAPIMAC**(Ma et al., 2025): A consistency-aware padding method for incomplete and misaligned multimodal clustering. It uses self-repellent greedy anchor search to identify anchor samples and applies consistency-aware padding to fill incomplete multimodal data before downstream alignment or classification.

B.2.2 Baseline Adaptation

For experimental rigor, baseline models are implemented using official releases or high-fidelity re-implementations following the original specifications. We strictly adhere to the hyperparameter configurations and optimization settings recommended in the original papers. To eliminate performance discrepancies stemming from feature extraction, all baselines utilize the same pre-extracted input features (e.g., CLIP, RoBERTa, BERT, COVAREP, Facet) and linear projection layers to a unified latent space ($d = 512$) as described in the Implementation Details section, ensuring a fair comparison of their respective reasoning and completion mechanisms.

(1) Adaptation of Non-Generative Methods.

These methods emphasize robustness through architectural design rather than explicit reconstruction.

Robult and **StableMiss**: These architectures employ native multi-branch processing. When evaluating on datasets with fewer modalities (e.g., MNIST, Hateful Memes), we adopt a branch deactivation strategy—only branches corresponding to available modalities are activated, while redundant ones are suppressed. This modular approach ensures that the core alignment and discriminative logic remain consistent with their original 2- or 3-modal formulations.

Memory-Driven Prompt: This model leverages memory-enhanced prompting to compensate for missing information under incomplete multimodal settings. For fair comparison, we follow its original prompting mechanism and adapt only the input projection layer when the feature dimensionality

differs across datasets. The core memory-driven prompt construction and fusion strategy are kept unchanged.

HME: HME is originally designed for multimodal sentiment analysis with incomplete audio, visual, and textual modalities. For adaptation to different modality configurations, we instantiate only the modality-specific enhancement branches corresponding to the available modalities in each dataset. When the target dataset contains fewer modalities, unavailable branches are removed rather than filled with artificial inputs. The high-order enhancement and uncertainty-aware fusion modules are kept unchanged, so that the model preserves its original mechanism for exploiting cross-modal dependencies under missing-modality conditions.

(2) Adaptation of Generative and Completion Methods. These models focus on reconstructing missing information either at the signal or latent level.

MD²N and **IMDer**: These models are designed as multimodal completion frameworks and support parallel processing of multiple streams, such as $\{A, V, T\}$. Their generative or reconstruction modules are modality-agnostic, allowing them to be adapted to different modality configurations by changing only the input and output projection layers. In single-modality scenarios such as MNIST, they operate in a degenerate single-branch mode, where the reconstruction objective is applied directly to the observed signal space without modifying the core completion mechanism.

CAPIMAC: CAPIMAC is originally proposed for incomplete multi-modal alignment and completion through consistency-aware padding. To adapt it to our evaluation setting, we use the same pre-extracted modality features and project them into the unified 512-dimensional latent space before applying its padding and alignment modules. For datasets with different modality counts, the padding module is instantiated only for the missing modalities defined by the corresponding masking protocol. The consistency-aware alignment objective is retained, allowing CAPIMAC to reduce the discrepancy between incomplete and completed multimodal representations without changing its core completion mechanism.

Knowledge Bridger: This model employs large language models (LLMs) as a universal semantic relay. For modalities beyond its original dual-modal scope, we follow the “knowledge bridging”

paradigm: extra-modal features are converted into textual attribute descriptions and injected into LLM prompts. This enables natural semantic fusion in the latent space without retraining.

SMCMSA and **DiC-MoR**: These frameworks address sequential alignment and cross-modal consistency. We adapt the input channels of their alignment layers to match the modality count of the target dataset. This dimensional self-adaptation is a standardized configuration that maintains performance stability across diverse multimodal combinations.

B.3 Implementation Details

B.3.1 Network Specifications

Perception encoders project multimodal inputs into a 512-dimensional unified latent space. We utilize 1024-d CLIP and 768-d RoBERTa features for Hateful Memes, and 768-d BERT, 74-d COVAREP, and 35-d Facet features for CMU-MOSEI. The MST module employs a **6-layer Transformer** encoder (8 heads, $d_{ff} = 2048$, dropout=0.1) on top of the discrete codebook ($V = 1024$). The HySSM control module uses a 64-dimensional latent control state. Latent features are normalized before diffusion conditioning to improve numerical stability.

Models are trained for 100 epochs with $r = 0.5$ and a batch size of 64. We use **AdamW** ($LR = 1 \times 10^{-4}$) with a weight decay of 1×10^{-2} and gradient clipping at 1.0. For the Llama-3-8B (Grattafiori et al., 2024) backbone in the MST module, we apply a separate **QLoRA** warm-up stage with a rank of 16 and alpha of 32 to align soft evidence tokens to the LLM input space. During the main HySSM optimization, the LLM planner is frozen; gradients update the semantic encoder, modality-specific emitters, and lightweight adapters where applicable. Following Eq. (5), the loss coefficients are set to $\lambda_1 = 0.5$ and $\lambda_2 = 0.1$ to balance signal-level alignment and logical consistency.

Table 6: Detailed hyperparameters for our HySSM.

Parameter	Value	Notes
Optimizer	AdamW	Weight decay 1×10^{-2}
Batch Size	64	Distributed training
HySSM State (N)	64	Latent dimensionality
λ_1, λ_2	0.5, 0.1	Grounding & consistency
QLoRA Rank	16	Parameter-efficient tuning
DDIM Steps	50	Signal reconstruction

B.3.2 Data Preprocessing and Evaluation

Heterogeneous signals undergo **Z-score normalization** to ensure signal-level consistency. For CMU-MOSEI, COVAREP features are extracted with a 20ms frame shift. Textual tokenization is performed via BPE with a sequence limit of 128. Testing is conducted across the full range $r \in [0.1, 0.9]$, and all results are reported as the average of five independent runs to ensure representational resilience.

B.4 Formalization of Missing Scenarios

Signal-level and Modality-level Masking. The experimental protocol defines two distinct degradation modes to probe the robustness frontier of HySSM. **Signal-level degradation** (e.g., MNIST) is operationalized via grid-based patch masking, simulating packet loss or localized sensor corruption. **Modality-level missingness** (e.g., Hateful Memes and CMU-MOSEI) involves the asymmetric dropout of entire feature streams, mimicking sensor offline scenarios or privacy-restricted omissions. These protocols collectively construct a fragmented observation space, necessitating the model to circumvent superficial statistical correlations in favor of intrinsic logical dependencies.

Agnostic Masking Shift (AMS) Construction.

To quantify the out-of-distribution (OOD) generalization, we induce a spectral shift between training and evaluation masking distributions. During training, the missing rate is fixed at $r = 0.5$ to expose the model to the state space with the highest combinatorial entropy, thereby maximizing the learning of intrinsic logic. In the testing phase, the rate r is dynamically shifted across $[0.1, 0.9]$. This shift validates the efficacy of the MST module in steering signal manifold regression via logical priors when navigating beyond the empirical distribution of the training set.

C Additional Experimental Results

Table 7: Ablation study on the Sample-and-Reweigh strategy. Values averaged across $r \in [0.1, 0.9]$. **Bold** denotes best.

Variants	MNIST	Hateful Memes	CMU-MOSEI
	(RMSE / Acc)	(AUROC / Acc)	(F1 / Acc)
(i) Weighted Integration (Ours)	0.117 / 89.0	68.0 / 63.2	69.5 / 72.1
(ii) Uniform Mean	0.121 / 88.3	66.5 / 61.9	68.2 / 70.8
(iii) Single Best	0.130 / 87.2	64.8 / 60.2	66.4 / 69.1
(iv) Argmax Generation	0.142 / 85.4	62.1 / 58.5	64.5 / 66.8

Validation of Sample-and-Reweigh Strategy.

We evaluate the bypass mechanism for the non-differentiable bottleneck by comparing four strategies: (i) **Weighted Integration (ours)**; (ii) **Uniform Mean**; (iii) **Single Best**; and (iv) **Argmax Generation**. As shown in Table 7, Weighted Integration effectively outperforms other variants; for instance, it surpasses Argmax Generation by 5.0% in F1 score on the complex CMU-MOSEI dataset. Analysis reveals that our strategy smooths the variance of discrete sampling and captures higher posterior diversity via weights ω_n . In contrast, variants (iii) and (iv) often fall into low-probability generation traps by ignoring the tail of the posterior distribution, resulting in a significant performance slump in AUROC on Hateful Memes. This validates the importance of the reweighting mechanism for optimizing the end-to-end semantic encoder under multimodal uncertainty.

Table 8: Ablation study on the training masking strategy. Values averaged across $r \in [0.1, 0.9]$. **Bold** denotes best.

Variants	MNIST	Hateful Memes	CMU-MOSEI
	(RMSE / Acc)	(AUROC / Acc)	(F1 / Acc)
(i) Dynamic Randomized (Ours)	0.117 / 89.0	68.0 / 63.2	69.5 / 72.1
(ii) Fixed Pattern Training	0.128 / 87.5	64.1 / 58.2	65.2 / 67.5
(iii) No Masking	0.165 / 80.2	59.8 / 52.1	61.4 / 63.6

Robustness of Training Masking Strategy.

We compare different masking schemes to verify their impact on model robustness, including: (i) **Dynamic Randomized Masking (ours)**; (ii) **Fixed Pattern Training**; and (iii) **No Masking**. As shown in Table 8, our strategy achieves the most stable performance across random missing scenarios. In contrast, training without masking leads to a sharp decline in F1 score on CMU-MOSEI (from 69.5% to 61.4%), indicating that models trained only on complete data fail to generalize to incomplete test distributions. These results suggest that dynamically partitioning $\mathcal{O}_i$ and $\mathcal{T}_i$ is not merely a data augmentation technique but forces the model to learn deep inter-modal dependencies rather than surface mappings. In the Hateful Memes task, the robustness gained from dynamic masking is particularly evident, with an AUROC improvement of approximately 3.9% over Fixed Pattern Training, further validating the importance of randomized diversity in building robust multimodal representations.