MUNI: Multimodal Unified Latent Diffusion for Coherent Any-to-Any Generation

Kyeongmin Yeo* **Yunhong Min*** **Minhyuk Sung**
KAIST
{aaaaa, dbsghd363, mhsung}@kaist.ac.kr

Abstract

We introduce MUNI, an end-to-end multimodal latent diffusion framework for any-to-any generation that unifies subset-conditioned cross-modal generation and unconditional joint sampling through a shared stochastic latent. Existing multimodal generative models are largely LLM-based, which limits leveraging modality-specific generators and requires text-paired data for training. Recent diffusion- and flow-based any-to-any extensions take a different direction but still rely on text-aligned embeddings, fully-paired training, or matched-dimensionality deterministic mappings. MUNI rests on two complementary contributions, one architectural and one in the training objective. First, we extend latent diffusion to multimodal any-to-any generation *end-to-end*: instead of the standard two-stage recipe that precomputes a frozen latent space and then fits a prior over it, MUNI jointly trains modality-specific encoders, expressive decoders, and a single shared flow-based prior under one objective. Second, we identify that the standard aggregation rules of multimodal variational inference are insufficient once coupled with a learned prior and expressive decoders. A suitable shared latent must simultaneously satisfy *coherence* across generated modalities, *predictive sufficiency* of subset latents, and *minimality* of the latent content. We propose a routed training objective whose structural choices align the latent with these criteria and admit a minimal-sufficiency characterization in the realizable setting. Experiments on PolyMNIST-Quadrant-Labels and a large-scale image-text-audio benchmark show MUNI matching or exceeding the strongest baselines on conditional generation while opening its largest margins on unconditional coherence. Project page: muni-proj.github.io.

1 Introduction

Diffusion- and flow-based models [20, 47, 33, 25] have become the dominant paradigm for generative modeling across many modalities and have set the state of the art in cross-modal conditional generation, including text-to-image [44, 1], text-to-audio [14, 23, 34], and image-to-video [6, 58]. Our key question is how to move beyond these one-to-one settings: how can the capability of diffusion- and flow-based models be effectively extended to *any-subset-conditioned* multimodal generation?

Concretely, given a set of data modalities, we consider a model that takes any subset of those modalities as input and generates the remaining ones. The conditioning subset ranges from the empty set—corresponding to unconditional joint generation of all modalities—to any proper subset of the full collection.

Despite the emergence of large foundational multimodal generative models [4, 36, 60, 56], most such models adopt large language models as their backbone, representing every modality as discrete tokens in a unified sequence. This design makes it difficult to leverage modality-specialized generative

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

architectures and pretrained generators—such as latent diffusion for images [44, 13] and flow matching for audio [23, 14, 34]—which operate in continuous spaces and currently lead in their respective modalities. It also ties each modality to what can be expressed through language, limiting both the fidelity of per-modality representations and the model's ability to capture direct cross-modal correlations that do not pass through text. Finally, this paradigm requires every training example to be paired with text, which is not the case for many natural multimodal sources, such as uncaptioned audio-visual recordings [7, 16] or scientific multimodal measurements [40, 5].

Following this spirit, recent work has explored diffusion- and flow-based models for multimodal generation, demonstrating the promise of this direction [50, 31, 3]. However, these approaches still have limitations: relying on text-aligned embedding spaces that may lose modality information not expressed in language [50]; requiring conditional paths to be learned for every subset configuration and depending on fully-paired data for stable training [31]; or imposing matched dimensionality across modalities and learning deterministic source-to-target mappings, despite the non-unique correspondences of multimodal data [3].

In this work, we propose a novel diffusion-based framework for any-to-any multimodal generation that does not rely on text-aligned embeddings, fully-paired training data, or pre-embedding for dimensionality matching. Our key ideas are twofold: (1) we extend latent diffusion [44] to the multimodal setting by jointly learning the unified prior together with all modality-specific components; and (2) we propose a novel training objective for the underlying variational inference component, addressing the limitations of existing aggregation rules when coupled with a diffusion or flow-based prior and expressive decoders.

Latent diffusion is the de facto standard for image and video generation [44, 13], but it follows a two-stage recipe in which a latent space is fixed by a pretrained encoder *before* a diffusion prior is fitted over it. Recent work [17] departs from this separation by training the latent prior *jointly* with the encoder and decoder under a single objective, so that the latent space is shaped by both reconstruction and prior modeling at once. Our *first contribution* extends this jointly trained framework to multimodal any-to-any generation by introducing modality-specific encoders, modality-specific expressive decoders, and a single unified flow-based prior shared across modalities.

Under this extension, jointly learning the latent embedding and the decoders aligns with multimodal variational inference, a literature that has primarily studied how unimodal posterior experts should be aggregated, including product of experts (PoE) [55], mixture of experts (MoE) [46], mixture of products of experts (MoPoE) [49], and their variants [52]. Our *second contribution* identifies that aggregation rules alone are insufficient once coupled with a diffusion or flow-based prior and expressive decoders, and introduces a novel training objective designed around three latent-space criteria: (i) *coherence-sufficiency* for unconditional co-generation, (ii) *predictive sufficiency* for cross-modal generation, and (iii) *minimality* of the latent content. Existing objectives do not enforce these criteria jointly: latents inferred from partial observations may support conditional generation, but need not preserve dependencies among unobserved modalities required for unconditional co-generation, while self-reconstruction rewards encoding modality-private details that need not belong to the shared latent. `MUNI` addresses these failures through non-mixture aggregation, target-detached self-reconstruction, and prior learning on leave-one-out latents.

We validate `MUNI` on a controlled PolyMNIST-Quadrant-Labels benchmark and a large-scale image-text-audio any-to-any benchmark. Across both, `MUNI` matches or exceeds the strongest multimodal-VAE and any-to-any generalist baselines on conditional generation while opening its largest margins precisely where our analysis predicts: on unconditional coherence. On image-text-audio, `MUNI` achieves the strongest generalist many-to-one performance, with audio-image alignment gains of up to $+7.14$ AIS over the next-best generalist, and improves unconditional co-generation over the only directly comparable generalist by $+5.57$ CLIP, $+10.60$ CLAP, and $+31.85$ AIS. These results show that `MUNI` scales to large image-text-audio any-to-any generation while retaining strong coherence against frontier generalist baselines.

2 Background and Problem Setup

2.1 Multimodal Generation with Unified Latents

Let $X_{1:M} = (X_1, \ldots, X_M)$ denote the random variables associated with M modalities, with joint data distribution $p_{\text{data}}(x_{1:M})$. Multimodal generation requires a model to represent the family of

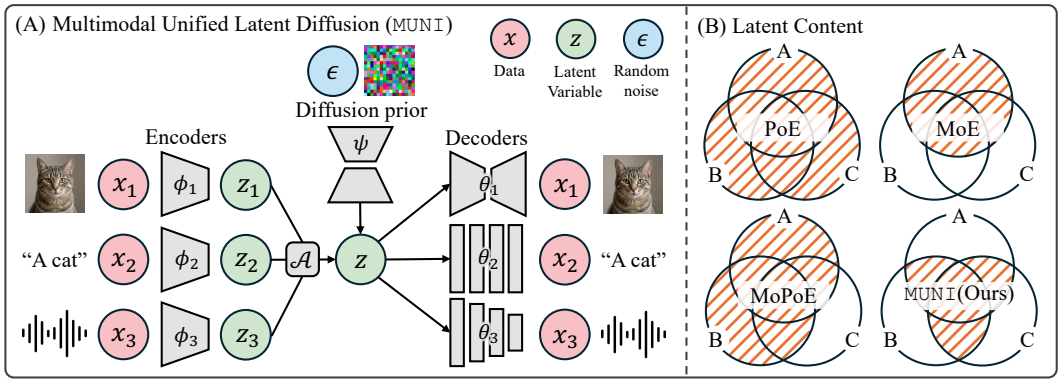


Figure 1: **Overview of MUNI and latent content. (A)** MUNI uses modality-specific encoders and decoders with a shared latent z and a learned prior. **(B)** Schematic of the modality-shared and modality-specific factors retained in the latent under PoE, MoE, MoPoE, and MUNI.

conditional distributions

$$p_{\text{data}}(x_{S^c} \mid x_S), \qquad S \subseteq [M], \tag{1}$$

where S is the observed subset and $S^c = [M] \setminus S$ is the set of target modalities. Given x_S, the model should sample x_{S^c} jointly, preserving consistency with the observed modalities. Especially, the case $S = \emptyset$ corresponds to unconditional co-generation, in which the model samples coherent tuples from the full joint distribution $p_{\text{data}}(x_{1:M})$.

Recent diffusion- and flow-based any-to-any models extend continuous generative modeling to multimodal generation [50, 31, 3]. Existing approaches, however, often rely on text-aligned conditioning representations [50], require learning conditional paths across all conditioning-target modality subsets rather than inferring a shared latent variable [31], or use invertible flows that require matched dimensionality and impose deterministic one-to-one correspondence between the latent and each modality [3]. This motivates a probabilistic shared-latent framework that can be inferred from arbitrary observed subsets, decoded stochastically into each modality, and sampled from a learned prior for unconditional co-generation.

Latent diffusion models use a learned latent space for high-fidelity generation. They move diffusion modeling from data space to latent space, where a prior models latent samples and a decoder maps them back to data. In the standard two-stage pipeline, a variational autoencoder is trained first, and a diffusion prior is then fitted to the frozen latent distribution. Unified Latents (UL) [17] instead jointly trains the encoder $\mathcal{E}_\phi(z_0 \mid x)$, latent prior $q_\psi(z_0)$, and decoder $p_\theta(x \mid z_0)$:

$$\mathcal{L}_{\text{UL}}(\phi, \psi, \theta) = \mathbb{E}_{x \sim p_{\text{data}}} \mathbb{E}_{z_0 \sim \mathcal{E}_\phi(\cdot \mid x)} \left[\mathcal{L}_x(\theta; x, z_0) + \beta \mathcal{L}_z(\psi; z_0) \right], \tag{2}$$

where $\mathcal{L}_z$ and $\mathcal{L}_x$ are the diffusion evidence lower bound (ELBO) terms for the latent prior and conditional decoder, respectively. By optimizing reconstruction and prior modeling in the same latent space, UL avoids treating the autoencoder representation as a fixed target for the prior. The latent is therefore shaped not only for accurate decoding, but also to remain tractable for the prior to model.

We extend this idea to multimodal generation by introducing subset encoders and modality-specific expressive decoders. Specifically, for each non-empty subset $A \subseteq [M]$, we introduce a subset encoder $q_{\phi,A}(z \mid x_A)$, a shared learned prior $q_\psi(z)$, and modality-specific decoders

$$p_\theta(x_{1:M} \mid z) = \prod_{m=1}^{M} p_{\theta_m}(x_m \mid z). \tag{3}$$

At inference time, conditional generation from S is represented by

$$p_{\theta,\phi}(x_{S^c} \mid x_S) = \int q_{\phi,S}(z \mid x_S) \prod_{m \in S^c} p_{\theta_m}(x_m \mid z)\, dz, \tag{4}$$

while unconditional co-generation is represented by

$$p_{\theta,\psi}(x_{1:M}) = \int q_\psi(z) \prod_{m=1}^{M} p_{\theta_m}(x_m \mid z)\, dz. \tag{5}$$

Table 1: Posterior aggregation rules and their alignment with the latent-space criteria in Sec. 3. Coh., Pred., and Min. denote coherence, predictivity, and minimality, respectively. Here, $q_{\mathrm{PoE}}(z \mid x_A) = \prod_{m \in A} q_{\phi_m}(z \mid x_m)$.

Method	Aggregation (Training)	Aggregation (Inference)	Coh.	Pred.	Min.
MVAE (PoE) [55]	$\prod_{m\in[M]} q_{\phi_m}(z \mid x_m)$	$\prod_{m\in S} q_{\phi_m}(z \mid x_m)$	✓	△	×
MMVAE (MoE) [46]	$\frac{1}{M}\sum_{m\in[M]} q_{\phi_m}(z \mid x_m)$	$\frac{1}{\lvert S\rvert}\sum_{m\in S} q_{\phi_m}(z \mid x_m)$	×	×	×
MoPoE [49]	$\frac{1}{2^M-1}\sum_{A\in\mathcal{P}^*([M])} q_{\mathrm{PoE}}(z \mid x_A)$	$\prod_{m\in S} q_{\phi_m}(z \mid x_m)$	×	✓	×
HELVAE [52]	$\left(\sum_{m\in[M]} \sqrt{q_{\phi_m}(z \mid x_m)}\right)^2$	$\left(\sum_{m\in S} \sqrt{q_{\phi_m}(z \mid x_m)}\right)^2$	✓	△	×

The same latent variable therefore mediates both subset-conditioned generation and unconditional joint sampling, while expressive decoders can model modality-specific variation. Fig. 1 (A) illustrates this construction.

A direct multimodal UL objective trains every subset latent to decode all modalities and uses the same latents to supervise the prior:

$$\mathcal{L}_{\mathrm{MMUL}} = \mathbb{E}_{x_{1:M}\sim p_{\mathrm{data}}} \sum_{A\in\mathcal{P}^*([M])} w_A \mathbb{E}_{z_A\sim q_{\phi,A}(\cdot\mid x_A)} \left[\sum_{m=1}^{M} \mathcal{L}_x^m(\theta_m; x_m, z_A) + \beta \mathcal{L}_z(\psi; z_A) \right], \tag{6}$$

where $\mathcal{P}^*([M]) = \{A : A \subseteq [M], A \neq \emptyset\}$ and w_A is a subset sampling weight. This objective is a natural multimodal analogue of UL. However, applying the prior loss to every subset fits the prior to a mixture of subset latents, including latents inferred from insufficient evidence to determine dependencies among the missing modalities. Such prior samples need not yield coherent unconditional co-generation. In addition, self-reconstruction terms for $m \in A$ can encourage the shared latent to store modality-specific details that are unnecessary for cross-modal generation.

2.2 Connection to Multimodal Variational Inference

This multimodal factorization based on unified latents closely resembles the latent-variable structure of multimodal VAEs [55, 46, 49, 38, 52, 9]. Both families introduce subset posteriors $q_{\phi,A}(z \mid x_A)$ for partial observations and decode a shared latent through modality-specific likelihood factors. Most multimodal VAE formulations, however, rely on a fixed Gaussian prior and relatively simple likelihood decoders, which limits generation fidelity. Multimodal UL instead retains the learned latent prior and expressive conditional decoders of the latent-diffusion setting.

In multimodal variational inference, subset posteriors are parameterized by aggregating unimodal posterior experts $q_{\phi_m}(z \mid x_m)$. Tab. 1 summarizes representative aggregation rules. The next section analyzes how these choices shape the induced subset latents, which properties are required for coherent multimodal generation, and why aggregation alone does not specify which latents should serve as targets for prior learning.

3 Latent-Space Criteria for Coherent Multimodal Generation

The direct objective in Eq. 6 and the aggregation rules in Tab. 1 both induce subset latents, but do not determine what information those latents should carry. Under the factorized decoder in Eq. 3, modalities are decoded independently conditioned on z, so cross-modal dependence must be represented in the latent. This yields three latent-space criteria: prior targets must support coherent joint sampling, subset latents must predict missing modalities while explaining dependencies within the observed subset, and the shared latent should avoid modality-private variation that expressive decoders can model.

Coherence-sufficient latents. For unconditional co-generation, factorized decoding preserves coherence only when the latent explains the dependencies among modalities. We measure the residual dependence by conditional total correlation and require

$$\mathrm{TC}(X_{1:M} \mid Z) = 0. \tag{7}$$

We call such a latent *coherence-sufficient*: conditioned on it, factorized decoding does not discard cross-modal dependence.

Predictively sufficient subset latents. For conditional generation from X_A, the subset latent Z_A must retain the information in X_A needed to predict each missing modality, while also explaining dependencies within the observed subset:

$$\mathrm{TC}(X_A \mid Z_A) = 0, \tag{8}$$

$$I(X_A; X_m \mid Z_A) = 0 \qquad \forall m \in A^c. \tag{9}$$

The mutual-information condition states that Z_A is a sufficient statistic of X_A for each target modality X_m. The total-correlation condition requires the latent to capture dependencies within the observed subset, which is necessary when such latents are used as prior targets for coherent joint sampling.

Minimal latent content. The preceding conditions specify the information that the latent must preserve for coherence and conditional alignment. Among the latents that satisfy these conditions, however, minimizing the information retained from X_A yields a simpler latent marginal for the prior to model and prevents the shared latent from absorbing modality-specific variation. The desired subset latent is therefore

$$Z_A^\star = \arg \min_{\substack{Z_A:\, \mathrm{TC}(X_A|Z_A)=0 \\ I(X_A;X_m|Z_A)=0\ \forall m\in A^c}} I(X_A; Z_A). \tag{10}$$

This criterion imposes a separation of roles: the shared latent captures the information needed to relate modalities, while modality-specific variation is modeled by $p_{\theta_m}(x_m \mid z)$.

Fig. 1 (B) visualizes the shared and modality-specific factors retained by PoE, MoE, MoPoE, and `MUNI`, while Tab. 1 summarizes their alignment with the three criteria. PoE and HELVAE aggregate all observed experts into a single posterior, making them structurally compatible with the coherence criterion. Their predictivity is only partial, however, because the standard objective does not explicitly train each subset latent to predict held-out modalities. In contrast, mixture-based aggregation, as in MMVAE, can draw the latent from a single expert, so the resulting representation may not contain enough information for either conditional alignment or coherent joint decoding. MoPoE differs from MMVAE at inference time: although it is trained with a mixture over subset products, it uses the full product posterior for conditional generation, which satisfies the predictivity requirement. Its prior, however, is still trained on mixture components that include smaller-subset latents, which need not be valid targets for co-generation. Finally, none of these aggregation rules enforces minimality, since self-reconstruction encourages the shared latent to store modality-specific details. Thus, aggregation specifies how evidence is combined, but not which subset latents are valid targets for prior learning or how modality-private information should be excluded from the shared latent.

These limitations motivate the objective introduced in Sec. 4, while Apps. A and B provide further analysis of the prior-learning targets and the corresponding latent-space constraints.

4 Multimodal Unified Latent Diffusion (`MUNI`)

`MUNI` modifies the direct multimodal UL objective in Eq. 6 according to the latent-space criteria in Sec. 3. For each modality m, we use a unimodal expert encoder $q_{\phi_m}(z \mid x_m)$ and an expressive decoder $p_{\theta_m}(x_m \mid z)$. For a non-empty subset A, the subset posterior is obtained by applying a non-mixture aggregation rule to the available unimodal experts: $q_{\phi,A}(z \mid x_A) = \mathcal{A}(\{q_{\phi_j}(z \mid x_j)\}_{j\in A})$, where $\mathcal{A}_A$ is either product or Hellinger aggregation. Unlike mixture-based aggregation, this forms a single posterior from all observed modalities.

For a reconstruction target m, we use a target-dependent subset posterior. If $m \in A$, the target expert is detached before aggregation; otherwise the usual subset posterior is used:

$$\bar{q}_{\phi_j}^{(m)}(z \mid x_j) = \begin{cases} \mathrm{sg}\left[q_{\phi_j}(z \mid x_j)\right], & j = m, \\ q_{\phi_j}(z \mid x_j), & j \neq m, \end{cases} \qquad q_{\phi,A}^{(m)}(z \mid x_A) = \mathcal{A}\left(\{\bar{q}_{\phi_j}^{(m)}(z \mid x_j)\}_{j\in A}\right), \tag{11}$$

where $\mathrm{sg}[\cdot]$ is the stop-gradient operator. Thus, self-reconstruction still trains the decoder, but does not reward the target encoder for transmitting modality-specific details.

The training objective is

$$\mathcal{L}_{\text{ours}} = \mathbb{E}_{x_{1:M}\sim p_{\text{data}}} \sum_{A\in\mathcal{P}^*([M])} w_A \Bigg[\sum_{m=1}^{M} \mathbb{E}_{z\sim q^{(m)}_{\phi,A}(\cdot|x_A)} \mathcal{L}^m_x(\theta_m; x_m, z) + \beta\, \mathbf{1}[|A| \geq M-1] \mathbb{E}_{z\sim q_{\phi,A}(\cdot|x_A)} \mathcal{L}_z(\psi; z) \Bigg]. \tag{12}$$

The reconstruction term covers both held-out prediction and target-detached self-reconstruction, while the prior term is applied only to full-modality and leave-one-out subset latents. In practice, Eq. 12 is optimized by Monte Carlo sampling subsets and reconstruction targets.

Proposition 4.1 (Minimal sufficient latent, informal)**.** *Under realizability and optimization assumptions, Eq. 12 selects subset latents that satisfy predictive sufficiency while approaching the minimality criterion in Eq. 10. On the subset routes used for prior learning, these latents are coherence-sufficient and therefore provide valid targets for the learned prior.*

Apps. A and B detail the objective design and provide a formal version and proof of Proposition 4.1. App. C gives the concrete prior loss $\mathcal{L}_z$ and the modality-specific reconstruction losses $\mathcal{L}^m_x$ used in our experiments.

At inference time, the same encoders and decoders are used for every conditioning subset. Given an observed subset S, we sample a latent vector from the subset posterior $z \sim q_{\phi,S}(z \mid x_S)$ and generate each target modality from $p_{\theta_m}(x_m \mid z)$ for $m \in S^c$. When $S = \emptyset$, we instead sample $z \sim q_\psi(z)$ and decode all modalities.

5 Related Work

Any-to-any generation. Recent diffusion- and flow-based frameworks support generation beyond fixed modality pairs. CoDi [50] composes modality-specific diffusion models through a shared text embedding space, enabling flexible input-output combinations but relying on a representation that may lose information not captured by text. Its separately trained encoders and decoders can also weaken input-output alignment. OmniFlow [31] formulates any-to-any generation with rectified flows over conditioning-target modality subsets, but learning paths across all subset configurations increases data and training demands. FlowBind [3] maps each modality to a shared space through invertible flows. This requires all modalities to have the same dimensionality, and generation is deterministic given the source modality rather than sampled from a conditional distribution. `MUNI` instead learns a shared latent variable: subset posteriors are used for conditional generation, modality decoders model target variation, and a prior model is used for unconditional co-generation.

Native multimodal generation. Native multimodal approaches generate heterogeneous modalities within a single model, keeping each modality in its own state space. Multiflow [2] combines discrete and continuous flows for protein sequence-structure co-design, Diffuse Everything [43] jointly diffuses discrete text and continuous images, and Transfusion [62] combines text autoregression with image diffusion in one transformer. These models learn cross-modal dependencies from samples in which all modalities are observed together. `MUNI` couples modalities through a shared latent inferred from any observed subset, so it learns to generate image-text-audio triplets from pairwise data alone.

Multimodal variational inference. Multimodal VAEs model multiple modalities through a shared latent variable and modality-specific decoders, with the main inference question being how to infer the latent from arbitrary observed subsets. MVAE [55] uses product-of-experts aggregation, MMVAE [46] uses mixture-of-experts aggregation, and MoPoE [49] mixes products over modality subsets. Prior analyses show that these choices induce trade-offs among joint generation, cross-modal prediction, and generation quality [10]. We use this literature as a reference for posterior aggregation, but build on the unified-latent formulation with a learned prior and expressive decoders, rather than a standard multimodal VAE with a fixed Gaussian prior and simple likelihood decoder. Moreover, prior multimodal VAE studies have largely focused on controlled or relatively simple benchmarks; to our knowledge, `MUNI` is the first to scale this probabilistic framework to practical image-text-audio any-to-any generation.

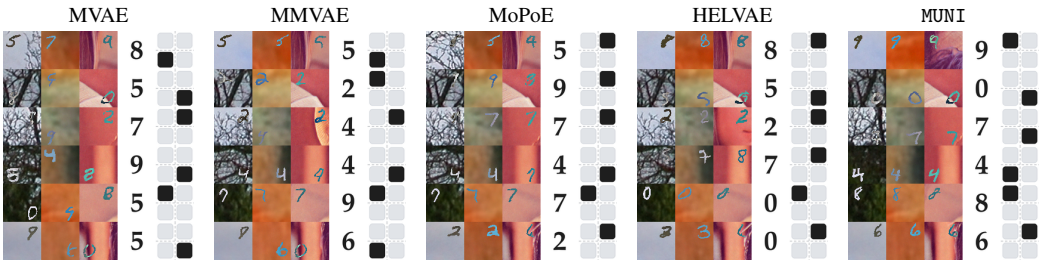


Figure 2: **Unconditional co-generation on PolyMNIST-Quadrant-Labels.** Each row within a method panel shows one tuple $(M_1, M_2, M_3, \text{digit}, \text{quadrant})$ jointly sampled from its prior. Coherent samples preserve digit identity and quadrant position across the generated image and label modalities.

Private latents and expressive models. Another line of work separates shared and modality-specific information with private latent variables, as in DMVAE [11], MMVAE+ [38], MMVAE++ [37], and IDMVAE [61]. These private latents are natural choices when decoders cannot represent target-modality variation without encoding it explicitly. Other methods improve multimodal VAEs with stronger generative components, such as a second-stage diffusion prior in ShaLa [9], diffusion decoders in Diff-MVAE [53], an energy-based prior in MVEBM [59], or mixture-structured latent spaces in CMVAE [39]. These extensions improve capacity, but do not specify which subset latents should train the prior or how self-reconstruction should affect the encoder and the corresponding latent. `MUNI` keeps a single shared latent, assigns modality-specific variation to expressive decoders, and uses the criteria in Sec. 3 to determine prior learning and self-reconstruction routes.

6 Experiments

In this section, we evaluate `MUNI` on two main benchmarks. We first describe the baselines and shared experimental setup in Sec. 6.1. We then evaluate on a standard PolyMNIST-based benchmark with mixed image and label modalities in Sec. 6.2, followed by a large-scale image-text-audio benchmark in Sec. 6.3. In both settings, we interpret the results through the latent-space criteria in Sec. 3, testing whether the observed patterns support our analysis and objective design. App. D further provides a controlled Gaussian-mixture experiment with known shared-private structure, making the predicted baseline failure modes and the role of our proposed design more explicit.

6.1 Common Setups

Baselines. Across all benchmarks, we compare against four representative multimodal VAE baselines that span the dominant choices of posterior aggregation: MVAE [55] (product of experts, PoE), MMVAE [46] (mixture of experts, MoE), MoPoE [49] (mixture of products of experts, generalizing PoE and MoE), and HELVAE [52] (Hellinger aggregation, a symmetric interpolation between PoE and MoE). For MVAE, we also train partially observed PoE posteriors with subset-level terms, matching the role of subset ELBOs in the original formulation. On the image-text-audio benchmark in Sec. 6.3, we additionally compare against three recent diffusion- and flow-based any-to-any generalists: CoDi [50], OmniFlow [31], and FlowBind [3].

Implementation. The original multimodal VAE baselines are typically implemented with a fixed Gaussian prior and deterministic likelihood decoders. For a fair comparison, we instantiate all methods under the same architecture and generative framework as `MUNI`. Specifically, all methods use a learned flow-based prior and expressive modality decoders, so that performance differences primarily reflect aggregation and prior-routing choices rather than model capacity. Full details about training, evaluation, and architecture are provided in App. E.

6.2 PolyMNIST-Quadrant-Labels

Setup. We construct PolyMNIST-Quadrant-Labels from PolyMNIST-Quadrant [61] by aligning the quadrant position across image modalities and adding digit and quadrant labels as discrete modalities. Each example contains three 64×64 RGB image modalities M_1, M_2, M_3 with distinct backgrounds, together with a digit label and a quadrant label. The image modalities share digit identity and quadrant

position but differ in background, so coherent generation requires preserving the shared labels while leaving background variation to the modality decoders. The dataset contains 220K training and 10K held-out evaluation samples, and MUNI uses product-based aggregation.

Metrics. We evaluate three tasks: *single-label→image*, *multi-label→image*, and *unconditional* co-generation. The first conditions on either the digit or quadrant label, the second conditions on both labels, and the third samples all five modalities jointly from the prior. For conditional generation, we report *alignment*: the fraction of generated images whose digit or quadrant is correctly recovered by a held-out verifier, averaged over image modalities M_1, M_2, M_3. For multi-label→image, a sample is counted as correct only when both labels are recovered. For unconditional co-generation, we report *coherence*, which requires verifier predictions on the three generated images to agree with one another and with the generated digit and quadrant labels.

Results. Tab. 2 summarizes verifier accuracy across the four metrics. On Single-L $\rightarrow$ I, MUNI remains competitive with the strongest baselines, and on Multi-L $\rightarrow$ I it achieves the best joint alignment. The largest gap appears in unconditional co-generation, where MUNI achieves the best coherence by a noticeable margin. This pattern is consistent with the criteria in Sec. 3. MMVAE samples from a mixture of single-modality experts, so the latent can ignore part of the observed subset and fail the predictivity criterion in Eqs. 8–9, explaining its weak Multi-L $\rightarrow$ I alignment. MoPoE uses the full product posterior at inference, which explains its strong conditional alignment; however, its prior is trained on mixture components that include smaller-subset latents, violating the coherence criterion in Eq. 7. MVAE and HELVAE aggregate all observed experts, but their standard objectives do not enforce cross-reconstruction or the minimality criterion in Eq. 10, limiting overall performance. MUNI aligns more directly with all three criteria, yielding strong conditional alignment together with the strongest unconditional coherence. Fig. 2 shows the same trend, with MUNI producing more consistent digit identity and quadrant position across jointly generated modalities.

Table 2: **PolyMNIST-Quadrant-Labels.** Single-L $\rightarrow$ I and Multi-L $\rightarrow$ I report verifier accuracy on generated images conditioned on a single label (Digit or Quadrant) or on both labels jointly. **Bold** and underline mark the best and second-best results.

	Single-L $\rightarrow$ I		Multi-L $\rightarrow$ I	Uncond.
Method	Digit ↑	Quadrant ↑	Both ↑	Coherence ↑
MVAE [55]	0.4137	<u>0.6803</u>	0.7053	0.0079
MMVAE [46]	<u>0.9279</u>	**0.9999**	0.1630	0.3167
MoPoE [49]	0.9199	**0.9999**	<u>0.9283</u>	0.3943
HELVAE [52]	**0.9297**	**0.9999**	0.7654	<u>0.4246</u>
MUNI **(Ours)**	0.9131	**0.9999**	**0.9346**	**0.4841**

6.3 Image-Text-Audio Any-to-Any Generation

Setup. We evaluate MUNI on a large-scale image-text-audio any-to-any benchmark, following the data setup of FlowBind [3]. Training uses three pairwise datasets across text, image, and audio: text-image (242K from LAION-COCO [30] and 30K from Flickr-30k [41]), text-audio (91K from AudioCaps v2 [27]), and image-audio (184K from VGGSound [7]). For many-to-one and unconditional co-generation, where no standard text-image-audio triplet benchmark exists, we follow the synthetic triplet protocol of FlowBind by generating images with FLUX.1-dev [1] from text-audio pairs in the AudioCaps v2 [27] test set to obtain (text, image, audio) triplets.

We report two variants of MUNI with aggregation rules compatible with the predictivity criterion: MUNI uses product aggregation, while MUNI† uses Hellinger aggregation, following HELVAE [52]. Our main comparisons are the generalist diffusion- and flow-based baselines from Sec. 6.1: CoDi, OmniFlow, and FlowBind. To isolate aggregation and prior-learning effects from model capacity, we also include MMVAE and MoPoE under the same backbone and compute as MUNI. The MVAE and HELVAE baselines are excluded because their standard training requires fully observed triplets, which are unavailable in this training setup. We further report LLM-based generalists and task-specific specialists as reference rows in App. E.

Metrics. We focus the main paper on two settings where aggregation and prior-routing effects are most visible. (i) *Many-to-one alignment* conditions on two source modalities, generates the third, and reports cross-modal alignment to each source separately, using CLIPScore [18] for text-image, CLAPScore [12] for text-audio, and Audio-Image Similarity (AIS) [54] for image-audio. (ii) *Unconditional coherence* jointly generates all three modalities and reports pairwise alignment

Table 3: **Many-to-one cross-modal alignment on image-text-audio.** MUNI uses PoE aggregation and MUNI† uses Hellinger aggregation. **Bold** and underline mark the best and second-best generalist results.

Category	Method	(I+A)→T		(T+A)→I		(T+I)→A	
		CLIP ↑	CLAP ↑	CLIP ↑	AIS ↑	CLAP ↑	AIS ↑
Generalists	CoDi [50]	24.05	33.72	24.98	85.52	11.06	65.31
	OmniFlow [31]	24.73	36.26	26.41	81.51	13.50	63.55
	FlowBind [3]	27.54	39.56	25.23	86.44	26.83	80.15
	MUNI **(Ours)**	25.20	**39.75**	25.44	**93.42**	26.88	**87.29**
	MUNI† **(Ours)**	**27.60**	36.61	**26.43**	90.94	**30.16**	86.89
Multimodal VAEs	MMVAE [46]	26.57	35.24	25.70	88.22	25.33	82.49
	MoPoE [49]	27.81	36.02	26.25	86.25	30.96	84.07



Figure 3: **Qualitative results of MUNI† (left) and MUNI (right) in unconditional co-generation of image-text-audio.** Each example shows a (text, audio, image) triple jointly sampled from the prior.

between generated modalities using the same scores. Due to space constraints, full *one-to-one fidelity* and *alignment* results are provided in App. E.

Results. On many-to-one generation (Tab. 3), MUNI and MUNI† achieve the best generalist result on every metric across the three input compositions. The gains are especially clear on audio-image alignment: MUNI improves AIS from 86.44 to 93.42 for (T+A)→I and from 80.15 to 87.29 for (T+I)→A over the next-best generalist. This pattern is consistent with the predictivity criterion in Eq. 9: CoDi's text-aligned conditioning can miss information not expressed in text, OmniFlow must learn conditional paths separately across subset configurations, making reliable multi-source prediction harder, and FlowBind maps a source to a deterministic target rather than sampling from $p(x_{S^c} \mid x_S)$, limiting its ability to represent the true conditional distribution. Among the multimodal VAE baselines, MMVAE remains weaker due to mixture-based single-expert sampling, while MoPoE is competitive conditionally because it uses the full product posterior at inference.

Table 4: **Unconditional coherence on image-text-audio.** CoDi [50] and FlowBind [3] are excluded since they do not support unconditional joint sampling. **Bold** and underline mark the best and second-best generalist results.

Method	T-I CLIP ↑	T-A CLAP ↑	A-I AIS ↑
OmniFlow [31]	21.17	14.23	50.95
MUNI **(Ours)**	**26.76**	23.73	81.24
MUNI† **(Ours)**	26.74	**24.83**	**82.80**
MMVAE [46]	25.52	20.06	76.68
MoPoE [49]	25.83	21.47	74.37

Unconditional co-generation (Tab. 4) shows a larger separation. Among the generalist baselines, only OmniFlow supports unconditional co-generation, but it is trained primarily through conditional generation paths and yields weak co-generation: MUNI† improves over it by $+5.57$ CLIP on T-I, $+10.60$ CLAP on T-A, and $+31.85$ AIS on A-I. Both MUNI variants also outperform MMVAE and MoPoE on all three pairwise alignments, matching the PolyMNIST trend and supporting the coherence criterion in Eq. 7. Qualitative unconditional samples from MUNI are shown in Fig. 3. Full one-to-one results, extended baseline comparisons, component-wise ablations, and one-to-many evaluations are provided in Apps. E.2–E.4.

7 Conclusion

We introduced MUNI, a multimodal unified-latent framework for any-to-any generation. MUNI extends Unified Latents to arbitrary subset conditioning, providing a single probabilistic interface for both conditional generation and unconditional co-generation. By identifying predictivity, coherence, and minimality as criteria for the shared latent, MUNI addresses the failure modes of existing multimodal VAE methods through three objective choices: non-mixture aggregation, target-detached self-reconstruction, and prior learning on leave-one-out latents. Experiments on PolyMNIST-Quadrant-Labels and image-text-audio generation show that MUNI preserves competitive conditional generation while substantially improving unconditional coherence. Our results further show that the multimodal variational-inference perspective can scale beyond controlled benchmarks when paired with expressive generative components. Overall, the results support the main message of this work: coherent any-to-any generation depends not only on how modality evidence is aggregated, but also on what information the shared latent is trained to contain.

Acknowledgements

This work was supported by the NRF grant (RS-2026-25486000); IITP grants (RS-2024-00399817, RS-2025-25441313, RS-2025-25443318, RS-2026-25526850, RS-2026-25617720) funded by MSIT, Korea; the Industrial Technology Innovation Program grant (RS-2025-02317326) funded by MOTIE, Korea; the InnoCORE program of MSIT (AI Meta-Scientist, N10260110); and the DRB-KAIST SketchTheFuture Research Center.

Competing Interests

The authors declare no competing financial interests relevant to this work.

References

[1] Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024.

[2] A. Campbell, J. Yim, R. Barzilay, T. Rainforth, and T. Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *ICML*, 2024.

[3] Y. Cha, S. Kim, J. Kwon, and S. Hong. FlowBind: Efficient any-to-any generation with bidirectional flows. In *ICLR*, 2026.

[4] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[5] S. N. Chandrasekaran, J. Ackerman, E. Alix, D. M. Ando, J. Arevalo, M. Bennion, N. Boisseau, A. Borowa, J. D. Boyd, L. Brino, et al. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, 2023.

[6] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang, C. Weng, and Y. Shan. VideoCrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[7] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. VGGSound: A large-scale audio-visual dataset. In *ICASSP*, 2020.

[8] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[9] J. Cui, Y.-Y. Chen, Y. Zhang, and M. Klenk. ShaLa: Multimodal shared latent generative modelling. In *AAAI*, 2026.

[10] I. Daunhawer, T. M. Sutter, K. Chin-Cheong, E. Palumbo, and J. E. Vogt. On the limitations of multimodal VAEs. In *ICLR*, 2022.

[11] I. Daunhawer, T. M. Sutter, R. Marcinkevičs, and J. E. Vogt. Self-supervised disentanglement of modality-specific and shared factors improves multimodal generative models. In *German Conference on Pattern Recognition*, 2020.

[12] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang. CLAP: Learning audio concepts from natural language supervision. In *ICASSP*, 2023.

[13] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[14] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons. Stable Audio Open. In *ICASSP*, 2025.

[15] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[16] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.

[17] J. Heek, E. Hoogeboom, T. Mensink, and T. Salimans. Unified Latents (UL): How to train your latents. *arXiv preprint arXiv:2602.17270*, 2026.

[18] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021.

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017.

[20] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[22] Hugging Face. Stable unCLIP. `https://huggingface.co/docs/diffusers/api/pipelines/stable_unclip`, 2025.

[23] C.-Y. Hung, N. Majumder, Z. Kong, A. Mehrish, A. Zadeh, C. Li, R. Valle, B. Catanzaro, and S. Poria. TangoFlux: Super fast and faithful text to audio generation with flow matching and Clap-ranked preference optimization. In *ICLR*, 2026.

[24] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *CVPR*, 2024.

[25] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

[26] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi. Fréchet Audio Distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019.

[27] C. D. Kim, B. Kim, H. Lee, and G. Kim. AudioCaps: Generating captions for audios in the wild. In *NAACL*, 2019.

[28] D. P. Kingma, T. Salimans, B. Poole, and J. Ho. Variational diffusion models. In *NeurIPS*, 2021.

[29] B.-K. Kwon, Q. Dai, H. Lee, C. Luo, and T.-H. Oh. JointDiT: Enhancing RGB-depth joint modeling with diffusion transformers. In *ICCV*, 2025.

[30] LAION e.V. reLAION-COCO. `https://huggingface.co/datasets/laion/relaion-coco`, 2025.

[31] S. Li, K. Kallidromitis, A. Gokul, Z. Liao, Y. Kato, K. Kozuka, and A. Grover. OmniFlow: Any-to-any generation with multi-modal rectified flows. In *CVPR*, 2025.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[33] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *ICLR*, 2023.

[34] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *ICML*, 2023.

[35] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`, 2024.

[36] J. Lu, C. Clark, S. Lee, Z. Zhang, S. Khosla, R. Marten, D. Hoiem, and A. Kembhavi. Unified-IO 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *CVPR*, 2024.

[37] K. Märtens and C. Yau. Disentangling shared and private latent factors in multimodal variational autoencoders. In *Machine Learning in Computational Biology (MLCB)*, 2024.

[38] E. Palumbo, I. Daunhawer, and J. E. Vogt. MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises. In *ICLR*, 2023.

[39] E. Palumbo, L. Manduchi, S. Laguna, D. Chopard, and J. E. Vogt. Deep generative clustering with multimodal diffusion variational autoencoders. In *ICLR*, 2024.

[40] L. Parker, F. Lanusse, S. Golkar, L. Sarra, M. Cranmer, A. Bietti, M. Eickenberg, G. Krawezik, M. McCabe, R. Morel, R. Ohana, M. Pettee, B. Régaldo-Saint Blancard, K. Cho, S. Ho, and The Polymathic AI Collaboration. AstroCLIP: A cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 2024.

[41] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015.

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[43] K. Rojas, Y. Zhu, S. Zhu, F. X.-F. Ye, and M. Tao. Diffuse Everything: Multimodal diffusion models on arbitrary state spaces. In *ICML*, 2025.

[44] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[45] H. Schechter Vera, S. Dua, B. Zhang, et al. EmbeddingGemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

[46] Y. Shi, N. Siddharth, B. Paige, and P. H. S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *NeurIPS*, 2019.

[47] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[48] K. Sung-Bin, A. Senocak, H. Ha, and T.-H. Oh. Sound2Vision: Generating diverse visuals from audio through cross-modal latent alignment. *arXiv preprint arXiv:2412.06209*, 2024.

[49] T. M. Sutter, I. Daunhawer, and J. E. Vogt. Generalized multimodal ELBO. In *ICLR*, 2021.

[50] Z. Tang, Z. Yang, C. Zhu, M. Zeng, and M. Bansal. Any-to-any generation via composable diffusion. In *NeurIPS*, 2023.

[51] R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015.

[52] H. Vo and I. Valera. Hellinger multimodal variational autoencoders. In *AISTATS*, 2026.

[53] D. Wesego and P. Rooshenas. Multimodal ELBO with diffusion decoders. In *CVPRW*, 2026.

[54] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello. Wav2CLIP: Learning robust audio representations from CLIP. In *ICASSP*, 2022.

[55] M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. In *NeurIPS*, 2018.

[56] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua. NExT-GPT: Any-to-any multimodal LLM. In *ICML*, 2024.

[57] Y. Xing, Y. He, Z. Tian, X. Wang, and Q. Chen. Seeing and Hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024.

[58] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, D. Yin, Y. Zhang, W. Wang, Y. Cheng, B. Xu, X. Gu, Y. Dong, and J. Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025.

[59] S. Yuan, J. Cui, H. Li, and T. Han. Learning multimodal latent generative models with energy-based prior. In *ECCV*, 2024.

[60] J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li, H. Yan, J. Fu, T. Gui, T. Sun, Y.-G. Jiang, and X. Qiu. AnyGPT: Unified multimodal LLM with discrete sequence modeling. In *ACL*, 2024.

[61] Y. Zhang, Y. Shen, and W. Wang. Disentanglement of variations with multimodal generative modeling. In *ICLR*, 2026.

[62] C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025.

Societal Impacts

Our work enables any-to-any multimodal generation from partially observed data, reducing the need for expensive fully observed multimodal tuples. However, stronger image-text-audio generation may also increase risks of synthetic-media misuse, including misinformation, impersonation, and privacy or copyright violations.

Limitations and Future Work

MUNI uses a general objective across all conditioning settings and does not explicitly exploit known structure among modalities. In practice, cross-modal relations can be highly asymmetric: for example, text-to-speech generation contains substantial variation, whereas speech-to-text is nearly deterministic. Incorporating such modality-specific structure into route sampling, loss design, or architecture choices could make the framework more effective for particular multimodal domains. Extending MUNI to exploit these known relationships is an important direction for future work.

A Additional Analysis

A.1 Objective Components

Relative to the direct cross-modal UL objective, Eq. 12 modifies three components: how subset posteriors are aggregated, which posterior is used for each reconstruction term, and which subset latents are used for prior learning.

Aggregation. Given a modality subset A, the subset posterior is constructed as

$$q_{\phi,A}(z \mid x_A) = \mathcal{A}\left(\{q_{\phi_j}(z \mid x_j)\}_{j\in A}\right), \tag{13}$$

where $\mathcal{A}$ is instantiated as product aggregation or Hellinger aggregation. Both choices form a single posterior from all observed experts, unlike mixture rules that may sample from only part of the observed subset. This makes the aggregation rule compatible with the predictivity criterion in Eq. 9, although aggregation alone does not enforce the criterion.

Target-routed reconstruction posterior. For reconstruction target m, Eq. 12 uses

$$\bar{q}^{(m)}_{\phi_j}(z \mid x_j) = \begin{cases} \mathrm{sg}\left[q_{\phi_j}(z \mid x_j)\right], & j = m, \\ q_{\phi_j}(z \mid x_j), & j \neq m, \end{cases} \qquad q^{(m)}_{\phi,A}(z \mid x_A) = \mathcal{A}\left(\{\bar{q}^{(m)}_{\phi_j}(z \mid x_j)\}_{j\in A}\right). \tag{14}$$

If $m \notin A$, no target expert is present and $q^{(m)}_{\phi,A}$ reduces to the usual subset posterior $q_{\phi,A}$. If $m \in A$, the forward pass still uses the target expert, but this expert is detached in the backward pass. Consequently, each encoder expert is updated only by reconstruction losses whose target is a different modality. The encoder is therefore trained to preserve information useful for cross-modal prediction, rather than details useful only for reconstructing its own modality. This removes the self-copying pressure that would otherwise place modality-specific information in the shared latent, supporting the minimality criterion in Eq. 10.

Latents for prior learning. The prior term in Eq. 12 is gated as

$$\mathbf{1}[|A| \geq M-1]\, \mathbb{E}_{z\sim q_{\phi,A}(\cdot|x_A)} \mathcal{L}_z(\psi; z). \tag{15}$$

Thus, prior learning is applied only to full-modality and leave-one-out subset latents, since smaller subset latents can under-specify the joint structure needed for coherent unconditional co-generation. For any non-empty proper subset $A \subsetneq [M]$,

$$\mathrm{TC}(X_{1:M} \mid Z_A) = \mathrm{TC}(X_A \mid Z_A) + \mathrm{TC}(X_{A^c} \mid Z_A) + I(X_A; X_{A^c} \mid Z_A). \tag{16}$$

The second term captures dependence among the modalities absent from the conditioning set, which cannot in general be determined by a subset latent Z_A inferred only from X_A. For a leave-one-out subset $A_i = [M] \setminus \{i\}$, this term vanishes identically because A_i^c contains a single modality. Hence, if Z_{A_i} makes the observed modalities conditionally independent and is sufficient for the held-out modality, then all terms in Eq. 16 vanish, giving $\mathrm{TC}(X_{1:M} \mid Z_{A_i}) = 0$. This motivates applying the prior loss to leave-one-out and full-modality latents.

B Proof of Proposition 4.1

We formalize the route-wise population argument underlying Proposition 4.1. The stop-gradient operator leaves the forward latent unchanged, but determines which encoder experts receive reconstruction gradients on each route. The proof therefore characterizes the sufficient latent selected by this routing rule.

Fix a non-empty subset $A \subseteq [M]$ and let Z_A be a candidate latent inferred from X_A. For a target modality m, define the evidence that can train the encoder on that route (A, m) as

$$Y_{A,m} = \begin{cases} X_A, & m \notin A, \\ X_{A\setminus\{m\}}, & m \in A. \end{cases} \tag{17}$$

For missing targets, all observed modalities in A receive prediction gradients. For self targets, the target expert is detached, so only the non-target experts in $A\setminus\{m\}$ can be trained by the reconstruction of X_m.

We define the conditional predictive information

$$\rho_{A,m}(Z_A) = I(Y_{A,m}; X_m \mid Z_A). \tag{18}$$

Thus, for $m \notin A$,

$$\rho_{A,m}(Z_A) = I(X_A; X_m \mid Z_A), \tag{19}$$

while for $m \in A$,

$$\rho_{A,m}(Z_A) = I(X_{A\setminus\{m\}}; X_m \mid Z_A). \tag{20}$$

We define the set of predictively sufficient latents for subset A as

$$\mathcal{S}_A = \{Z_A : \rho_{A,m}(Z_A) = 0 \text{ for all } m \in [M]\}. \tag{21}$$

For missing targets, the condition $\rho_{A,m}(Z_A) = 0$ is exactly $I(X_A; X_m \mid Z_A) = 0$. For self targets, the conditions $\rho_{A,m}(Z_A) = 0$ for all $m \in A$ are equivalent to $\mathrm{TC}(X_A \mid Z_A) = 0$, which is proved by the following lemma:

Lemma B.1 (Conditional sufficiency implies conditional factorization)**.** *Suppose that, for every modality $m \in A$,*

$$I(X_{A\setminus\{m\}}; X_m \mid Z_A) = 0. \tag{22}$$

Then

$$\mathrm{TC}(X_A \mid Z_A) = 0. \tag{23}$$

Proof. Write $A = \{a_1, \ldots, a_K\}$. By the chain rule for conditional total correlation,

$$\mathrm{TC}(X_A \mid Z_A) = \sum_{k=2}^{K} I(X_{a_k}; X_{a_1}, \ldots, X_{a_{k-1}} \mid Z_A). \tag{24}$$

For each k, the variables $(X_{a_1}, \ldots, X_{a_{k-1}})$ are a subset of $X_{A\setminus\{a_k\}}$. Therefore,

$$I(X_{a_k}; X_{a_1}, \ldots, X_{a_{k-1}} \mid Z_A) \le I(X_{a_k}; X_{A\setminus\{a_k\}} \mid Z_A) = 0, \tag{25}$$

where the last equality follows from the assumption. Hence every term in the chain-rule expansion is zero, and $\mathrm{TC}(X_A \mid Z_A) = 0$. □

Therefore, the sufficient set can equivalently be written as

$$\mathcal{S}_A = \{Z_A : \mathrm{TC}(X_A \mid Z_A) = 0,\ I(X_A; X_m \mid Z_A) = 0 \text{ for all } m \in A^c\}. \tag{26}$$

Assumptions. We make the following idealized assumptions. Throughout this proof, all losses are population risks, i.e., expectations over the data distribution.

Model capacity and optimization. We assume that the encoder and decoder families are sufficiently expressive and that the population objectives are optimized ideally. For any route (A, m), if

$$\rho_{A,m}(Z_A) > 0, \tag{27}$$

then the encoder experts receiving gradients on that route, together with the decoder for modality m, admit an update that decreases the corresponding expected target loss. Conversely, when $\rho_{A,m}(Z_A) =$

0 and the decoder is optimal given Z_A, the route is target-wise optimal. We also assume that the ideal latent $Z_A^\star$ defined in Eq. 10 is representable by the encoder family.

Prior ordering. For any latent Z_A, define the optimal prior cost for its marginal distribution as

$$\mathcal{C}_A(Z_A) = \inf_{\psi} \mathbb{E}[\mathcal{L}_z(\psi; Z_A)]. \tag{28}$$

We assume that, within the sufficient set $\mathcal{S}_A$, larger retained information from X_A implies larger optimal prior cost:

$$I(X_A; Z) < I(X_A; Z') \quad \Longrightarrow \quad \mathcal{C}_A(Z) < \mathcal{C}_A(Z'), \qquad Z, Z' \in \mathcal{S}_A. \tag{29}$$

A simple case where this ordering holds is maximum-likelihood training with a prior family that can realize the marginal density of Z_A. Then

$$\inf_{\psi} \mathbb{E}[-\log q_\psi(Z_A)] = H(Z_A) = I(X_A; Z_A) + H(Z_A \mid X_A). \tag{30}$$

Thus, when $H(Z_A \mid X_A)$ is fixed across the latents being compared, minimizing the optimal prior cost over $\mathcal{S}_A$ is equivalent to minimizing $I(X_A; Z_A)$. When the prior is trained with diffusion or flow-matching losses, we use Eq. 29 as the corresponding idealized ordering assumption for the prior loss.

Small prior weight. We assume that β is sufficiently small that the prior loss cannot compensate for a reducible target loss. Consequently, solutions are first restricted to target-wise optimal, where no target loss can be further decreased by the corresponding feasible update. The prior cost then selects among these latents. Under this assumption, the target losses enforce sufficiency, while the prior term selects among sufficient latents according to $\mathcal{C}_A$.

Proposition B.2 (Minimal sufficient latent)**.** *Let $\mathcal{B}_A$ be the set of latents for which no target loss can be further reduced by a feasible update of the encoder experts receiving gradients on the corresponding route, together with the corresponding decoder. Under the assumptions above, let $\widehat{Z}_A$ be a minimizer of the prior cost over $\mathcal{B}_A$:*

$$\widehat{Z}_A = \operatorname*{argmin}_{Z_A \in \mathcal{B}_A} \mathcal{C}_A(Z_A) \tag{31}$$

Then $\widehat{Z}_A$ solves the minimal sufficient problem

$$\widehat{Z}_A = \operatorname*{argmin}_{Z_A \in \mathcal{S}_A} I(X_A; Z_A). \tag{32}$$

Moreover, if $|A| \geq M - 1$, then $\widehat{Z}_A$ is coherence-sufficient:

$$\mathrm{TC}(X_{1:M} \mid \widehat{Z}_A) = 0. \tag{33}$$

Proof. We first show that every element of $\mathcal{B}_A$ is sufficient. Suppose $Z_A \in \mathcal{B}_A$ but $\rho_{A,m}(Z_A) > 0$ for some target modality m. By the model capacity and optimization assumption, the encoder experts receiving gradients on that route, together with the decoder for modality m, admit a feasible update that decreases the corresponding expected target loss. This contradicts the definition of $\mathcal{B}_A$. Hence

$$\rho_{A,m}(Z_A) = 0 \quad \text{for all } m \in [M], \tag{34}$$

and therefore

$$\mathcal{B}_A \subseteq \mathcal{S}_A. \tag{35}$$

Since $Z_A^\star$ is assumed to be representable by the encoder family, it is feasible under the model class. By definition, $Z_A^\star \in \mathcal{S}_A$, so $\rho_{A,m}(Z_A^\star) = 0$ for all $m \in [M]$. Under the idealized optimization assumption, no corresponding target loss can be further decreased by the feasible updates. Hence $Z_A^\star \in \mathcal{B}_A$.

Let $\widehat{Z}_A$ be a solution of Eq. 31. Since $Z_A^\star \in \mathcal{B}_A$, prior selection gives

$$\mathcal{C}_A(\widehat{Z}_A) \leq \mathcal{C}_A(Z_A^\star). \tag{36}$$

Since $\widehat{Z}_A, Z_A^\star \in \mathcal{S}_A$, the contrapositive of Eq. 29 gives

$$I(X_A; \widehat{Z}_A) \leq I(X_A; Z_A^\star). \tag{37}$$

On the other hand, $Z_A^\star$ minimizes $I(X_A; Z_A)$ over $\mathcal{S}_A$, and $\widehat{Z}_A \in \mathcal{S}_A$. Therefore the two inequalities are tight, and $\widehat{Z}_A$ also solves Eq. 32.

It remains to show coherence-sufficiency for the subset latents used as prior targets. If $A = [M]$, then $\widehat{Z}_A \in \mathcal{S}_A$ directly gives

$$\mathrm{TC}(X_{1:M} \mid \widehat{Z}_A) = 0. \tag{38}$$

Now consider the leave-one-out case $A = A_i = [M] \setminus \{i\}$. Since $A_i^c = \{i\}$, sufficiency gives

$$I(X_{A_i}; X_i \mid \widehat{Z}_{A_i}) = 0. \tag{39}$$

The self-target conditions inside A_i imply, by Lemma B.1,

$$\mathrm{TC}(X_{A_i} \mid \widehat{Z}_{A_i}) = 0. \tag{40}$$

Using the subset decomposition in Eq. 16,

$$\mathrm{TC}(X_{1:M} \mid \widehat{Z}_{A_i}) = \mathrm{TC}(X_{A_i} \mid \widehat{Z}_{A_i}) + \mathrm{TC}(X_i \mid \widehat{Z}_{A_i}) + I(X_{A_i}; X_i \mid \widehat{Z}_{A_i}). \tag{41}$$

Since X_i is a single modality, $\mathrm{TC}(X_i \mid \widehat{Z}_{A_i}) = 0$. Combining this with Eqs. 39 and 40 in the decomposition above yields $\mathrm{TC}(X_{1:M} \mid \widehat{Z}_{A_i}) = 0$, as desired.

Thus full-modality and leave-one-out latents are coherence-sufficient for joint generation. □

C Details of Reconstruction and Prior Losses

Eq. 12 writes the reconstruction and prior terms abstractly as $\mathcal{L}_x^m$ and $\mathcal{L}_z$. This section gives the concrete losses used in our experiments. The prior term is the analogue of the KL term in a multimodal VAE:

$$D_{\mathrm{KL}}\left(q_{\phi,A}(z \mid x_A) \,\|\, q_\psi(z)\right) = \mathbb{E}_{z\sim q_{\phi,A}}\left[-\log q_\psi(z)\right] - H\left(q_{\phi,A}(\cdot \mid x_A)\right). \tag{42}$$

As in Unified Latents [17], the negative log-density under the learned prior is represented by a diffusion/flow objective. However, since directly optimizing the diffusion ELBO is ineffective for training the prior flow model [28], we train the prior with the standard conditional flow-matching loss and apply the ELBO-correct gradient only to the encoder.

Flow prior. Let $q_{\phi,A}$ be a diagonal Gaussian posterior and sample

$$z \sim \mathcal{N}(\mu_{\phi,A}(x_A), \sigma^2_{\phi,A}(x_A) I). \tag{43}$$

The flow prior uses the linear path

$$z_u = uz + (1-u)\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad u \sim p(u). \tag{44}$$

The prior network is trained with the standard conditional flow-matching loss

$$\ell_z(\psi; z, u, \epsilon) = \left\| v_\psi(z_u, u) - (z - \epsilon) \right\|_2^2. \tag{45}$$

For the encoder, however, we use the ELBO-corrected gradient for the prior likelihood, following the derivation of Unified Latents [17]. For the linear path in Eq. 44,

$$\mathrm{SNR}(u) = \frac{u^2}{(1-u)^2}, \qquad \frac{1}{2}\frac{d}{du}\mathrm{SNR}(u) = \frac{u}{(1-u)^3}. \tag{46}$$

Thus, the model-dependent part of the continuous-time ELBO weights the clean-latent prediction error as

$$\mathcal{J}_{\mathrm{ELBO}}(\psi; z) = \int_0^1 \frac{u}{(1-u)^3} \mathbb{E}_\epsilon\left[\|\hat{z}_0(z_u, u) - z\|_2^2\right] du, \tag{47}$$

where

$$\hat{z}_0(z_u, u) = z_u + (1-u) v_\psi(z_u, u). \tag{48}$$

Since

$$\|\hat{z}_0 - z\|_2^2 = (1-u)^2 \|v_\psi(z_u, u) - (z - \epsilon)\|_2^2, \tag{49}$$

sampling $u \sim p(u)$ gives the ELBO-correct importance factor for the velocity loss:

$$\omega_{\mathrm{ELBO}}(u) = \frac{u}{(1-u)p(u)}. \tag{50}$$

The factor $1/p(u)$ corrects for non-uniform timestep sampling.

In implementation, we apply this weight only to the encoder-side gradient using a gradient-scaling operator gs_λ. It is the identity in the forward pass and scales the backward gradient by λ:

$$\mathrm{gs}_\lambda(z) = z \quad \text{in the forward pass,} \qquad \frac{\partial\, \mathrm{gs}_\lambda(z)}{\partial z} = \lambda I \quad \text{in the backward pass.} \tag{51}$$

For the prior route, we set

$$\lambda_z(u) = \gamma_z\, \omega_{\mathrm{ELBO}}(u), \tag{52}$$

where γ_z is the configured `encoder_grad_scale`. This scalar controls the encoder-side strength of the prior term relative to reconstruction, and its values are reported in Tab. 7.

The entropy term of the KL in Eq. 42 is analytic for the diagonal Gaussian posterior. Let $\lambda_z(u) = \gamma_z \omega_{\mathrm{ELBO}}(u)$, $\tilde{z} = \mathrm{gs}_{\lambda_z(u)}(z)$, and $\tilde{z}_u = u\tilde{z} + (1-u)\epsilon$. We use

$$\mathcal{L}_z(\psi, \phi; A) = \mathbb{E}_{\substack{z\sim q_{\phi,A}(\cdot|x_A) \\ \epsilon\sim\mathcal{N}(0,I),\, u\sim p(u)}} \left[\left\| v_\psi(\tilde{z}_u, u) - (\tilde{z} - \epsilon) \right\|_2^2 \right] - \gamma_z H\left(q_{\phi,A}(\cdot \mid x_A)\right). \tag{53}$$

Since gs_λ is the identity in the forward pass, the prior network is trained with the standard conditional flow-matching loss, while the encoder receives the ELBO-correct gradient scale $\lambda_z(u)$ and the analytic entropy term.

Decoder losses. The reconstruction loss $\mathcal{L}_x^m$ is chosen according to the target modality. For a flow decoder with continuous target x_m, we use the linear path

$$x_{m,u} = u x_m + (1-u)\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad u \sim p_m(u), \tag{54}$$

and train with the conditional flow-matching loss

$$\mathcal{L}_x^m(\theta_m; x_m, z) = \mathbb{E}_{u,\epsilon}\left[\left\| v_{\theta_m}(x_{m,u}, u; z) - (x_m - \epsilon) \right\|_2^2 \right]. \tag{55}$$

For categorical label modalities, we use cross-entropy,

$$\mathcal{L}_x^m(\theta_m; y_m, z) = -\log p_{\theta_m}(y_m \mid z), \tag{56}$$

and for autoregressive text decoders, we use teacher forcing,

$$\mathcal{L}_x^m(\theta_m; x_{m,1:T}, z) = -\sum_{t=1}^{T} \log p_{\theta_m}(x_{m,t} \mid x_{m,<t}, z). \tag{57}$$

D Controlled Gaussian-Mixture Experiment

We additionally conduct a controlled synthetic experiment with a known shared-private structure. Unlike the main benchmarks, this setting gives direct control over which factors are shared by all modalities, which factors are shared by only a subset of modalities, and which factors are private to a single modality. This allows us to directly examine how each method handles shared and private factors in conditional and unconditional generation.

Setup. We construct a three-modality benchmark with modalities M_1, M_2, M_3. The data are generated from seven independent ground-truth factors: a 10-dimensional factor shared by all modalities z_{123}, three 10-dimensional pair-shared factors z_{12}, z_{13}, z_{23}, and three 20-dimensional modality-private factors z_1, z_2, z_3. Each factor is sampled from a randomly generated Gaussian mixture with 64 components. Each observation $x_i \in \mathbb{R}^{50}$ is the concatenation of the factors involving modality M_i:

$$\begin{aligned} x_1 &= (z_{123}, z_{12}, z_{13}, z_1), \\ x_2 &= (z_{123}, z_{12}, z_{23}, z_2), \\ x_3 &= (z_{123}, z_{13}, z_{23}, z_3). \end{aligned}$$

Table 5: **Controlled Gaussian-mixture benchmark.** Alignment is the shared-coordinate MSE between conditioning and target modalities for conditional generation. For unconditional co-generation, fidelity is measured by SWD to the ground-truth modality marginal, and coherence is the shared-coordinate MSE across jointly sampled modalities. Lower is better. **Bold** marks the best result and underline marks the second-best. A dash indicates that the corresponding setting is not evaluated for that method.

	1→1	2→1	Uncond.	
Method	Align. ↓	Align. ↓	Fid. ↓	Coh. ↓
MVAE [55]	0.2265	0.3233	1.5996	0.2801
MMVAE [46]	**0.0267**	5.0647	0.2080	2.0496
MoPoE [49]	0.0558	0.0604	0.2720	0.6297
HELVAE [52]	0.6325	2.1096	0.4345	4.3642
MMVAE+ [38]	0.0276	5.0413	2.8070	2.0432
ShaLa [9]	7.2122	-	0.8075	2.5984
`MUNI` **(Ours)**	0.0508	**0.0414**	**0.1300**	**0.0394**

Additional baselines. We include MMVAE+ [38] to test whether explicit private latents remain useful with expressive decoders, and ShaLa [9] as an existing baseline with an expressive prior. All methods use the same encoder and decoder architectures. For ShaLa, we replace the original Gaussian decoder with the same flow-matching decoder used by the other baselines. Since ShaLa does not support multi-conditioned generation, we do not evaluate it on the $2 \to 1$ setting.

Metrics. We evaluate three generation tasks: *one-to-one*, *many-to-one*, and *unconditional*. For the conditional tasks, we report *alignment*, the mean squared error (MSE) on the shared coordinates between each conditioning modality and the generated target. For unconditional co-generation, we report *fidelity* as the sliced Wasserstein distance (SWD) between generated samples and the ground-truth marginal of each generated modality, and *coherence* as the shared-coordinate MSE across pairs of jointly sampled modalities. All metrics are averaged across supported source-target configurations, and lower values are better.

Implementation details. We train on 10^6 procedurally generated training examples and evaluate on a held-out test split. All reported toy results are computed with $N = 20{,}000$ samples per task. SWD is computed with 128 random projections.

All methods use the same backbone family: modality-specific MLP encoders, an MLP flow-matching prior, and modality-specific MLP flow-matching decoders. The latent has dimension 48 for MVAE [55], MMVAE [46], MoPoE [49], HELVAE [52], and `MUNI`; MMVAE+ [38] uses the same 48-dimensional shared latent together with an additional 24-dimensional private latent per modality. Apart from method-specific components such as MMVAE+'s private latents and ShaLa's two-stage conditional prior, the baselines share the same MLP encoder, flow prior, and flow decoder backbones. Flow timesteps are sampled from a logit-normal distribution during training. At evaluation, we use a fixed 50-step sampler budget for both the learned prior and the modality decoders, with the same sampler settings used across methods. Other details are summarized in Tab. 7.

Results. Tab. 5 reports the quantitative results. MoE-based methods, including MMVAE and MMVAE+, achieve strong $1 \to 1$ alignment, but degrade sharply in the $2 \to 1$ setting and in unconditional coherence. This suggests that mixture-based aggregation does not reliably fuse evidence from multiple observed modalities or support coherent joint sampling. MoPoE improves conditional alignment through product-based aggregation at inference, but its unconditional coherence remains limited as expected.

MVAE and HELVAE are less competitive overall. Their shared latent is pressured to carry both shared content and modality-specific variation, which hurts performance under limited latent capacity. MMVAE+ introduces private latents, but does not meaningfully improve over MMVAE in this benchmark, indicating that explicit private latents provide limited benefit once the decoders are expressive. ShaLa uses an expressive prior but does not support multi-conditioned generation; in the evaluated settings, it remains weak on $1 \to 1$ alignment and unconditional coherence under the same training budget.

Table 6: **PolyMNIST-Quadrant-Labels.** Alignment is verifier accuracy on generated images for conditional label-to-image tasks. Coherence is strict agreement between verifier predictions on M_1, M_2, M_3 and the generated labels under unconditional sampling. Higher is better. **Bold** marks the best result and underline marks the second-best. A dash indicates an unsupported setting.

	Single-L → I		Multi-L → I	Uncond.
Method	Digit ↑	Quad. ↑	Align. ↑	Coh. ↑
MVAE [55]	0.4137	0.6803	0.7053	0.0079
MMVAE [46]	0.9279	**0.9999**	0.1630	0.3167
MoPoE [49]	0.9199	**0.9999**	0.9283	0.3943
HELVAE [52]	0.9297	**0.9999**	0.7654	0.4246
MMVAE+ [38]	**0.9368**	**0.9999**	0.1668	0.2783
ShaLa [9]	0.8976	0.9915	-	0.4539
`MUNI` **(Ours)**	0.9131	**0.9999**	**0.9346**	**0.4841**

`MUNI` achieves the best $2 \rightarrow 1$ alignment and the strongest unconditional performance with the lowest fidelity and coherence errors. These results support the latent-space criteria in Sec. 3: combining all observed evidence supports predictivity in many-to-one generation, while restricting prior learning to leave-one-out latents improves coherence in unconditional co-generation.

E More Experimental Details and Results

We provide additional implementation details and extended results for Sec. 6.

E.1 PolyMNIST-Quadrant-Labels

Architecture. All compatible methods use the same backbone family. The shared latent is 32-dimensional, while MMVAE+ uses an additional 32-dimensional private latent for each modality. The image modalities use a shared modality-conditioned ResNet encoder with modality-specific input and posterior heads, and a shared NCSN++ flow-matching decoder whose target modality is specified by a modality label. The digit and quadrant modalities use MLP encoders and classifier decoders. Please refer to Tab. 7 for additional architectural details.

Training. Image flow decoders use logit-normal timestep sampling following SD3 [13], while latent priors use uniform timestep sampling. See Tab. 7 for more training details.

Evaluation. All metrics are computed on the full 10K held-out split using the same definitions as in Sec. 6.2. A frozen verifier predicts digit and quadrant labels from generated images. At evaluation time, we use 50 prior sampling steps and 25 image-decoder sampling steps.

Results. Tab. 6 extends the main comparison with MMVAE+ [38] and ShaLa [9]. MMVAE+ improves single-label digit alignment, but remains weak under multi-label conditioning and unconditional co-generation, suggesting that explicit private latents provide limited additional benefit once the modality decoders are expressive. ShaLa achieves the second-best unconditional coherence, but does not support multi-label conditioning and is less competitive under single-label conditioning. `MUNI` achieves the best multi-label alignment and the best unconditional coherence while remaining competitive under single-label conditioning.

E.2 Image-Text-Audio Any-to-Any Generation

Additional baselines. In addition to the generalist any-to-any baselines reported in the main paper, we include task-specific specialists for all six one-to-one generation directions:

- **T→I:** FLUX.1-dev [1], a 12B rectified-flow transformer for prompt-conditioned image synthesis.
- **I→T:** LLaVA-NeXT [35], a visual instruction-tuned large multimodal model with improved high-resolution image understanding, OCR, and reasoning capabilities.

- **T→A:** TangoFlux [23], an efficient flow-matching text-to-audio model trained with CLAP-ranked preference optimization for improved text-audio alignment.
- **A→T:** Qwen2-Audio [8], a large audio-language model that accepts speech, music, and environmental audio as input and produces natural-language responses.
- **I→A:** Seeing and Hearing [57], a visual-guided audio generation method that aligns pretrained visual and audio diffusion models through a shared latent space.
- **A→I:** Sound2Vision [48], an audio-to-image method that maps audio representations into a visual latent space and uses a pretrained image generator to synthesize plausible scenes from sound.

These specialist models are optimized for individual translation directions and are therefore reported as reference rows rather than direct comparison to generalist any-to-any systems. For example, FLUX.1 prioritizes perceptual quality and human preference and often produces stylized images [29], which FID can penalize [24]. We also include UnifiedIO2-L [36] as an LLM-based generalist baseline. UnifiedIO2-L tokenizes images, text, audio, and other modalities into a shared autoregressive sequence space, but it does not support audio-image translation. Diffuse Everything [43] is a native multimodal model trained only on images and text, and we evaluate it on text-image translation.

Dataset. We adopt the pairwise training data setup of FlowBind [3], using paired text, image, and audio data without any fully observed triplets. For text-image training, we use 242K LAION-COCO [30] pairs, filtered to examples with aesthetic scores above 5.0, together with 30K Flickr-30k [41] image-text pairs. For text-audio training, we use 91K AudioCaps v2 [27] examples, which provide natural-language captions for YouTube audio clips. For audio-image training, we use 184K VGGSound [7] examples, a large-scale audio-visual dataset collected from YouTube videos.

For one-to-one evaluation, text-to-image and image-to-text are evaluated on the MS-COCO [32] evaluation split, while text-audio, audio-text, image-audio, and audio-image tasks are evaluated on held-out test splits of the corresponding pairwise datasets.

Implementation details. We follow FlowBind [3] to define multimodal latent learning in the aligned representation spaces of pretrained modality-specific encoders and decoders. For images, we use CLIP [42] as the encoder and Stable-UnCLIP [22] as the pretrained image generation backbone. For audio, we use CLAP [12] as the encoder and AudioLDM [34] as the pretrained audio generation backbone. For text, we use EmbeddingGemma [45] as the encoder and Gemma3-1B [15] as the autoregressive text decoder. This design allows us to train over compact semantic representations.

Unlike FlowBind [3], which maps all modalities into a fixed-dimensional shared latent space, our framework allows modality-specific decoder heads with different output dimensions. We use this flexibility to fine-tune the autoregressive text decoder for stronger text generation. Specifically, we fine-tune the Gemma3-1B text decoder using LoRA [21] adapters with rank $r = 16$, scaling factor $\alpha = 32$, and dropout rate 0.05. All other encoders, decoders, and the prior network are implemented as MLPs with residual connections.

For encoder aggregation, we instantiate two variants: `MUNI` uses product-of-experts aggregation, while `MUNI`† uses Hellinger aggregation. For fair comparisons, we reproduce the multimodal VAE baselines [46, 49] using the same pretrained encoders and decoders, optimization settings, and compute budget as `MUNI`. Other details are summarized in Tab. 7.

Training. We use modality-specific cross-reconstruction losses. For text, we train the autoregressive decoder with cross-entropy loss on the first 64 examples of each effective batch of 1024, weighted by 0.05, and add a representation alignment loss between the final 768-dimensional output of the projection MLP and the target EmbeddingGemma [45] vector. For image and audio, we use flow matching and MSE losses, respectively. We jointly train the encoders, decoders, and flow prior by sampling one feasible reconstruction or prior-learning route per iteration from the modalities observed in the batch. In Monte Carlo training, we sample both reconstruction and prior-learning routes with a 3:2 sampling ratio, which empirically improves conditional generation while preserving prior coherence.

Sampling. For image generation, we use Euler sampling with CFG scale 4.0 and 25 inference steps. For text generation, we use greedy sampling.

Evaluation. We measure fidelity with FID [19] for images, FAD [26] for audio, and CIDEr [51] for text. For cross-modal alignment, we use CLIPScore [18] for text-image pairs, CLAPScore [12] for text-audio pairs, and AIS [54] for audio-image pairs.

One-to-one results. Tabs. 8 and 9 report the full one-to-one fidelity and alignment results. In fidelity, MUNI and MUNI† achieve the best or second-best generalist performance on most directions, with particularly large gains in text-to-image generation (14.40 FID vs. 20.13 for FlowBind [3]) and strong results on image-to-audio and audio-to-image generation (1.90 FAD and 22.97 FID). For image-to-text, MUNI† obtains competitive CIDEr scores, while UnifiedIO2-L [36] and FlowBind achieve the best scores on image-to-text and audio-to-text, respectively. The high CIDEr score of UnifiedIO2-L is expected, as it relies on an LLM-based decoder specialized for text generation.

For alignment, MUNI† is strongest on audio-related directions, achieving the best generalist CLAP score for text-to-audio (31.81) and the best AIS scores for image-to-audio and audio-to-image (90.11 and 95.61). It also remains competitive on text-image alignment, and both MUNI variants outperform the image-text model Diffuse Everything on all four text-image metrics. Overall, these results show that our framework remains competitive on standard one-to-one generation while also providing stronger many-to-one aggregation and unconditional co-generation, as shown in the main text.

Table 7: **Training configuration across experiments.** Hyperparameters are shared between MUNI and the multimodal VAE baselines under matched-compute comparisons. Reported runtimes are measured on a single RTX PRO 6000 GPU.

	Gaussian-mixture	**PolyMNIST-Quadrant-Labels**	**Image-Text-Audio**
Optimization			
Optimizer		AdamW (β_1=0.9, β_2=0.95)	
Learning rate	4×10^{-3}	6×10^{-4}	4×10^{-4}
LR schedule		constant, 5-epoch linear warmup	
Weight decay	0.0	0.0	0.0
Grad clip (ℓ_2)	1.0	1.0	1.0
Epochs	200	300	200
Batch size	4000	1024	1024
Total Runtime	$\sim$10 min	$\sim$3 h	$\sim$3 h
EMA decay	0.999	0.999	0.999
Encoders			
Architecture	MLP	ResNet (images), MLP (labels)	MLP
Aggregation	PoE	PoE	{PoE, Hellinger}
Total params	4.6 M	8.8 M	90 M
Prior			
Architecture	MLP	MLP	MLP
Encoder grad scale	0.1; 0.01 for ShaLa [9]	0.1	0.01
Total params	3.2 M	12.6 M	30 M
Decoders			
Architecture	MLP	NCSN++ (images), MLP (labels)	MLP (image, audio), Transformer (text)
Total params	14.0 M	10.7 M	140 M

E.3 Component-Wise Ablations

We change one component of MUNI at a time: extending prior supervision from full-modality and leave-one-out subsets to arbitrary non-empty subsets, replacing non-mixture aggregation with MoE, and removing target detachment during self-reconstruction. All variants use the same data, architecture, optimization, and evaluation protocol as the corresponding full model.

The PolyMNIST results in Tab. 10 are consistent with our analysis in App. A.1. Prior supervision on arbitrary subsets preserves multi-label alignment but reduces unconditional coherence, since latents inferred from smaller subsets need not capture dependencies among the missing modalities. MoE sharply reduces multi-label alignment, as it does not combine all observed evidence in one

Table 8: **Full comparison of one-to-one generation fidelity on image-text-audio.** Specialists and Diffuse Everything are shown as reference rows, and multimodal VAE baselines use the same backbone and compute as MUNI. Gray marks UnifiedIO2-L results, which do not support audio-image translation and are not ranked. **Bold** and underline mark the best and second-best generalist results.

Category	Method	T→I FID ↓	I→T CIDEr ↑	T→A FAD ↓	A→T CIDEr ↑	I→A FAD ↓	A→I FID ↓
Specialists	FLUX.1 [1]	24.81	–	–	–	–	–
	LLaVA-NeXT [35]	–	44.60	–	–	–	–
	TangoFlux [23]	–	–	2.47	–	–	–
	Qwen2-Audio [8]	–	–	–	21.39	–	–
	Seeing & Hearing [57]	–	–	–	–	7.47	–
	Sound2Vision [48]	–	–	–	–	–	25.59
Image-text model	Diffuse Everything [43]	16.69	11.65	–	–	–	–
Generalists	UnifiedIO2-L [36]	22.03	91.36	7.13	15.18	–	–
	CoDi [50]	24.97	11.97	8.67	7.70	13.92	53.60
	OmniFlow [31]	20.52	26.60	4.08	31.42	4.87	97.56
	FlowBind [3]	20.13	42.57	4.64	**54.41**	2.12	28.47
	MUNI **(Ours)**	14.73	57.72	**3.52**	43.72	1.98	**22.61**
	MUNI† **(Ours)**	**14.40**	**58.31**	**3.52**	47.38	**1.90**	22.97
Multimodal VAEs	MMVAE [46]	14.94	59.34	3.55	47.43	1.92	23.09
	MoPoE [49]	14.60	59.74	3.46	50.34	2.04	23.57

Table 9: **Full comparison of one-to-one cross-modal alignment on image-text-audio.** Specialists and Diffuse Everything are shown as reference rows, and multimodal VAE baselines use the same backbone and compute as MUNI. Gray marks UnifiedIO2-L results, which do not support audio-image translation and are not ranked. **Bold** and underline mark the best and second-best generalist results.

Category	Method	T→I CLIP ↑	I→T CLIP ↑	T→A CLAP ↑	A→T CLAP ↑	I→A AIS ↑	A→I AIS ↑
Specialists	FLUX.1 [1]	30.36	–	–	–	–	–
	LLaVA-NeXT [35]	–	29.47	–	–	–	–
	TangoFlux [23]	–	–	46.96	–	–	–
	Qwen2-Audio [8]	–	–	–	45.45	–	–
	Seeing & Hearing [57]	–	–	–	–	51.96	–
	Sound2Vision [48]	–	–	–	–	–	97.40
Image-text model	Diffuse Everything [43]	29.09	21.90	–	–	–	–
Generalists	UnifiedIO2-L [36]	29.73	29.97	15.22	35.48	–	–
	CoDi [50]	29.70	26.05	16.72	32.72	59.45	84.47
	OmniFlow [31]	**30.76**	27.48	30.07	**44.74**	76.52	64.22
	FlowBind [3]	28.07	29.14	27.13	41.80	89.23	93.94
	MUNI **(Ours)**	29.94	29.05	31.45	39.38	89.92	95.59
	MUNI† **(Ours)**	29.92	**29.18**	**31.81**	39.45	**90.11**	**95.61**
Multimodal VAEs	MMVAE [46]	29.68	29.14	30.09	38.30	90.05	95.11
	MoPoE [49]	29.59	29.05	31.43	38.98	90.82	94.20

Table 10: **Component-wise ablations on PolyMNIST-Quadrant-Labels.** Metrics follow Tab. 2. **Bold** and underline mark the best and second-best results.

	Single-L → I		Multi-L → I	Uncond.
Variant	Digit ↑	Quadrant ↑	Both ↑	Coherence ↑
MUNI	0.9131	0.9999	0.9346	**0.4841**
All-subset prior supervision	0.9253	**1.0000**	**0.9389**	0.3988
MoE aggregation	**0.9288**	**1.0000**	0.1647	0.3148
No target detachment	0.8938	0.9998	0.9212	0.4612

Table 11: **Component-wise ablations of MUNI† on image-text-audio.** Metrics follow Tabs. 3 and 4. **Bold** and underline mark the best and second-best results.

	(I+A)→T		(T+A)→I		(T+I)→A		Uncond.		
Variant	CLIP ↑	CLAP ↑	CLIP ↑	AIS ↑	CLAP ↑	AIS ↑	T-I ↑	T-A ↑	A-I ↑
MUNI†	**27.60**	**36.61**	**26.43**	**90.94**	**30.16**	**86.89**	**26.74**	**24.83**	**82.80**
All-subset prior supervision	26.26	35.04	26.42	90.74	29.35	85.80	26.13	22.42	81.93
MoE aggregation	26.26	36.20	25.55	87.83	23.76	83.19	26.28	22.25	75.00
No target detachment	25.82	34.87	26.24	89.59	28.74	85.91	25.21	21.70	77.25

latent sample. Removing target detachment restores the incentive to encode target-private details for self-reconstruction and slightly lowers all four metrics.

On image-text-audio, MUNI† outperforms all three ablations on every many-to-one alignment and unconditional coherence metric (Tab. 11). Replacing non-mixture aggregation with MoE causes the largest alignment losses when generating an image or audio target from two sources, consistent with its failure to combine both sources in one latent sample.

E.4 One-to-Many Generation

Our coherence analysis concerns full-modality and leave-one-out latents used for prior learning. When several targets are missing, their shared dependencies need not be determined by the observed subset, so this guarantee does not extend to one-to-many generation. However, MUNI infers a latent from any observed subset, so it can generate the targets one at a time by autoregressive (AR) decoding without retraining. Given source a and targets b, c, MUNI-AR first generates b from a and then re-encodes (a, b) to generate c. With exact conditional models, either order samples from the joint conditional $p(b, c \mid a) = p(b \mid a)\, p(c \mid a, b) = p(c \mid a)\, p(b \mid a, c)$.

Table 12: **One-to-many generation on image-text-audio.** Sources are drawn from the synthetic triplets used in Tab. 3. Within each route, the first two metrics measure source-target alignment and the third measures target-target coherence. Each route is written as $a \to (b + c)$. **Bold** and underline mark the best and second-best results.

	T→(I+A)			I→(T+A)			A→(T+I)		
Method	CLIP ↑	CLAP ↑	AIS ↑	CLIP ↑	AIS ↑	CLAP ↑	CLAP ↑	AIS ↑	CLIP ↑
CoDi [50]	26.11	16.99	63.42	25.83	59.97	7.42	31.83	82.81	23.41
OmniFlow [31]	25.49	26.93	76.94	22.12	79.45	14.78	**42.47**	70.56	22.70
FlowBind [3]	23.78	23.61	70.79	27.78	69.11	22.96	41.75	88.93	25.82
MUNI	**26.58**	31.09	81.67	**28.03**	85.02	25.98	39.35	93.24	25.18
MUNI-AR ($a \to b \to c$)	26.57	29.49	83.83	27.99	**88.48**	**33.90**	39.49	**93.90**	**25.94**
MUNI-AR ($a \to c \to b$)	25.98	**31.43**	**87.21**	25.60	85.00	24.29	39.76	93.23	25.54

Tab. 12 compares AR decoding with parallel decoding, which generates both targets independently given one latent of a. Even with parallel decoding, MUNI achieves higher target-target coherence than all baselines on T→(I+A) and I→(T+A) and remains competitive with FlowBind on A→(T+I). With AR decoding, MUNI achieves the highest coherence on all three routes.

F More Qualitative Results

We provide additional qualitative results for image-text-audio any-to-any generation in Figs. 4–11. MUNI† shows strong source alignment in one-to-one and many-to-one generation, and produces more coherent multimodal triples in unconditional co-generation. For qualitative examples involving audio, we include the corresponding raw audio files in the supplementary material. For tasks where audio is the target modality, i.e., text-to-audio, image-to-audio, and (text+image)-to-audio, we provide the full qualitative outputs, including the associated text and image results, in the supplementary material.

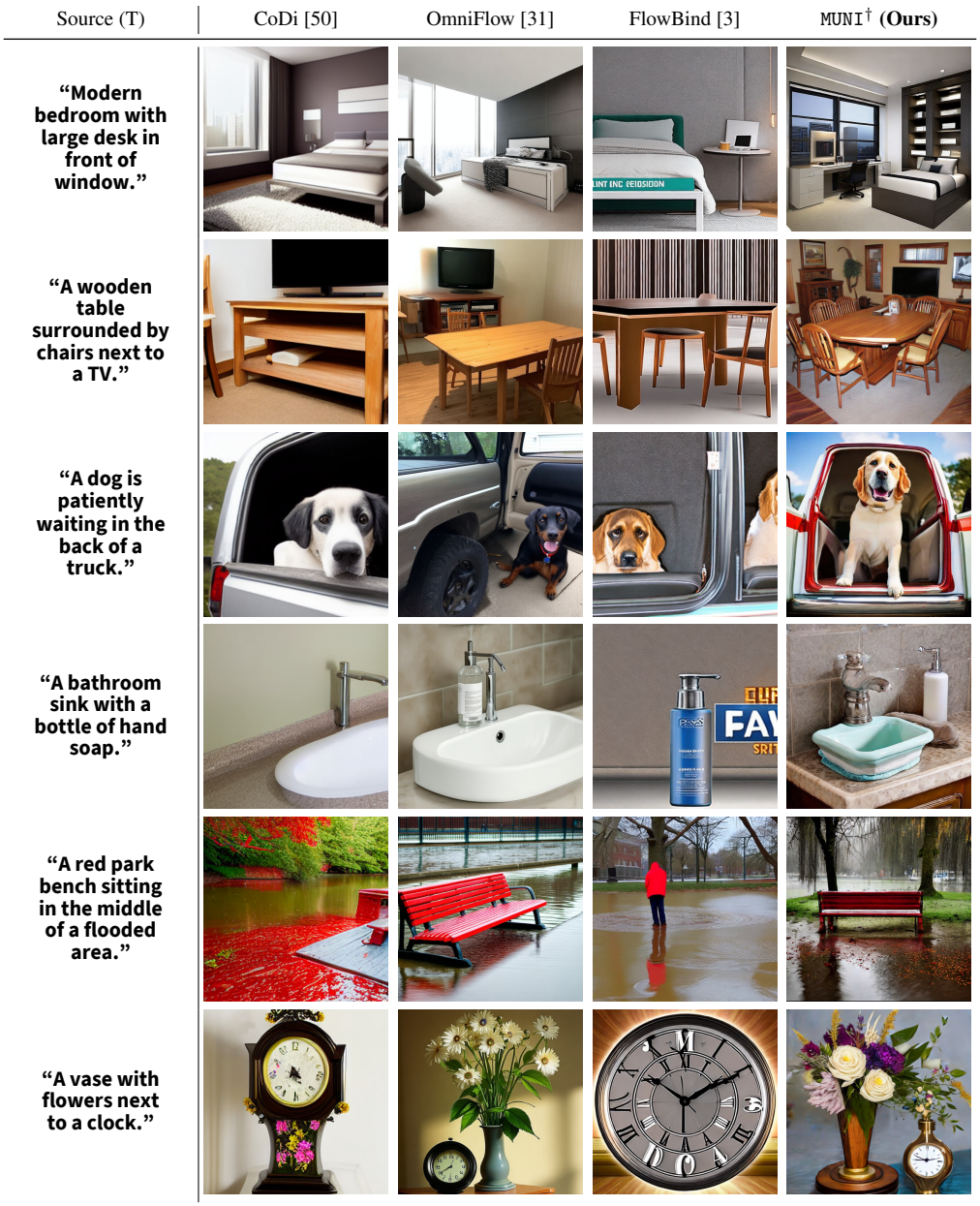


Figure 4: **Text-to-image generation.** Given a source text (left), each method generates an image. MUNI† produces an image semantically aligned with the source.

Source (I)	CoDi [50]	OmniFlow [31]	FlowBind [3]	MUNI† (Ours)
	"Mike mitt kicks his playoff pitch during the final baseball game of the year."	"A young boy throwing a baseball with a baseball bat."	"A boy in a white baseball uniform, shirt and visor is holding a baseball bat."	"A young boy in a baseball uniform is excited after catching a ball."
	"Dog and a bag lunch on coffee."	"A paper bag with a sandwich and a knife sitting on a wooden table."	"A helmet with books on it sitting on a table with a dog and coffee."	"A cup and donut on a table with books."
	"Apple with peanuts."	"A pan filled with fried rice and shrimp."	"A pan with some banana flour on it."	"A skillet with bananas and syrup being cooked."
	"On a hill with horses and bicyclists on a truck four people don't stop."	"A herd of cows walking down a road in a forest, with a truck parked nearby."	"A man riding a bucking cow is driving down a road past a herd of cows."	"A man in a tractor with cows on the road."
	"All women finals at the 2012 olympic skiing, multiple women in the 100 olympic."	"A person is skiing through a snowy landscape with footprints."	"A man skis down a snowy hill as he is sliding off of a race track."	"A skier in purple is sliding down the slope."
	"Two guys making some pizza on the pizza."	"Two children are sitting at a table with dishes."	"A little girl is cooking pizza with a plate, while a toddler is making it."	"Two little girls are making pizzas in the kitchen."

Figure 5: **Image-to-text generation.** Given a source image (left), each method generates a caption. MUNI† produces a caption that accurately describes the source.

Source (A)	CoDi [50]	OmniFlow [31]	FlowBind [3]	MUNI† **(Ours)**
 *(Baby crying with women talking.)*	**"A baby is crying and mother is born."**	**"A woman and baby talking."**	**"A infant crying and a woman speaking."**	**"A woman is talking to a baby while he cries."**
 *(A person talks; a man talks while typing on a keyboard.)*	**"People typing."**	**"A man speaks and types on a keyboard."**	**"A woman speaks and then a person talks keyboard."**	**"Two people talking and typing on a computer keyboard."**
 *(Motorboat engine vibrating.)*	**"One dog coming to turn for a speeding car are speeding away."**	**"Wind blows and an engine hums."**	**"A boat motor is cruising then slows somewhat loudly."**	**"A motorboat engine running and revving."**
 *(A cat makes a sound, then wind blows.)*	**"Cat starts a cat < baby."**	**"A cat meows and whines."**	**"Loud meowing of cat, then meowing."**	**"A cat is howling and the wind is blowing."**
 *(A man talks to an audience, then crowds applaud.)*	**"The man laughs to congratulate the man who says that the election is about to yell."**	**"Man speaking into a microphone."**	**"A man speaks with people talking then he applauds."**	**"A man is speaking to an audience while others clap."**
 *(Fireworks exploding.)*	**"Jump happening in a fire over a falling lake and blue rainbow."**	**"Rain falls and a pop occurs."**	**"Multiple popping and eruption."**	**"Fireworks burst in the sky and then a burst."**

Figure 6: **Audio-to-text generation.** Given a source audio (left), each method generates a caption.

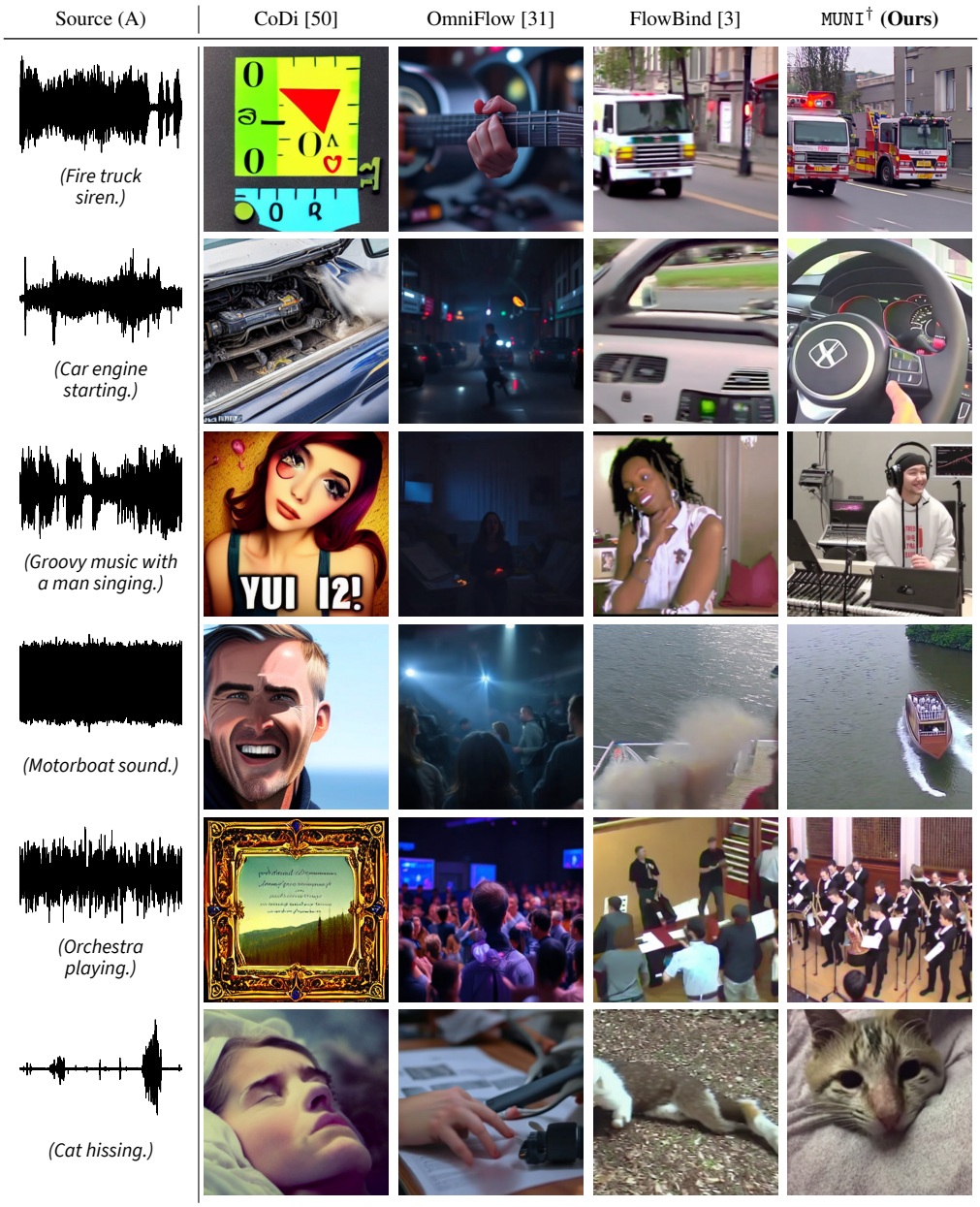


Figure 7: **Audio-to-image generation.** Given a source audio (left), each method generates an image.

Source (I)	Source (A)	CoDi [50]	OmniFlow [31]	FlowBind [3]	MUNI† **(Ours)**
	 *(Car engine sound with a man talking.)*	**"A man stops for a train about to go on a car being driven by a man."**	**"A man sits in a truck, working on an engine."**	**"The vehicle motor is driving down the road."**	**"A man walking next to a small car."**
	 *(A man whistles and is playing a guitar.)*	**"A man is speaking to a person in the city talking while walking for the man."**	**"A man talking with a crowd talking in the background."**	**"A man sitting at a park bench while music is playing."**	**"A man playing an acoustic guitar and another person sitting on the bench."**
	 *(Helicopter sound.)*	**"Airplane."**	**"A helicopter hovering in the distance, with."**	**"A helicopter flying over with muffled helicopter."**	**"A helicopter is flying through the air."**
	 *(Boat engine sound with a man talking.)*	**"Man driving in a blue car race to turn mid-way on his car engine, is speeding."**	**"A man is working on a boat in the water."**	**"A man motor is on a boat in the water boating."**	**"A man in sunglasses is driving on the water."**
	 *(A man is speaking with crowd noise in background.)*	**"A man is jaying after a dog says to his dog greeting. as he walks on the dog."**	**"A man talking in a movie, with a blurry background."**	**"A man speaking with someone."**	**"A man is talking to some people and there is a dog."**
	 *(Car siren.)*	**"Red traffic car for a road red."**	**"A darkened city street with a siren blaring in the distance,."**	**"A police siren is signaling while sirens are."**	**"A police car is going by with sirens."**

Figure 8: **(Image + Audio) → Text generation.** Given an image and an audio clip as joint sources (left), each method generates a caption.

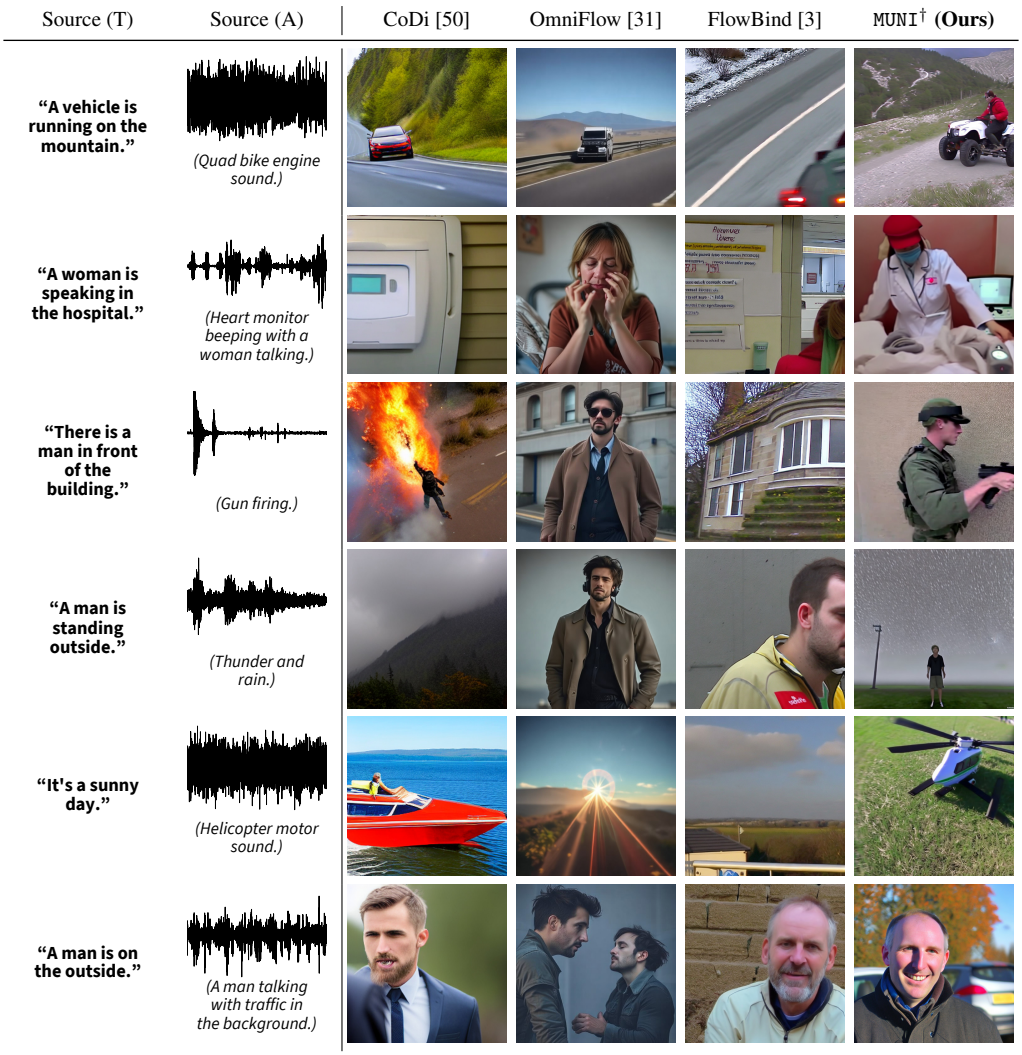


Figure 9: **(Text + Audio) → Image generation.** Given a text and an audio clip as joint sources (left), each method generates an image.

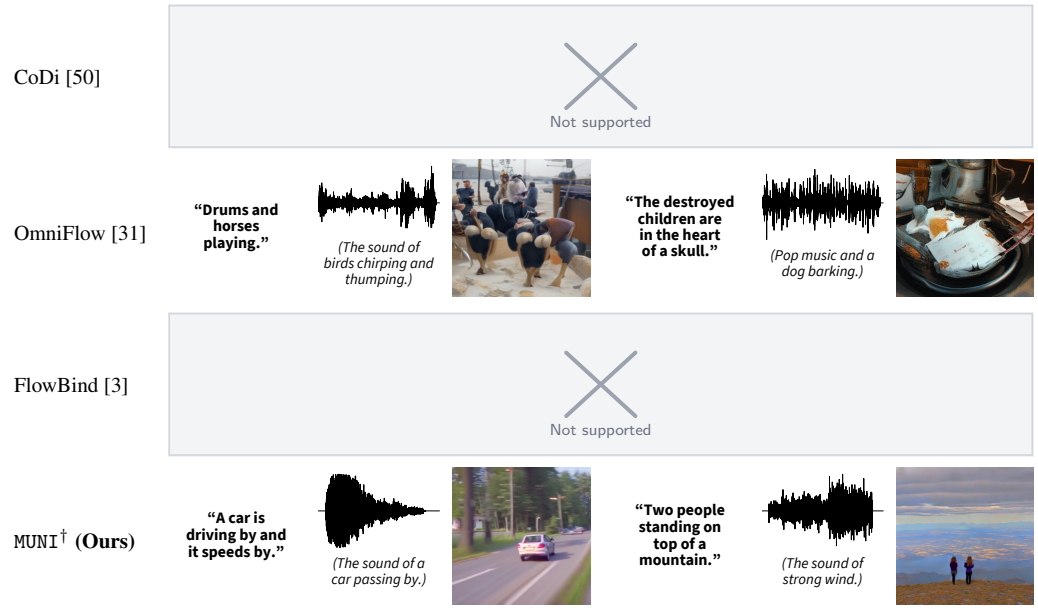


Figure 10: **Unconditional co-generation among generalists.** Each row shows two jointly sampled (text, audio, image) triples from a method's prior. CoDi [50] and FlowBind [3] do not support fully unconditional joint sampling.

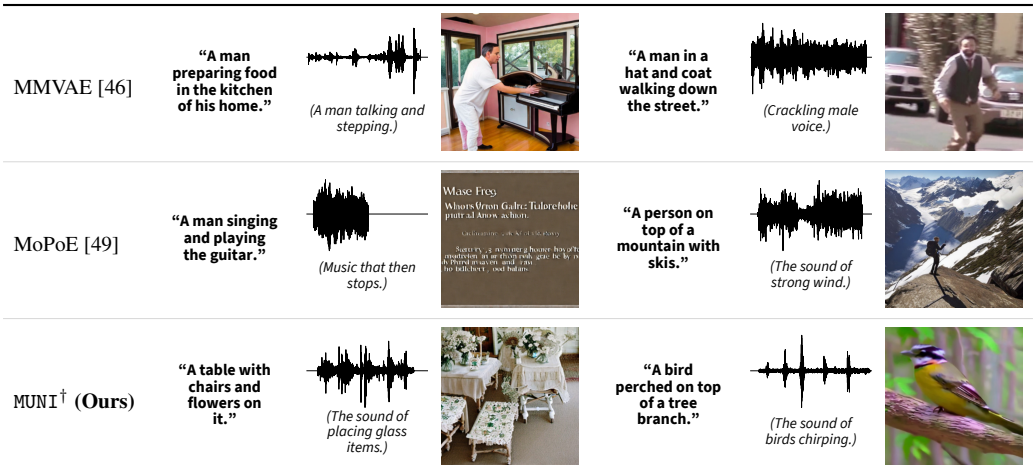


Figure 11: **Unconditional co-generation against multimodal VAE baselines.** Each row shows two jointly sampled (text, audio, image) triples. The multimodal VAE baselines share the same backbone and compute as MUNI†.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims match the proposed method, theory, and experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations and future directions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Appendix B states the assumptions and provides the full proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper reports the datasets, evaluation protocols, architectures, and key training hyperparameters needed to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code and datasets are publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental details and hyperparameters are in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars or statistical significance tests, following common practice in large-scale any-to-any generative model evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Training configurations and runtime estimates are reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential benefits and misuse risks.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release high-risk model checkpoints or raw scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used assets are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The benchmark construction and evaluation protocol are documented.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs not used in core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.