# MUNITE: UNIFIED MULTIMODAL LATENT INFERENCE FOR ANY-TO-ANY MULTIMODAL GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We introduce MUNITE, a latent-variable framework for flexible any-to-any multimodal generation that treats encoding and latent generation as the same inference problem under different amounts of observed evidence. Given any subset of modalities, MUNITE models the conditional distribution over the latent representation associated with the complete observation. Full observation recovers deterministic encoding, no observation recovers the latent marginal, and intermediate subsets define conditional latent inference, all within a single conditional flow model. A shared latent sample captures variation that must remain consistent across generated targets, while modality-specific generative decoders model the remaining uncertainty independently. To learn these conditional distributions from incomplete training examples, we extend conditional flow matching through self-distillation: predictions conditioned on richer available observations supervise the same model conditioned on smaller subsets at the same intermediate latent state. When the richer-evidence trajectory follows the exact conditional flow, this provides the same expected learning signal as full-target denoising. Across PolyMNIST-D-Q, FFHQ64, and image–text–audio, MUNITE achieves competitive or better generation quality and source–target alignment, with higher joint-generation coherence. In particular, it attains the highest coherence in all one-to-many and unconditional image–text–audio comparisons, showing the effectiveness of unified latent inference across diverse multimodal settings.

## 1 INTRODUCTION

Amid rapid progress in large foundation models for multimodal generation, an important complementary direction is to develop flexible any-to-any multimodal generative models without requiring a large language model as the multimodal backbone. Many practical applications may depend on only a subset of the capabilities offered by general-purpose language models, often without advanced text reasoning, coding, or mathematical problem solving, and in some cases with little or no text processing. Such applications call for models specialized to the modalities at hand that still support arbitrary combinations of observed and generated modalities.

A natural alternative is to build multimodal generation around a shared latent space. Latent space modeling has long been successful for images and videos (Rombach et al., 2022; Blattmann et al., 2023), has recently been explored for language generation through diffusion models (Lovelace et al., 2023), and is now emerging as a promising basis for any-to-any multimodal generation (Yeo et al., 2026; Cha et al., 2026). Recent approaches further blur the boundary between representation learning and generation by jointly learning modality encoders, decoders, and the latent generative model, rather than fixing the latent representation before training the generator (Heek et al., 2026; Yeo et al., 2026). This suggests a broader possibility: instead of treating multimodal encoding and latent generation as separate components, can they be viewed as instances of the same underlying modeling problem?

We answer this question by **unifying these roles within a single conditional model**. Our key observation is that full-observation encoding, partial-observation latent inference, and unconditional latent generation correspond to different amounts of evidence about the same latent representation (Figure 1). Given an arbitrary subset of modalities, our model learns the conditional distribution over latent representations of complete observations. When all modalities are observed, this distribution



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

collapses to a point mass, recovering deterministic multimodal encoding. When no modality is observed, it becomes the marginal latent distribution, recovering unconditional generation. Every intermediate subset corresponds to conditional latent inference, so encoding, conditional inference, and unconditional generation form a spectrum indexed by the observed modality subset. This view generalizes the Generative Encoder of UNITE (Duggal et al., 2026), whose single network covers only the two endpoints of full and absent observation, to every level of partial observation. The inference model is coupled with modality-specific generative decoders through a shared latent: one latent sample captures the variation that must be shared across generated targets, while the decoders model the remaining modality-specific uncertainty.

This formulation also creates a natural mechanism for self-distillation across observation levels. A partially observed training example does not provide the latent representation of its complete observation as a direct training target. Instead, the model conditioned on a richer set of available modalities provides a clean-latent target for the same model conditioned on a smaller subset, evaluated at the same intermediate flow state. Under the assumptions developed in Section 2.3, we show that this self-distillation objective yields the same expected gradient as denoising against the unavailable full-observation latent. This allows partially observed examples to contribute directly to learning the desired conditional distribution.

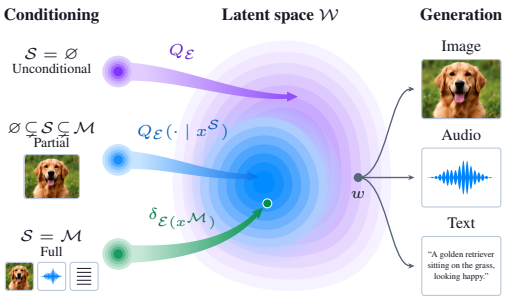



Figure 1: **Latent inference and decoding.** Left: latent distributions under different observation subsets. Right: factorized decoding from a latent value.

We evaluate MUNITE on PolyMNIST-D-Q, FFHQ64, and image–text–audio generation, which together span categorical attributes, dense visual modalities, and heterogeneous text, image, and audio, with training data that mix fully and partially observed examples. MUNITE attains the best unconditional fidelity and coherence on PolyMNIST-D-Q and FFHQ64 and the highest coherence in every one-to-many and unconditional image–text–audio comparison, while remaining competitive on one-to-one and many-to-one generation. These results show that a single conditional latent-inference model can support the full range of observation regimes in any-to-any generation, improving joint coherence while retaining the benefits of modality-specific generative decoders.

## 2 Method

MUNITE learns any-to-any generation from a mixture of fully and partially observed multimodal data. We first introduce a single latent inference model that, for any observed subset of modalities, captures the conditional distribution over latent representations of complete observations. We pair it with modality-specific generative decoders, so that the shared latent carries cross-modal dependencies and the decoders model the remaining uncertainty (Section 2.1). We learn the latent representation through cross-modal reconstruction (Section 2.2) and its evidence-conditioned distributions through self-distillation, which extends conditional flow matching to partially observed examples (Section 2.3). An auxiliary contrastive loss further encourages alignment across observation subsets (Section 2.4).

### 2.1 Shared-Latent Generative Framework

Let $\mathcal{M} = \{1, \dots, M\}$ index the modalities and $X^\mathcal{M} \sim P_{\text{data}}$ denote a complete observation. We learn a latent representation $W$ of complete observations together with a single model $\mathcal{E}$ that operates on an arbitrary observed subset $\mathcal{S} \subseteq \mathcal{M}$. A partial observation $x^\mathcal{S}$ generally admits multiple possible completions $x^\mathcal{M}$, which in turn induce a conditional distribution $Q_\mathcal{E}(\cdot \mid x^\mathcal{S})$ over their latent representations.

At the two extremes, the same model $\mathcal{E}$ reduces to familiar operations. When $\mathcal{S} = \mathcal{M}$, $\mathcal{E}$ acts as a deterministic encoder, yielding a unique latent representation $w = \mathcal{E}(x^\mathcal{M}) \in \mathbb{R}^D$. Accordingly, the conditional distribution collapses to $Q_\mathcal{E}(\cdot \mid x^\mathcal{M}) = \delta_w$. When $\mathcal{S} = \emptyset$, the model instead generates from the latent marginal $Q_\mathcal{E}(\mathrm{d}w)$. For an intermediate subset $\emptyset \subsetneq \mathcal{S} \subsetneq \mathcal{M}$, it models $Q_\mathcal{E}(\mathrm{d}w \mid x^\mathcal{S})$

(Figure 1, left). We parameterize $\mathcal{E}$ as a conditional flow model that transports a standard Gaussian to the corresponding distribution $Q_{\mathcal{E}}(\cdot \mid x^{\mathcal{S}})$ for any observed subset $\mathcal{S}$.

**A generative-encoder perspective.** Motivated by the success of latent diffusion models (Rombach et al., 2022), Unified Latents (Heek et al., 2026) proposed jointly learning latent representations of the data and a generative model of their distribution. UNITE (Duggal et al., 2026) took this idea further by introducing a *Generative Encoder* that uses the same model to encode an observed image and, when no image is observed, to generate samples from the latent distribution. Multimodal data add a natural intermediate regime in which only a subset of modalities is observed, and MUNITE extends the generative-encoder view to it by modeling the conditional distribution of the full-observation latent given an arbitrary observed subset. Full-observation encoding and unconditional latent generation then become the endpoints $\mathcal{S} = \mathcal{M}$ and $\mathcal{S} = \emptyset$ of a family of latent-inference problems $\{Q_{\mathcal{E}}(\cdot \mid x^{\mathcal{S}})\}_{\mathcal{S}\subseteq\mathcal{M}}$.

To generate multimodal observations, we pair the latent inference model $\mathcal{E}$ with modality-specific generative decoders $\{\mathcal{D}_{\psi_m}\}_{m\in\mathcal{M}}$. As in multimodal VAEs (Wu & Goodman, 2018; Shi et al., 2019), the decoders generate independently conditioned on a shared latent. For this factorization to preserve the dependencies in the data, we require $W$ to be *coherence-sufficient* (Yeo et al., 2026):

$$P_{\text{data}}(\mathrm{d}x^{\mathcal{M}} \mid W = w) = \prod_{m\in\mathcal{M}} P_{\text{data}}(\mathrm{d}x^{m} \mid W = w). \tag{1}$$

The latent thus accounts for cross-modal dependence, leaving the remaining modality-specific variation to the decoders. For disjoint input and target sets $\mathcal{S}, \mathcal{T} \subseteq \mathcal{M}$, marginalizing over the latent gives

$$P_{\text{data}}(\mathrm{d}x^{\mathcal{T}} \mid x^{\mathcal{S}}) = \int \prod_{m\in\mathcal{T}} P_{\text{data}}(\mathrm{d}x^{m} \mid W = w)\, Q_{\mathcal{E}}(\mathrm{d}w \mid x^{\mathcal{S}}). \tag{2}$$

We train each decoder to model the conditional distribution of its modality given $W$. At inference, we use $\mathcal{E}$ to sample one latent $w$ conditioned on $x^{\mathcal{S}}$ and generate the target modalities in parallel with $\{\mathcal{D}_{\psi_m}(w)\}_{m\in\mathcal{T}}$.

### 2.2 Learning the Shared Representation

We learn the shared representation by reconstructing observed modalities through their generative decoders. We write $\mathcal{E}_\theta(w_t, t; x^{\mathcal{S}})$ for the clean-latent prediction of the model $\mathcal{E}$ with parameters $\theta$. For a training example $x^{\mathcal{A}}$, let the nonempty set $\mathcal{A} \subseteq \mathcal{M}$ denote its available modalities. For reconstruction, we sample a target modality $m \in \mathcal{A}$ and a conditioning subset $\mathcal{S} \subseteq \mathcal{A}$ as described in Appendix C.2. Given a Gaussian query $\epsilon \sim \mathcal{N}(0, I_D)$ at $t = 0$, the reconstruction objective is

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{(\mathcal{A},x^{\mathcal{A}}),\mathcal{S},m,\epsilon}\left[\ell_m\left(\mathcal{D}_{\psi_m}\left(\mathcal{E}_\theta(\epsilon, 0; x^{\mathcal{S}})\right), x^{m}\right)\right]. \tag{3}$$

Here $\ell_m$ denotes the training objective of decoder $m$, such as cross-entropy for text and labels or a denoising objective for diffusion and flow decoders. With $\mathcal{S} = \mathcal{M}$, this trains each decoder on the full-observation latents required by Equation 2, while smaller subsets further stabilize the learning of the shared representation.

To favor information useful across modalities, we use target-detached self-reconstruction (Yeo et al., 2026). When the target is also an input ($m \in \mathcal{S}$), it still contributes to the latent, but the reconstruction gradients through its token stream are stopped. Consequently, the target's own reconstruction loss cannot train its token stream to pass target-private information to the latent. We describe the implementation in Section 2.5.

### 2.3 Conditional Flow Matching with Partial Observations

We train $\mathcal{E}_\theta$ to model $Q_{\mathcal{E}_\theta}(\cdot \mid x^{\mathcal{S}})$ by conditional flow matching along a linear interpolant between standard Gaussian noise and clean latents. Since the clean target can be computed only from a complete example, we extend the objective to partial observations through self-distillation.

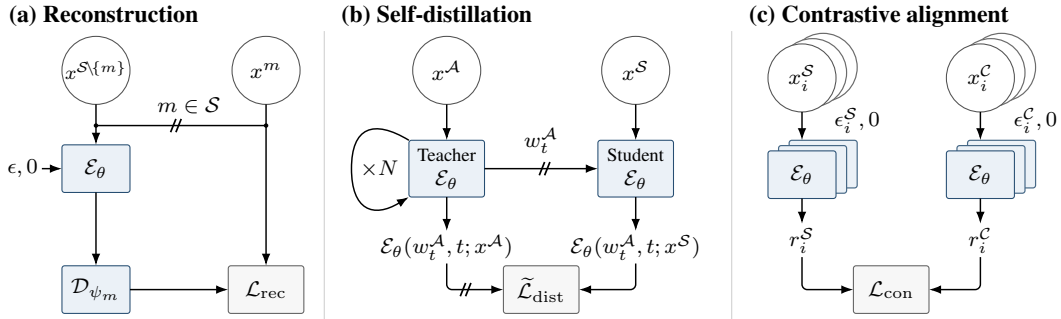



Figure 2: **MUNITE training objectives.** (a) Reconstruction learns a shared representation, using target detaching to discourage copying modality-private information. (b) Self-distillation uses detached predictions conditioned on more modalities to supervise subset-conditioned predictions at the same noisy latent state, with both branches using the same model. (c) An auxiliary contrastive loss encourages semantic alignment between complementary modality subsets using in-batch negatives.

**Fully observed examples.** The full-observation prediction provides the clean target $w = \mathcal{E}_\theta(\epsilon, 0; x^{\mathcal{M}})$, where $\epsilon \sim \mathcal{N}(0, I_D)$. We use the same Gaussian sample as the initial flow state, setting $w_0 = \epsilon$, and sample $t \sim \mathcal{U}[0, 1)$. The interpolant and the velocity field implied by the clean prediction are

$$w_t = (1-t)w_0 + t\,\mathrm{sg}(w), \qquad v_\theta(w_t, t; x^{\mathcal{S}}) = \frac{\mathcal{E}_\theta(w_t, t; x^{\mathcal{S}}) - w_t}{1-t}, \tag{4}$$

where sg denotes stop-gradient. Writing $W, W_0, W_t$ for the corresponding random variables, the conditional flow matching objective (Lipman et al., 2023) becomes

$$\mathcal{L}_{\mathrm{CFM}} = \mathbb{E}\big[\|v_\theta(W_t, t; X^{\mathcal{S}}) - (\mathrm{sg}(W) - W_0)\|_2^2\big] = \mathbb{E}\left[\frac{\|\mathcal{E}_\theta(W_t, t; X^{\mathcal{S}}) - \mathrm{sg}(W)\|_2^2}{(1-t)^2}\right]. \tag{5}$$

Clean-latent prediction allows encoding and denoising to share their output space, as in JiT and UNITE (Li & He, 2026; Duggal et al., 2026). For optimization stability, we drop the factor $(1-t)^{-2}$ and train with the unweighted denoising loss

$$\widetilde{\mathcal{L}}_{\mathrm{CFM}} = \mathbb{E}\big[\|\mathcal{E}_\theta(W_t, t; X^{\mathcal{S}}) - \mathrm{sg}(W)\|_2^2\big].$$

At a fixed encoder and time $t$, both weightings have the population-optimal prediction (Lipman et al., 2023)

$$\mu_{\mathcal{S}}(w_t, t; x^{\mathcal{S}}) := \mathbb{E}[W \mid W_t = w_t,\ X^{\mathcal{S}} = x^{\mathcal{S}}]. \tag{6}$$

With full observations, this reduces to $\mathcal{E}(x^{\mathcal{M}})$, recovering encoding as a one-step prediction (Section 2.1). As in UNITE, the Gaussian query keeps the input format common to all observation levels; since the reconstruction target does not depend on it, the trained models effectively ignore it (Appendix A.1).

**Partially observed examples.** When $\mathcal{A} \subsetneq \mathcal{M}$, we cannot compute $\mathcal{E}(x^{\mathcal{M}})$. The available modalities nevertheless constrain its conditional distribution. For nested subsets $\mathcal{S} \subsetneq \mathcal{A}$, the tower property gives, at each fixed $t$,

$$\mathbb{E}\big[\mu_{\mathcal{A}}(W_t, t; X^{\mathcal{A}}) \mid W_t, X^{\mathcal{S}}\big] = \mu_{\mathcal{S}}(W_t, t; X^{\mathcal{S}}). \tag{7}$$

A clean prediction using $x^{\mathcal{A}}$ can therefore supervise the prediction using $x^{\mathcal{S}}$, provided their common noisy state follows the full-latent intermediate distribution. We approximate it by integrating $\mathcal{E}_\theta$ conditioned on all available modalities up to the sampled time, using $N$ Euler steps with $N = 4$ for training efficiency:

$$\frac{\mathrm{d}w_u^{\mathcal{A}}}{\mathrm{d}u} = v_\theta(w_u^{\mathcal{A}}, u; x^{\mathcal{A}}), \qquad w_0^{\mathcal{A}} = w_0, \qquad 0 \le u \le t. \tag{8}$$

At the resulting state, the same model serves as the teacher, conditioned on $x^{\mathcal{A}}$, and as the student, conditioned on $x^{\mathcal{S}}$, giving the unweighted distillation objective

$$\widetilde{\mathcal{L}}_{\mathrm{dist}} = \mathbb{E}\left[\left\|\mathcal{E}_\theta(W_t^{\mathcal{A}}, t; X^{\mathcal{S}}) - \mathrm{sg}\left(\mathcal{E}_\theta(W_t^{\mathcal{A}}, t; X^{\mathcal{A}})\right)\right\|_2^2\right]. \tag{9}$$

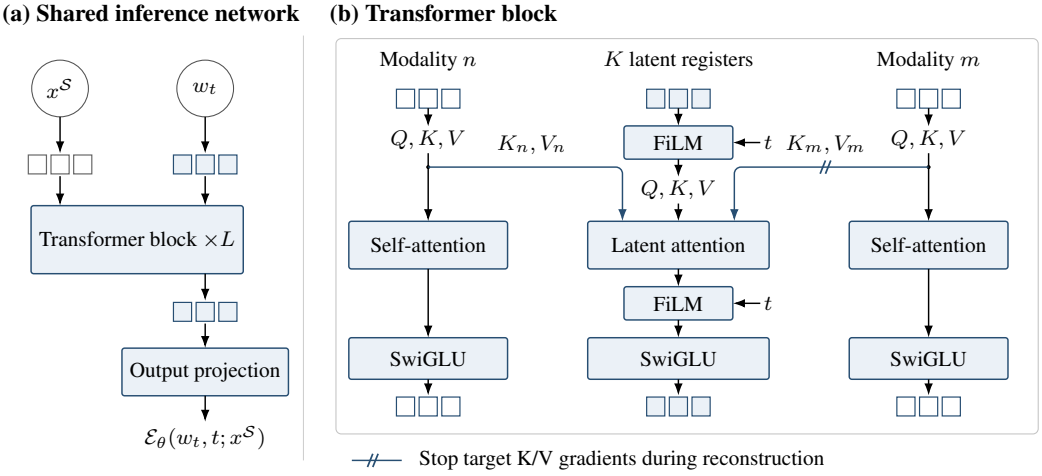



Figure 3: **Latent inference architecture.** A shared Transformer performs latent inference under arbitrary observation subsets by aggregating modality information in latent registers. Keeping modality streams separate allows reconstruction gradients to be stopped on the target's key/value paths to the registers (slashes), while retaining its information in the forward pass.

For a complete example, the teacher target is its encoding $w$ and the state is the analytic interpolant in Equation 4, recovering $\widetilde{\mathcal{L}}_{\mathrm{CFM}}$.

**Proposition 1** (Gradient equivalence)**.** *Let $\mathcal{S} \subsetneq \mathcal{A}$, let $x^{\mathcal{A}}$ be drawn from $P_{\mathrm{data}}(\mathrm{d}x^{\mathcal{A}})$, and treat the target latent distribution and teacher as fixed. If the teacher predicts $\mu_{\mathcal{A}}$ and its trajectory reproduces the conditional time marginals of $W_t$, then $\widetilde{\mathcal{L}}_{\mathrm{dist}}$ and $\widetilde{\mathcal{L}}_{\mathrm{CFM}}$ yield identical expected gradients through the subset-conditioned prediction.*

The proof follows from the orthogonality of conditional expectation (Appendix B). While the equivalence holds at a fixed representation with an exact teacher, in practice the current model serves as the teacher and its conditional flow is integrated numerically.

### 2.4 Contrastive Alignment across Observation Subsets

Reconstruction and denoising train $\mathcal{E}$ to retain cross-modal information and to model its conditional distributions, but neither directly shapes the latent geometry to reflect semantic similarity. We therefore add an auxiliary contrastive objective that promotes semantic alignment across observation subsets. For a batch of $B$ examples with available set $\mathcal{A}$, choose a nonempty proper subset $\mathcal{S} \subsetneq \mathcal{A}$ and its complement $\mathcal{C} = \mathcal{A} \setminus \mathcal{S}$. Let $r_i^{\mathcal{S}}$ and $r_i^{\mathcal{C}}$ be the flattened, $\ell_2$-normalized outputs of $\mathcal{E}_\theta(\epsilon_i^{\mathcal{S}}, 0; x_i^{\mathcal{S}})$ and $\mathcal{E}_\theta(\epsilon_i^{\mathcal{C}}, 0; x_i^{\mathcal{C}})$, using independent Gaussian queries. With temperature $\tau_c > 0$, the directional InfoNCE loss (van den Oord et al., 2018) is

$$\mathcal{L}_{\mathcal{S}\to\mathcal{C}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp((r_i^{\mathcal{S}})^\top r_i^{\mathcal{C}}/\tau_c)}{\sum_{j=1}^{B} \exp((r_i^{\mathcal{S}})^\top r_j^{\mathcal{C}}/\tau_c)}. \tag{10}$$

We average the two directions, $\mathcal{L}_{\mathrm{con}} = (\mathcal{L}_{\mathcal{S}\to\mathcal{C}} + \mathcal{L}_{\mathcal{C}\to\mathcal{S}})/2$. As in Section 2.2, these one-step predictions from partial observations are not samples of the target latent distribution, but because they share the backbone parameters, aligning them stabilizes the internal representation of $\mathcal{E}$.

### 2.5 Model Architecture

We implement $\mathcal{E}$ as a shared Transformer that predicts a clean latent from a latent state and an arbitrary subset of observed modalities. Figure 3 illustrates the architecture. Observed modalities are represented as separate token streams, while $K$ latent registers receive the Gaussian query or noisy latent state and aggregate information from the observed streams at every block. The final register states are projected to the clean latent prediction.

Within each block, modality streams attend only within themselves, whereas the latent registers attend to all observed streams and to one another, following the latent-attention construction of Perceiver (Jaegle et al., 2021). This separation prevents direct mixing between modality streams and allows their features to be reused across denoising steps.

For reconstruction target $m$, we detach its projected keys and values on every modality-to-register attention edge. The forward computation is unchanged, but reconstruction gradients can no longer flow from the registers into the target stream. Since modality streams do not attend to one another, no other modality offers an indirect path. This realizes the target-detached reconstruction of Section 2.2 without target-specific encoders. Appendix C.1 gives further implementation details.

### 2.6 Training and Inference

We train $\mathcal{E}$ and the modality-specific generative decoders with

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{dist}}\widetilde{\mathcal{L}}_{\text{dist}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}. \tag{11}$$

The reconstruction loss updates $\mathcal{E}$ and the decoder of the sampled target, while the distillation and contrastive losses update only $\mathcal{E}$. Algorithm 1 in Appendix C.2 gives the full training procedure.

At inference, one latent drawn by integrating $v_\theta$ from Gaussian noise is shared by all target decoders, as the factorization in Equation 2 requires.

### 2.7 Relation to Multimodal Latent-Variable Models

Among multimodal VAEs, which likewise decode modalities independently from a shared latent (Wu & Goodman, 2018; Shi et al., 2019), JNF (Senellart & Allassonnière, 2025) is particularly close to our formulation: it first learns a shared latent with a joint VAE trained on complete observations, then fits a separate normalizing-flow posterior for each modality to samples from the joint encoder. Writing these posteriors as $q_m(z \mid x^m)$, inference from a subset $\mathcal{S}$ is based on their product,

$$q_{\text{JNF}}(z \mid x^{\mathcal{S}}) \propto p(z)^{1-|\mathcal{S}|} \prod_{m\in\mathcal{S}} q_m(z \mid x^m), \tag{12}$$

from which JNF samples using HMC.

MUNITE instead directly models $Q_{\mathcal{E}}(\mathrm{d}w \mid x^{\mathcal{S}})$, the conditional distribution of the same full-observation latent for every $\mathcal{S}$, without composing modality-specific posteriors. Whereas JNF builds on a joint encoder trained on complete observations, MUNITE can also learn these conditionals from partially observed examples: the prediction under richer available evidence serves as a self-distillation target, justified by Equation 7 and Proposition 1.

MUNI (Yeo et al., 2026) shares our use of factorized generative decoders, target-detached reconstruction, and a single latent sample for all targets, but draws this sample from an aggregated posterior $q_{\phi,\mathcal{S}}(z \mid x^{\mathcal{S}})$ for an observed subset. The central condition of its predictive-sufficiency criterion is

$$I(X^{\mathcal{S}}; X^m \mid Z_{\mathcal{S}}) = 0 \qquad \forall m \notin \mathcal{S}, \tag{13}$$

which guarantees target-wise sufficiency but does not imply conditional independence among multiple missing targets. MUNITE instead samples $W \sim Q_{\mathcal{E}}(\cdot \mid x^{\mathcal{S}})$, where the full-observation latent $W$ is trained to satisfy Equation 1. The missing variation shared across targets is therefore sampled once in $W$, and marginalizing this latent yields the conditional joint in Equation 2. We formalize this distinction in Appendix A.3.

## 3 Related Work

Since Section 2.1 covers latent diffusion, Unified Latents, and UNITE (Rombach et al., 2022; Heek et al., 2026; Duggal et al., 2026), we focus here on multimodal generation.

**Any-to-any generation.** Any-to-any generation extends cross-modal synthesis beyond fixed modality pairs to arbitrary combinations of observed and target modalities (Tang et al., 2023; Bachmann

et al., 2024). Masked models such as 4M and 4M-21 learn these mappings by predicting target-modality tokens from observed-modality tokens with a shared Transformer (Mizrahi et al., 2023; Bachmann et al., 2024). Joint flow models instead couple modality trajectories during sampling, using continuous representations in OmniFlow and discrete tokens in NExT-OMNI (Li et al., 2025; Luo et al., 2026); our CFM and DFM baselines implement these two approaches in a controlled setting (Section 4). CoDi similarly couples modality-specific denoisers through cross-attention to aligned representations of the other modalities' noisy latents (Tang et al., 2023).

**Shared-latent generative models.** Multimodal VAEs model multiple modalities through a shared latent variable and modality-specific encoders and decoders (Wu & Goodman, 2018; Shi et al., 2019; Sutter et al., 2021; Palumbo et al., 2023; Vo & Valera, 2026). Building on this factorization, MUNI jointly trains modality-specific encoders, expressive generative decoders, and a shared flow prior, and introduces coherence sufficiency, predictive sufficiency, and minimality as criteria for the learned multimodal latent (Yeo et al., 2026). FlowBind takes a different latent-space approach, using bidirectional flows to connect modality representations through a common latent distribution learned from partially paired data (Cha et al., 2026). MUNITE instead directly models the conditional distribution of a coherence-sufficient full-observation latent under arbitrary observed subsets, replacing the diagonal-Gaussian parameterization of subset posteriors with conditional transport.

# 4 Experiments

We evaluate one-to-one, many-to-one, one-to-many, and many-to-many generation, as well as unconditional co-generation. Our central question is whether jointly generated targets agree on content that their inputs leave unspecified. We adapt the PolyMNIST–digit–quadrant (PolyMNIST-D-Q) and image–text–audio benchmarks of MUNI (Yeo et al., 2026) and introduce a new FFHQ64 benchmark with three image-like and two categorical modalities.

## 4.1 Experimental setup

We compare MUNITE with MUNI (Yeo et al., 2026) and joint continuous and discrete flow-matching baselines, CFM and DFM (Lipman et al., 2023; Gat et al., 2024). We train MUNI with its released training and sampling pipelines and retain its modality-specific decoder architectures in MUNITE. To compare shared-latent and joint generation, CFM and DFM implement the continuous and discrete flow-matching approaches used in OmniFlow and NExT-OMNI, respectively (Li et al., 2025; Luo et al., 2026). Both baselines use the same Transformer blocks as MUNITE but apply joint attention across all modality tokens. Since they do not use generative decoders, we increase the backbone depth to approximately match MUNITE's total number of trainable parameters. Because fully Transformer-based joint models may converge more slowly than latent models with generative decoders, we also train CFM and DFM for twice as many epochs as MUNITE (Appendix D.2). CFM represents categorical modalities with one-hot vectors, whereas DFM discretizes continuous modalities with separately trained, frozen VQ-VAEs (van den Oord et al., 2017). All four methods share the same training data, evaluation protocol, and frozen evaluators.

For the image–text–audio task, we additionally compare against FlowBind (Cha et al., 2026) and pretrained CoDi and OmniFlow (Tang et al., 2023; Li et al., 2025). Because CoDi and OmniFlow are pretrained on substantially larger datasets, we use their released models without retraining, take their single-target and unconditional scores from MUNI (Yeo et al., 2026), and evaluate their one-to-many coherence ourselves (Appendix E.3). The released FlowBind checkpoint is trained on less data than our benchmark, so we retrain it on our training data with its official pipeline.

The training data mix fully and partially observed examples (Table 4 in Appendix D). For PolyMNIST-D-Q and FFHQ64, we mask one or two of the three non-categorical modalities in three quarters of the training examples and keep the categorical labels. For image–text–audio, the corpora used by MUNI and FlowBind contain only modality pairs, so we add caption–image–audio triplets from E-MM1 (Broadbent et al., 2025). For evaluation, we use 10,000 complete PolyMNIST tuples, 5,000 complete FFHQ tuples, and the image–text–audio evaluation sets of MUNI (Yeo et al., 2026). Appendix F shows qualitative results.

Table 1: Results on PolyMNIST-D-Q (top) and FFHQ64 (bottom). Acc. and Coh. denote accuracy and coherence, $d$ and $q$ denote digit and quadrant, and scores for $m_i$ are averaged over the three views. FD uses verifier features, and Normal Err. is $1 - \cos$. Best results are in bold.

| Route | Input → output | Metric | MUNI | CFM | DFM | MUNITE |
|---|---|---|---|---|---|---|
| | | *PolyMNIST-D-Q* | | | | |
| Uncond. | All five modalities | FD ↓ | 3.0407 | 71.6626 | 19.9587 | **1.5230** |
| | | Digit Coh. (all) ↑ | 0.5799 | 0.0237 | 0.2973 | **0.9826** |
| | | Quad. Coh. (all) ↑ | 0.9041 | 0.9902 | 0.8222 | **1.0000** |
| $1 \to 1$ | $d \to m_i$ | Digit Acc. ↑ | 0.9843 | 0.2402 | 0.7224 | **0.9941** |
| | $q \to m_i$ | Quad. Acc. ↑ | **1.0000** | 0.9936 | 0.9372 | **1.0000** |
| $N \to 1$ | $(d, q) \to m_i$ | Both-label Acc. ↑ | 0.9925 | 0.2200 | 0.7325 | **0.9943** |
| $1 \to N$ | $d \to (m_0, m_1, m_2)$ | Quad. Coh. ↑ | 0.2528 | 0.9879 | 0.8798 | **1.0000** |
| | $q \to (m_0, m_1, m_2)$ | Digit Coh. ↑ | 0.1074 | 0.1597 | 0.4612 | **0.9904** |
| | | *FFHQ64* | | | | |
| Uncond. | All five modalities | RGB FID ↓ | 34.618 | 122.617 | 55.840 | **29.918** |
| | | Age Coh. ↑ | 0.5630 | 0.3068 | 0.4546 | **0.5692** |
| | | Gender Coh. ↑ | 0.9596 | 0.7328 | 0.8658 | **0.9610** |
| | | RGB–Seg. Coh. ↑ | 0.7762 | 0.7357 | 0.8084 | **0.8896** |
| | | RGB–Normal Err. ↓ | 0.0700 | 0.0603 | 0.0495 | **0.0491** |
| $1 \to 1$ | Age → RGB | Age Acc. ↑ | 0.5204 | 0.2506 | 0.4002 | **0.6274** |
| | Gender → RGB | Gender Acc. ↑ | 0.9684 | 0.6944 | 0.8518 | **0.9790** |
| $N \to 1$ | (Age, Gender) → RGB | Both Acc. ↑ | 0.5100 | 0.1842 | 0.3690 | **0.5898** |
| $M \to N$ | (Age, Gender) → | RGB–Seg. Coh. ↑ | 0.6038 | 0.7441 | 0.8142 | **0.8888** |
| | (RGB, Seg., Normals) | RGB–Normal Err. ↓ | 0.1106 | 0.0572 | **0.0492** | 0.0500 |

### 4.2 POLYMNIST-D-Q

The three $64 \times 64$ RGB views share digit identity and quadrant but differ in handwriting and background. Conditioning on either label leaves the other attribute unobserved, so joint generation requires the three views to agree on that missing attribute. For unconditional co-generation, coherence measures agreement among the three images and the jointly generated labels. Following MUNI (Yeo et al., 2026), we compute digit and quadrant accuracy and coherence using a frozen SegFormer verifier (Appendix D.4), while Fréchet distance (FD) in verifier-feature space measures image fidelity.

MUNITE retains strong single-image label accuracy while substantially improving agreement on the unobserved attribute (Table 1, top). Under quadrant conditioning, MUNI and MUNITE both reach 1.0000 quadrant accuracy for single images, yet the digit coherence of their jointly generated views is 0.1074 and 0.9904, respectively. Under digit conditioning, quadrant coherence likewise rises from 0.2528 to 1.0000. MUNI's scores are near the chance agreement of independent views, 0.10 and 0.25, so accurate single-view conditionals alone do not make the views agree (Section 2.7). Each MUNITE view still follows the data distribution of the missing attribute, and the views agree only by chance when decoded from separate latent samples (Appendix E.1).

MUNITE also achieves the lowest FD and the highest unconditional coherence for both labels. CFM and DFM attain high quadrant agreement on some routes, yet lag behind both latent models in digit accuracy and image fidelity despite matched trainable parameters and twice as many training epochs. Their joint backbones must model the handwriting and background of each view together with the shared attributes, which both latent models leave to image-specific decoders.

### 4.3 FFHQ64

We construct a five-modality task from FFHQ (Karras et al., 2019), combining $64 \times 64$ RGB faces with the age, gender, and segmentation annotations of FFHQ-Aging (Or-El et al., 2020) and pseudo-labeled surface normals. We evaluate RGB generation from either or both categorical attributes, joint generation of RGB, segmentation, and normals from both, and unconditional co-generation of all five modalities. Attribute accuracy measures agreement with the observed labels, and coherence measures

agreement among jointly generated modalities. RGB FID measures image fidelity, and normal error the geometric consistency between generated normals and RGB.

On FFHQ64, the gains extend from shared labels to dense cross-modal structure (Table 1, bottom). When conditioning on age and gender, MUNITE increases RGB–segmentation coherence from 0.6038 to 0.8888 and reduces normal error from 0.1106 to 0.0500 relative to MUNI. DFM attains a slightly lower normal error (0.0492 vs. 0.0500), but MUNITE attains much higher segmentation coherence (0.8888 vs. 0.8142). MUNITE also has the lowest unconditional RGB FID and the strongest unconditional coherence on all four cross-modal comparisons. CFM and DFM remain substantially weaker in RGB fidelity, as in PolyMNIST-D-Q.

### 4.4 IMAGE–TEXT–AUDIO

Table 2: Image–text–audio coherence between jointly generated modalities. $T$, $I$, and $A$ denote text, image, and audio. OmniFlow's unconditional scores are from Yeo et al. (2026, Table 4), and dashes denote unsupported unconditional generation. Higher is better, and best results are in bold.

| Method | $1 \to N$ coherence | | | Unconditional coherence | | |
|---|---|---|---|---|---|---|
| | $T \to (I, A)$ AIS | $I \to (T, A)$ CLAP | $A \to (T, I)$ CLIP | CLIP $(T, I)$ | CLAP $(T, A)$ | AIS $(I, A)$ |
| CoDi | 63.866 | 7.418 | 23.414 | — | — | — |
| OmniFlow | 77.027 | 14.781 | 22.697 | 21.17 | 14.23 | 50.95 |
| MUNI | 81.558 | 32.237 | 25.850 | 26.949 | 28.467 | 81.989 |
| CFM | 74.702 | 21.695 | 24.034 | 23.622 | 19.896 | 74.208 |
| DFM | 73.635 | 20.951 | 24.037 | 23.740 | 19.503 | 71.942 |
| FlowBind | 79.819 | 25.812 | 24.853 | — | — | — |
| MUNITE | **83.336** | **34.762** | **26.334** | **27.398** | **30.220** | **87.045** |

We follow the image–text–audio benchmark used by FlowBind and MUNI (Cha et al., 2026; Yeo et al., 2026), modeling fixed 768-dimensional features and evaluating the rendered images, captions, and audio. For one-to-one generation, we use FID, FAD, and CIDEr for fidelity and CLIP, CLAP, and Audio-Image Similarity (AIS) for cross-modal alignment. For many-to-one generation, we measure the alignment of the generated target with each source modality, and for one-to-many and unconditional co-generation, the alignment among jointly generated modalities, which we report as coherence.

MUNITE achieves the highest coherence in all three one-to-many directions and all three unconditional comparisons (Table 2). The improvement holds across text, image, and audio, indicating that the shared latent preserves dependencies among generated targets when their shared variation is not determined by the input. Ablating target detaching, contrastive alignment, or partial-data distillation leaves this one-to-many advantage over MUNI intact, whereas unconditional coherence relies most on target detaching and partial-data distillation (Appendix E.3.2). On the one-to-one and many-to-one routes, MUNITE is competitive with the compared baselines, including the externally pretrained models (Appendix E.3.1).

## 5 CONCLUSION AND FUTURE WORK

MUNITE unifies encoding and any-to-any multimodal generation as inference over a full-observation latent. Self-distillation extends the learning of its conditional distributions to incomplete training examples, and modality-specific decoders generate from a shared latent sample. Across the evaluated benchmarks, MUNITE improves coherence among jointly generated modalities while maintaining competitive single-target generation quality.

A natural next step is multimodal world modeling. The compact latent learned by MUNITE could serve as a shared world-state representation that can be inferred from arbitrary subsets of observations. Learning dynamics directly in this latent space would enable state estimation and future prediction under varying multimodal observations, while modality-specific decoders could render the predicted state into any desired modality.

## AI USE STATEMENT

We used generative AI tools to check for typographical errors, prepare figures and tables, identify weaknesses in the manuscript and suggest improvements, and generate descriptive labels for audio examples in qualitative figures. These labels serve only to present audio examples and are not part of our method. Our evaluation uses an existing benchmark unchanged, including its previously generated synthetic data. The authors reviewed and edited all AI-assisted outputs and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## ETHICS STATEMENT

Our experiments include face and audiovisual generation, which may reproduce biases present in the underlying datasets and can be misused to create deceptive synthetic media. In the FFHQ64 benchmark, age and gender are used only as dataset annotations for controlled evaluation and should not be interpreted as validated personal attributes. Generated examples should therefore be interpreted as synthetic outputs rather than representations of real individuals or verified attributes.

## REPRODUCIBILITY STATEMENT

Section 2 specifies the learning objectives, gradient routing, and inference procedure, and Appendix B proves Proposition 1. Appendix C describes the architectures, the sampling of training subsets, and the training update in Algorithm 1. Appendix D details how each benchmark and its observation groups are constructed, together with model sizes, optimization and inference budgets, and the evaluators. Appendix E defines all metrics and gives the sources of the external baseline results.

## REFERENCES

Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4M-21: An any-to-any vision model for tens of tasks and modalities. In *NeurIPS*, 2024.

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

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable Video Diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Jim Broadbent, Felix Cohen, Frederik Hvilshøj, Eric Landau, and Eren Sasoglu. EBind: A practical approach to space binding. *arXiv preprint arXiv:2511.14229*, 2025.

Yeonwoo Cha, Semin Kim, Jinhyeon Kwon, and Seunghoon Hong. FlowBind: Efficient any-to-any generation with bidirectional flows. In *ICLR*, 2026.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *ICASSP*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Shivam Duggal, Xingjian Bai, Zongze Wu, Richard Zhang, Eli Shechtman, Antonio Torralba, Phillip Isola, and William T. Freeman. End-to-end training for unified tokenization and latent denoising. *arXiv preprint arXiv:2603.22283*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *WACV*, 2025.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *NeurIPS*, 2024.

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

Jonathan Heek, Emiel Hoogeboom, Thomas Mensink, and Tim Salimans. Unified Latents (UL): How to train your latents. *arXiv preprint arXiv:2602.17270*, 2026.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. In *ICASSP*, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *ICML*, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *NAACL*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Interspeech*, 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. OmniFlow: Any-to-any generation with multi-modal rectified flows. In *CVPR*, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *CVPR*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *ICML*, 2023.

Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q. Weinberger. Latent diffusion for language generation. In *NeurIPS*, 2023.

Run Luo, Xiaobo Xia, Lu Wang, Longze Chen, Renke Shan, Jing Luo, Min Yang, and Tat-Seng Chua. NExT-OMNI: Towards any-to-any omnimodal foundation models with discrete flow matching. In *ICLR*, 2026.

David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively multimodal masked modeling. In *NeurIPS*, 2023.

Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *ECCV*, 2020.

Emanuele Palumbo, Imant Daunhawer, and Julia E. Vogt. MMVAE+: Enhancing the generative quality of multimodal VAEs without compromises. In *ICLR*, 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. EmbeddingGemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

Christoph Schuhmann, Andreas Köpf, Theo Coombes, Richard Vencu, Benjamin Trom, and Romain Beaumont. LAION COCO: 600M synthetic captions from LAION2B-en. https://laion.ai/blog/laion-coco/, 2022.

Agathe Senellart and Stéphanie Allassonnière. Bridging the inference gap in multimodal variational autoencoders. *Journal of Data Science, Statistics, and Visualisation*, 5(9), 2025.

Yuge Shi, N. Siddharth, Brooks Paige, and Philip H. S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *NeurIPS*, 2019.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021b.

Stability AI. Stable unCLIP 2.1. https://huggingface.co/docs/diffusers/api/pipelines/stable_unclip, 2023.

Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. Generalized multimodal ELBO. In *ICLR*, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *CVPR*, 2016.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In *NeurIPS*, 2023.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *CVPR*, 2015.

Huyen Thuc Khanh Vo and Isabel Valera. Hellinger multimodal variational autoencoders. In *AISTATS*, 2026.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2CLIP: Learning robust audio representations from CLIP. In *ICASSP*, 2022.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *NeurIPS*, 2018.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021.

Kyeongmin Yeo, Yunhong Min, and Minhyuk Sung. MUNI: Multimodal unified latent diffusion for coherent any-to-any generation. *arXiv preprint arXiv:2606.16408*, 2026.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Yijie Zhang, Yiyang Shen, and Weiran Wang. Disentanglement of variations with multimodal generative modeling. In *ICLR*, 2026.

# APPENDIX

## A FULL-OBSERVATION LATENTS AND CONDITIONAL GENERATION

This section expands on the formulation of Section 2.1. We first characterize the latent conditionals induced by the full-observation encoder, then derive the conditional joint in Equation 2 and contrast it with the target-wise criterion of MUNI.

### A.1 THE DISTRIBUTION INDUCED BY FULL OBSERVATIONS

For a fixed representation, $W = \mathcal{E}(X^{\mathcal{M}})$, so the conditional distributions in Section 2.1 are pushforwards of the data conditionals through the full-observation encoder:

$$Q_{\mathcal{E}}(\cdot \mid x^{\mathcal{S}}) = \mathcal{E}_{\#}\big[P_{\mathrm{data}}(\cdot \mid x^{\mathcal{S}})\big]. \tag{14}$$

With full observation this is $\delta_{\mathcal{E}(x^{\mathcal{M}})}$, and empty conditioning gives the marginal distribution of $W$.

**Query dependence.** As described in Section 2.3, the shared model also receives a Gaussian query under full observation, so the full-observation latent $\mathcal{E}_\theta(\epsilon, 0; x^{\mathcal{M}})$ need not be exactly deterministic. We therefore measure its dependence on the query by the variance ratio

$$r = \frac{\mathbb{E}_{x^{\mathcal{M}}} \operatorname{tr} \operatorname{Cov}_\epsilon \left[\mathcal{E}_\theta(\epsilon, 0; x^{\mathcal{M}})\right]}{\operatorname{tr} \operatorname{Cov}_{x^{\mathcal{M}}} \left[\mathbb{E}_\epsilon \, \mathcal{E}_\theta(\epsilon, 0; x^{\mathcal{M}})\right]},$$

estimated without bias from eight queries for each of 1,000 held-out complete observations (the 975 evaluation triplets for image–text–audio). The trained models give $r = 2.6 \times 10^{-5}$ on PolyMNIST-D-Q, $7.9 \times 10^{-5}$ on FFHQ64, and $6.5 \times 10^{-4}$ on image–text–audio, so the full-observation latent is deterministic for practical purposes.

### A.2 DERIVATION OF THE CONDITIONAL JOINT

The coherence-sufficiency condition in Equation 1 makes all modalities conditionally independent given $W$. Hence, for disjoint input and target sets $\mathcal{S}$ and $\mathcal{T}$,

$$P_{\mathrm{data}}(\mathrm{d}x^{\mathcal{T}} \mid W = w, x^{\mathcal{S}}) = \prod_{m \in \mathcal{T}} P_{\mathrm{data}}(\mathrm{d}x^m \mid W = w),$$

and integrating over the latent conditional in Equation 14 gives

$$\begin{aligned} P_{\mathrm{data}}(\mathrm{d}x^{\mathcal{T}} \mid x^{\mathcal{S}}) &= \int P_{\mathrm{data}}(\mathrm{d}x^{\mathcal{T}} \mid W = w, x^{\mathcal{S}})\, Q_{\mathcal{E}}(\mathrm{d}w \mid x^{\mathcal{S}}) \\ &= \int \prod_{m \in \mathcal{T}} P_{\mathrm{data}}(\mathrm{d}x^m \mid W = w)\, Q_{\mathcal{E}}(\mathrm{d}w \mid x^{\mathcal{S}}), \end{aligned} \tag{15}$$

which is Equation 2. One sample of $w$ must therefore be shared by all target decoders. Sampling a separate latent for each target would instead produce $\prod_{m \in \mathcal{T}} P_{\mathrm{data}}(\mathrm{d}x^m \mid x^{\mathcal{S}})$, discarding the dependence that remains after observing the inputs. Taking $\mathcal{S} = \emptyset$ covers unconditional co-generation.

### A.3 PREDICTIVE SUFFICIENCY AND JOINT COHERENCE

MUNI trains the latent of an observed subset toward predictive sufficiency (Yeo et al., 2026),

$$\mathrm{TC}(X^{\mathcal{S}} \mid Z_{\mathcal{S}}) = 0, \qquad I(X^{\mathcal{S}}; X^m \mid Z_{\mathcal{S}}) = 0 \quad \text{for all } m \notin \mathcal{S}. \tag{16}$$

The first condition, on the total correlation (TC), concerns dependence within the observed subset, and the second preserves its information about each missing modality. Neither requires the missing targets to be independent given $Z_{\mathcal{S}}$, as factorized decoding would. A latent can therefore satisfy Equation 16 while factorized decoding misses the joint conditional.

Consider independent uniform bits $U$ and $B$ with $X^1 = U$ and $X^2 = X^3 = (U, B)$, and let $\mathcal{S} = \{1\}$. The latent $Z_{\mathcal{S}} = U$ satisfies Equation 16, and pairing $U$ with an independent uniform bit reproduces

each conditional $P_{\mathrm{data}}(\mathrm{d}x^m \mid x^1)$ exactly. When the two targets are decoded independently in this way, however, they coincide only with probability $1/2$, whereas $X^2$ and $X^3$ always coincide in the data. This holds even though $I(X^1; (X^2, X^3) \mid Z_{\mathcal{S}}) = 0$, so preserving the source information about the joint target does not justify independent decoding; the latent must also resolve the dependence between the targets. The full-observation latent $W = (U, B)$ satisfies Equation 1, and its conditional distribution

$$Q_{\mathcal{E}}(\cdot \mid x^1 = u) = \tfrac{1}{2}\delta_{(u,0)} + \tfrac{1}{2}\delta_{(u,1)}$$

fixes the shared bit once, before either target is decoded, so Equation 15 recovers the correct joint conditional. MUNI imposes coherence sufficiency on the encodings used for prior learning, whereas MUNITE requires it of the full-observation latent whose conditional distribution is inferred for every input subset.

## B Derivation of partial-data self-distillation

We prove Proposition 1 for the unweighted objectives used in training. Throughout, the representation and the teacher are fixed, both objectives use the same distribution over times and conditioning subsets, and each partially observed example is a complete draw from $P_{\mathrm{data}}$ with its unavailable modalities removed independently of the data. The synthetic masking of PolyMNIST-D-Q and FFHQ64 matches this setting, whereas on image–text–audio the available modalities follow the source corpus.

### B.1 The teacher state

Fix $t \in [0, 1)$. Let $W_0 \sim \mathcal{N}(0, I_D)$ be independent of the complete observation and $W_t = (1-t)W_0 + tW$. Since $W - W_t = (1-t)(W - W_0)$, the clean-latent prediction $\mu_{\mathcal{A}}$ determines the velocity

$$\frac{\mu_{\mathcal{A}}(w_t, t; x^{\mathcal{A}}) - w_t}{1-t} = \mathbb{E}[W - W_0 \mid W_t = w_t, X^{\mathcal{A}} = x^{\mathcal{A}}],$$

which is the marginal velocity of the interpolation given $X^{\mathcal{A}}$. By Theorem 1 of Lipman et al. (2023), the ODE driven by this velocity from $W_0$ has the conditional time marginals of $W_t$, so an exact teacher gives

$$(W_t^{\mathcal{A}}, X^{\mathcal{A}}) \overset{d}{=} (W_t, X^{\mathcal{A}}). \tag{17}$$

A partial example thus supplies a correctly distributed evaluation state without revealing its full latent. Only these marginals must match; individual teacher trajectories may differ from the straight interpolation paths.

### B.2 Proof of Proposition 1

*Proof.* Fix $t$ and write $f_\theta = \mathcal{E}_\theta(W_t, t; X^{\mathcal{S}})$ and $\mu_{\mathcal{A}} = \mu_{\mathcal{A}}(W_t, t; X^{\mathcal{A}})$. Since $\mathcal{S} \subsetneq \mathcal{A}$, both are functions of $(W_t, X^{\mathcal{A}})$, and $\mathbb{E}[W - \mu_{\mathcal{A}} \mid W_t, X^{\mathcal{A}}] = 0$. The cross term in

$$\mathbb{E}\|f_\theta - W\|_2^2 = \mathbb{E}\|f_\theta - \mu_{\mathcal{A}}\|_2^2 + \mathbb{E}\|W - \mu_{\mathcal{A}}\|_2^2 + 2\,\mathbb{E}\langle f_\theta - \mu_{\mathcal{A}}, \mu_{\mathcal{A}} - W\rangle$$

therefore vanishes. By Equation 17, the first term on the right is the distillation loss at time $t$ with teacher prediction $\mu_{\mathcal{A}}$. Averaging over $t$ and the subsets gives

$$\widetilde{\mathcal{L}}_{\mathrm{CFM}} - \widetilde{\mathcal{L}}_{\mathrm{dist}} = \mathbb{E}\|W - \mu_{\mathcal{A}}(W_t, t; X^{\mathcal{A}})\|_2^2, \tag{18}$$

which does not depend on $\theta$, so the two objectives have the same gradient with respect to $\theta$. By the tower property, their common minimizer over functions of $(W_t, X^{\mathcal{S}})$ is $\mathbb{E}[\mu_{\mathcal{A}} \mid W_t, X^{\mathcal{S}}] = \mathbb{E}[W \mid W_t, X^{\mathcal{S}}] = \mu_{\mathcal{S}}(W_t, t; X^{\mathcal{S}})$. □

### B.3 Full observations and subset averaging

When $\mathcal{A} = \mathcal{M}$, the teacher target is $\mathcal{E}(x^{\mathcal{M}})$, and the state is the analytic interpolant in Equation 4. Under deterministic encoding, the residual in Equation 18 vanishes, and the target is independent of the Gaussian query, so that query can also serve as $W_0$.

The implementation averages the loss over one subset of each size (Algorithm 1) instead of drawing a single size per update. By linearity of expectation, this stratified average equals the objective under uniformly sampled sizes, so the derivation applies to it unchanged.

# C Architecture and training details

## C.1 Architecture

Images are divided into $8 \times 8$ patches as in ViT (Dosovitskiy et al., 2021), giving 64 tokens per $64 \times 64$ image, and each categorical modality uses a learnable embedding table. For image–text–audio, every modality is given as a precomputed 768-dimensional feature vector, following MUNI and FlowBind, and a linear layer maps it to eight tokens.

The final register states predict $K$ latent tokens of dimension $d$, with total dimension $D = Kd$. We use $K \times d = 4 \times 8$ for PolyMNIST-D-Q, $64 \times 8$ for FFHQ64, and $1 \times 768$ for image–text–audio.

Table 3 lists the widths and depths of the Transformer backbones, which use eight attention heads and FiLM timestep conditioning (Perez et al., 2018) where applicable. Table 7 compares the total trainable parameters, including the generative decoders.

Table 3: Transformer backbones.

| Model | Width | Depth | Time cond. |
|---|---|---|---|
| *PolyMNIST-D-Q* | | | |
| MUNITE | 256 | 7 | FiLM |
| CFM | 256 | 16 | FiLM |
| DFM | 256 | 24 | N/A |
| *FFHQ64* | | | |
| MUNITE | 512 | 11 | FiLM |
| CFM | 512 | 22 | FiLM |
| DFM | 512 | 33 | N/A |
| *Image–text–audio* | | | |
| MUNITE | 768 | 6 | FiLM |
| CFM | 768 | 26 | FiLM |
| DFM | 768 | 37 | N/A |

MUNI and MUNITE use the same generative-decoder families (Yeo et al., 2026). On PolyMNIST-D-Q, the three images share an NCSN++ decoder (Song et al., 2021b) of base width 64, and the labels use two-layer heads of width 256. On FFHQ64, the dense modalities use a spatial NCSN++ decoder of base width 128, and the labels use two-layer heads of width 512. Following common practice for image diffusion models (Esser et al., 2024), both methods train these image decoders with logit-normal timestep sampling. For image–text–audio, we adopt MUNI's decoders and their training details without modification. Image and audio features are decoded by a width-768, depth-10 flow and a width-768, depth-6 feedforward network, respectively. Text is decoded by the Gemma-3-1B (Gemma Team, 2025) feature-to-text decoder fine-tuned by FlowBind (Cha et al., 2026). A feedforward network maps the shared latent to its prefix, and the decoder itself is adapted only through LoRA (Hu et al., 2022).

## C.2 Training procedure

Each minibatch contains examples with the same set of available modalities $\mathcal{A}$, from which all conditioning subsets are drawn. Let $n = |\mathcal{A}|$. For reconstruction, a target $m$ is sampled uniformly from $\mathcal{A}$, and a conditioning subset $\mathcal{S} \subseteq \mathcal{A}$ is drawn by first sampling its size uniformly from $\{1, \dots, n\}$. On image–text–audio, this rule is mixed equally with one that draws $\mathcal{S}$ from $\mathcal{A} \setminus \{m\}$, with size uniform on $\{0, \dots, n-1\}$, to increase the share of cross-modal reconstruction. Contrastive splits draw the size of $\mathcal{S}$ uniformly from $\{1, \dots, n-1\}$. Algorithm 1 summarizes one update. We set $\lambda_{\text{rec}} = 1$, $\lambda_{\text{dist}} = 2$, $\lambda_{\text{con}} = 0.1$, and $\tau_c = 0.1$.

## C.3 Baselines

MUNI aggregates Gaussian modality encoders by a product of experts without a prior expert and samples this posterior directly for conditional generation (Yeo et al., 2026).

CFM and DFM use the Transformer blocks, width, and number of attention heads of MUNITE (Table 3), with more blocks so that their trainable parameters approximately match those of MUNITE, including its generative decoders (Table 7). DFM has no timestep conditioning and therefore needs

**Algorithm 1** MUNITE training step

**Require:** minibatch $x^{\mathcal{A}}$, teacher Euler steps $N$, learning rate $\eta$
Sample $m \in \mathcal{A}$, $\mathcal{S} \subseteq \mathcal{A}$, and $\epsilon \sim \mathcal{N}(0, I_D)$
$\mathcal{L}_{\text{rec}} \leftarrow \ell_m\big(\mathcal{D}_{\psi_m}(\mathcal{E}_\theta(\epsilon, 0; x^{\mathcal{S}})), x^m\big)$ ▷ target-detached
Sample $w_0 \sim \mathcal{N}(0, I_D)$ and $t \sim \mathcal{U}[0, 1)$
**if** $\mathcal{A} = \mathcal{M}$ **then**
  $\bar{w} \leftarrow \mathcal{E}_\theta(w_0, 0; x^{\mathcal{M}})$; $w_t \leftarrow (1-t)w_0 + t\bar{w}$; $\mathcal{K} \leftarrow \{0, \dots, M\}$
**else**
  $w_t \leftarrow w_0$
  **for** $j = 0, \dots, N-1$ **do**
    $w_t \leftarrow w_t + \frac{t}{N}\, v_\theta\big(w_t, \frac{jt}{N}; x^{\mathcal{A}}\big)$
  **end for**
  $\bar{w} \leftarrow \mathcal{E}_\theta(w_t, t; x^{\mathcal{A}})$; $\mathcal{K} \leftarrow \{0, \dots, |\mathcal{A}| - 1\}$
**end if**
$\widetilde{\mathcal{L}}_{\text{dist}} \leftarrow 0$
**for** $k \in \mathcal{K}$ **do**
  Sample $\mathcal{S}_k \subseteq \mathcal{A}$ with $|\mathcal{S}_k| = k$
  $\widetilde{\mathcal{L}}_{\text{dist}} \leftarrow \widetilde{\mathcal{L}}_{\text{dist}} + \frac{1}{|\mathcal{K}|}\big\|\mathcal{E}_\theta(\text{sg}(w_t), t; x^{\mathcal{S}_k}) - \text{sg}(\bar{w})\big\|_2^2$
**end for**
$\mathcal{L}_{\text{con}} \leftarrow 0$
**if** $|\mathcal{A}| > 1$ **then**
  Sample nonempty $\mathcal{S} \subsetneq \mathcal{A}$; $\mathcal{C} \leftarrow \mathcal{A} \setminus \mathcal{S}$
  $\mathcal{L}_{\text{con}} \leftarrow \frac{1}{2}(\mathcal{L}_{\mathcal{S}\to\mathcal{C}} + \mathcal{L}_{\mathcal{C}\to\mathcal{S}})$
**end if**
$\mathcal{L} \leftarrow \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{dist}}\widetilde{\mathcal{L}}_{\text{dist}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}$
$(\theta, \psi) \leftarrow (\theta, \psi) - \eta\, \nabla_{\theta,\psi}\mathcal{L}$

more blocks than CFM. CFM represents images and features as continuous patches or vectors and labels as one-hot vectors, and learns a straight flow from Gaussian noise (Lipman et al., 2023). DFM represents each label by a single token and each image or feature by frozen VQ-VAE codes, with 1,024-entry codebooks, 64 codes per image, and 32 codes per image–text–audio feature, and is trained with absorbing-mask corruption and clean-token prediction (Gat et al., 2024). During training, the numbers of source and target modalities, $s \geq 0$ and $r \geq 1$ with $s + r \leq n$, are sampled uniformly, and the available modalities are randomly assigned to the source, target, and unused sets.

# D Benchmark construction and comparison protocol

## D.1 Datasets

Table 4: Training observations grouped by available modality subset. For PolyMNIST-D-Q and FFHQ64, categorical labels are observed in all groups. $T$, $I$, and $A$ denote text, image, and audio, respectively.

| **PolyMNIST** $(+\, d, q)$ | | | **FFHQ64** $(+$ Age, Gender$)$ | | | **Image–Text–Audio** | | |
|---|---|---|---|---|---|---|---|---|
| Observed subset | Rows | % | Observed subset | Rows | % | Observed subset | Rows | % |
| $m_0, m_1, m_2$ | 55,000 | 25.0 | RGB, Seg., Norm. | 16,250 | 25.0 | $T, I$ | 272,703 | 37.8 |
| $m_0, m_1$ | 18,333 | 8.3 | RGB, Seg. | 5,417 | 8.3 | $I, A$ | 183,449 | 25.5 |
| $m_0, m_2$ | 18,334 | 8.3 | RGB, Norm. | 5,416 | 8.3 | $T, A$ | 91,028 | 12.6 |
| $m_1, m_2$ | 18,333 | 8.3 | Seg., Norm. | 5,417 | 8.3 | $T, I, A$ | 173,576 | 24.1 |
| $m_0$ | 36,667 | 16.7 | RGB | 10,833 | 16.7 | | | |
| $m_1$ | 36,666 | 16.7 | Seg. | 10,834 | 16.7 | | | |
| $m_2$ | 36,667 | 16.7 | Norm. | 10,833 | 16.7 | | | |
| Total | 220,000 | 100 | Total | 65,000 | 100 | Total | 720,756 | 100 |

PolyMNIST-D-Q, released by Yeo et al. (2026) as PolyMNIST-Quadrant-Labels and built on PolyMNIST-Quadrant (Zhang et al., 2026), contains three $64 \times 64$ RGB views with their digit (D) and quadrant (Q) labels. Each view composites a resized MNIST digit (LeCun et al., 1998) onto a

Table 5: Image–text–audio training data.

| Corpus | Modalities | Rows |
|---|---|---|
| LAION-COCO aesthetic subset | $T, I$ | 241,873 |
| VGGSound | $I, A$ | 183,449 |
| AudioCaps | $T, A$ | 91,028 |
| Flickr30k | $T, I$ | 30,830 |
| E-MM1 | $T, I, A$ | 173,576 |
| Total | | 720,756 |

Table 6: Image–text–audio evaluation sets.

| Routes | Population | Count |
|---|---|---|
| $T \leftrightarrow I$ | COCO | 30,000 |
| $T \leftrightarrow A$ | AudioCaps | 975 |
| $I \leftrightarrow A$ | VGGSound | 15,446 |
| $2 \rightarrow 1, 1 \rightarrow 2$ | AudioCaps-based triplets | 975 |
| Unconditional | Joint samples | 1,000 |

Table 7: Trainable parameters in millions, including learned encoders, priors, and decoders. Frozen foundation models and the frozen VQ-VAEs of DFM, which have 13.27M parameters on PolyMNIST-D-Q and FFHQ64, are excluded.

| Task | FlowBind | MUNI | CFM | DFM | MUNITE |
|---|---|---|---|---|---|
| PolyMNIST-D-Q | — | 32.09 | 20.09 | 20.57 | 19.33 |
| FFHQ64 | — | 130.99 | 105.11 | 107.27 | 106.64 |
| Image–text–audio | 567.97 | 328.90 | 281.21 | 281.16 | 278.72 |

modality-specific background by color inversion. The first 220,000 of the 230,000 tuples are used for training, and the last 10,000 are held out for evaluation. Because digits are drawn from both MNIST splits, the same digit image may appear in training and evaluation tuples.

FFHQ64 is constructed from FFHQ (Karras et al., 2019) and contains RGB, 19-class segmentation, surface normals, age in ten bins (0–2, 3–6, 7–9, 10–14, 15–19, 20–29, 30–39, 40–49, 50–69, and 70+ years), and binary gender. Age, gender, and segmentation labels are taken from FFHQ-Aging (Or-El et al., 2020), and surface normals are pseudo-labels from Marigold E2E-FT (Garcia et al., 2025). The first 65,000 aligned faces are used for training, and the last 5,000 are held out for evaluation.

The image–text–audio corpus (Table 5) contains 720,756 paired or triplet observations from the LAION-COCO aesthetic subset (Schuhmann et al., 2022), VGGSound (Chen et al., 2020), AudioCaps (Kim et al., 2019), Flickr30k (Young et al., 2014), and E-MM1 (Broadbent et al., 2025). The observed modalities of each example follow the original pairing of its corpus rather than synthetic masking. All models operate on fixed, normalized 768-dimensional features from EmbeddingGemma-300M (Schechter Vera et al., 2025) for text, CLIP ViT-L/14 (Radford et al., 2021) for images, and LAION CLAP HTSAT-base (Wu et al., 2023) for audio, whose features are mapped to 768 dimensions by the audio autoencoder released with FlowBind (Cha et al., 2026). For the three-modality observations, we use the caption–image–audio triplets of E-MM1 rated Good Match in its human-annotated 1M split, excluding any triplet that overlaps our evaluation sets or the other training corpora.

### D.2 Model sizes and training budgets

In Table 7, CFM and DFM approximately match the trainable parameters of MUNITE, whereas MUNI and FlowBind keep the larger architectures of their released implementations.

All methods are trained with Adam (Kingma & Ba, 2015) under the training budgets in Table 8. MUNI, MUNITE, CFM, and DFM are evaluated with EMA weights, with decay 0.999 on PolyMNIST-D-Q and 0.9995 on FFHQ64 and image–text–audio, whereas FlowBind follows its official pipeline without EMA.

### D.3 Inference budgets

In Table 9, CFM is given as many network evaluations as the latent flow and decoder of MUNITE combined, counting two forward passes per step with classifier-free guidance (Ho & Salimans, 2022). DFM samples one token at a time in random order, which avoids the factorization error of predicting several tokens in parallel. FlowBind (Cha et al., 2026) infers a latent from each source and averages them when several sources are given; it has no unconditional sampler.

Table 8: Training budgets. Batch sizes are global, and all methods use a constant learning rate. FlowBind's epoch count assumes 702 updates per epoch.

| Method | Epochs | Updates | Batch | LR |
|---|---|---|---|---|
| *PolyMNIST-D-Q* | | | | |
| MUNI, MUNITE | 100 | 85,600 | 256 | $1.5 \times 10^{-4}$ |
| CFM, DFM | 200 | 171,200 | 256 | $1.5 \times 10^{-4}$ |
| *FFHQ64* | | | | |
| MUNI, MUNITE | 500 | 252,000 | 128 | $1.5 \times 10^{-4}$ |
| CFM, DFM | 1,000 | 504,000 | 128 | $1.5 \times 10^{-4}$ |
| *Image–text–audio* | | | | |
| MUNITE | 100 | 70,200 | 1,024 | $4.0 \times 10^{-4}$ |
| MUNI, CFM, DFM | 200 | 140,400 | 1,024 | $4.0 \times 10^{-4}$ |
| FlowBind | $\approx 285$ | 200,000 | 1,024 | $1.0 \times 10^{-4}$ |

Table 9: Inference and decoding methods. Euler step counts equal the number of function evaluations, CFG denotes classifier-free guidance, and AR denotes autoregressive sampling. Rendering of image–text–audio features is not included.

| Method | Conditional | Unconditional | Decoding |
|---|---|---|---|
| *PolyMNIST-D-Q* | | | |
| MUNITE | Euler 50 | Euler 50 | Image Euler 50, CFG 1.5 |
| MUNI | Posterior sampling | Euler 50 | Image Euler 50, CFG 1.5 |
| CFM | Euler 150 | Euler 150 | — |
| DFM | Random-order AR | Random-order AR | VQ decoder |
| *FFHQ64* | | | |
| MUNITE | Euler 50 | Euler 50 | Image Euler 50, CFG 1.5 |
| MUNI | Posterior sampling | Euler 50 | Image Euler 50, CFG 1.5 |
| CFM | Euler 150 | Euler 150 | — |
| DFM | Random-order AR | Random-order AR | VQ decoder |
| *Image–text–audio* | | | |
| MUNITE | Euler 25 | Euler 25 | Image-feature Euler 25, CFG 4 |
| MUNI | Posterior sampling | Euler 25 | Image-feature Euler 25, CFG 4 |
| CFM | Euler 75 | Euler 75 | — |
| DFM | Random-order AR | Random-order AR | VQ decoder |
| FlowBind | Euler 10 per source | — | Euler 10 per target |

For the models trained on our benchmark, generated image and audio features are rendered by Stable unCLIP small (Stability AI, 2023), an unCLIP-style decoder (Ramesh et al., 2022), and by AudioLDM-m-full (Liu et al., 2023), respectively, with 20 denoising steps as in the benchmark. Because CFM and DFM have no modality-specific decoders, they generate normalized text features as FlowBind does, and all three decode them with the frozen FlowBind text decoder, on which the text decoders of MUNI and MUNITE are built (Appendix C.1).

### D.4 Verifiers

Following MUNI (Yeo et al., 2026), accuracy and coherence on PolyMNIST-D-Q and FFHQ64 are measured with verifiers that predict the labels and dense maps of generated images. Each verifier fine-tunes a pretrained SegFormer backbone (Xie et al., 2021), MiT-B0 for PolyMNIST-D-Q and MiT-B5 for FFHQ64, on the classification, segmentation, and normal-estimation tasks of its benchmark. The verifiers are trained on all 230,000 PolyMNIST-D-Q tuples and all 70,000 FFHQ faces, including the held-out ones, and are kept frozen when scoring all methods.

# E EVALUATION METRICS AND ADDITIONAL RESULTS

## E.1 POLYMNIST-D-Q

PolyMNIST-D-Q is evaluated on the 10,000 held-out tuples and on 10,000 unconditional samples. FD uses verifier features instead of Inception features and is averaged over the three views.

For label-to-image generation, accuracy compares the verifier's prediction for each generated view with the input label, and the $(d, q) \to m_i$ route counts a view as correct only if both labels match. Unconditional coherence is the fraction of samples in which the verifier assigns the generated label to all three views. One-to-many coherence is the agreement between pairs of generated views on the attribute that the input leaves out, the quadrant for $d \to (m_0, m_1, m_2)$ and the digit for $q \to (m_0, m_1, m_2)$, averaged over the three pairs and all samples.

Pairwise agreement alone does not show that the missing attribute is sampled correctly, since a model that always generates the same quadrant or digit for a given input would also agree fully. Table 10 therefore also compares the missing label of each generated view with its data conditional, which is uniform because every digit–quadrant pair occurs equally often. MUNITE stays close to the finite-sample reference for both attributes, so its agreement does not come from a collapsed prediction. For the missing quadrant, MUNI is equally close to uniform, yet its views agree only at chance level: the quadrant of each view is correctly distributed, but the three views choose their quadrants independently. When $m_1$ and $m_2$ are decoded from separately drawn latents (Indep.), MUNITE's agreement also falls to chance, while the accuracy of each view on the input label remains 0.995 and 1.000 for digit and quadrant inputs. The agreement among views is therefore carried by the shared latent sample.

## E.2 FFHQ64

FFHQ64 is evaluated on the 5,000 held-out tuples and on 5,000 unconditional samples, and RGB fidelity is measured by FID (Heusel et al., 2017) with InceptionV3 (Szegedy et al., 2016) features. Age and gender accuracy compare the verifier's predictions for a generated face with the input labels, requiring both to match on the two-label route, and unconditional age and gender coherence compare them with the generated labels. Generated dense maps are scored against the verifier's predictions for the generated face. RGB–segmentation coherence is the pixel accuracy over the 19 classes, and normal error is one minus the mean per-pixel cosine similarity. For $(\mathrm{Age}, \mathrm{Gender}) \to (\mathrm{RGB}, \mathrm{Seg.}, \mathrm{Normals})$, the inputs provide no reference maps, so these scores compare the generated targets with one another.

## E.3 IMAGE–TEXT–AUDIO

The evaluation follows the routes and evaluation sets of FlowBind and MUNI (Cha et al., 2026; Yeo et al., 2026), listed in Table 6. The 975 joint-evaluation triplets contain AudioCaps (Kim et al., 2019) captions and real audio with FLUX.1-generated images (Black Forest Labs, 2024), so the three modalities are not co-recorded. CIDEr is computed with one reference caption per COCO (Lin et al., 2014) or AudioCaps item.

Table 10: Missing attribute in PolyMNIST-D-Q one-to-many generation. TV is the total variation distance between the missing-label distribution of a generated view and the uniform data conditional, averaged over input labels and views; Ref. is its expected value for the same number of samples drawn from that conditional. Indep. is MUNITE with $m_1$ and $m_2$ decoded from separately drawn latents, keeping $m_0$ and all decoder noise fixed. Chance is the coherence of independent uniform views.

| | TV ↓ | | | Coherence ↑ | | | |
|---|---|---|---|---|---|---|---|
| Route | MUNI | MUNITE | Ref. | MUNI | MUNITE | Indep. | Chance |
| $d \to (m_0, m_1, m_2)$ | 0.029 | 0.029 | 0.022 | 0.2528 | 1.0000 | 0.2500 | 0.25 |
| $q \to (m_0, m_1, m_2)$ | 0.109 | 0.034 | 0.024 | 0.1074 | 0.9904 | 0.1014 | 0.10 |

Fidelity is measured by FID (Heusel et al., 2017) with InceptionV3 (Szegedy et al., 2016) features, FAD (Kilgour et al., 2019) with VGGish (Hershey et al., 2017) embeddings, and CIDEr (Vedantam et al., 2015). Source–target alignment is measured by CLIP ViT-B/32 (Radford et al., 2021) for text–image and LAION larger-CLAP (Wu et al., 2023) for text–audio, both different from the encoders used to extract the training features. CLIP and CLAP similarities and AIS are reported on a scale of 100.

AIS measures how highly a generated sample ranks its paired counterpart within a reference pool, using Wav2CLIP (Wu et al., 2022) audio and CLIP ViT-B/32 (Radford et al., 2021) image embeddings as in FlowBind and MUNI (Cha et al., 2026; Yeo et al., 2026). For $N$ generated samples with normalized embeddings $g_i$, the normalized embeddings $c_i$ of their counterparts, and a pool $R$ of counterpart embeddings spanning the evaluation set,

$$\mathrm{AIS} = \frac{100}{N|R|} \sum_{i=1}^{N} \sum_{r \in R} \mathbf{1}[\langle g_i, r \rangle < \langle g_i, c_i \rangle]. \tag{19}$$

For text-conditioned joint image–audio generation, each generated image is paired with its generated audio, and the pool contains the generated audios.

**External pretrained models.** CoDi (Tang et al., 2023) and OmniFlow (Li et al., 2025) are evaluated as released pretrained models. We measure their one-to-many coherence on the same 975 triplets and take their other scores from MUNI (Yeo et al., 2026), where they are reported to two decimal places. For text-conditioned joint image–audio generation, CoDi uses its default DDIM sampler (Song et al., 2021a) with 50 steps and guidance 7.5, and OmniFlow uses 28 steps with guidance 5.

### E.3.1 Single-target generation

Tables 11–13 report the one-to-one and many-to-one scores summarized in Section 4.4. MUNITE has the best score on three of the six one-to-one fidelity measures (Table 11). Relative to MUNI, it has higher text–audio and lower text–image alignment in both one-to-one and many-to-one generation (Tables 12 and 13). Since its single outputs are not uniformly better, the coherence gains in Table 2 are not explained by single-output quality.

Table 11: Image–text–audio one-to-one fidelity on the evaluation sets of Table 6, grouped by target modality. CoDi and OmniFlow scores are taken from Yeo et al. (2026, Table 8). Bold marks the best score in each column.

| | Image: FID ↓ | | Audio: FAD ↓ | | Text: CIDEr ↑ | |
|---|---|---|---|---|---|---|
| Method | $T \to I$ | $A \to I$ | $T \to A$ | $I \to A$ | $I \to T$ | $A \to T$ |
| CoDi | 24.97 | 53.60 | 8.67 | 13.92 | 11.97 | 7.70 |
| OmniFlow | 20.52 | 97.56 | 4.08 | 4.87 | 26.60 | 31.42 |
| MUNI | 15.615 | 22.786 | 3.507 | 2.246 | **44.400** | 51.794 |
| CFM | **13.455** | 28.971 | 3.564 | 2.105 | 13.085 | 19.729 |
| DFM | 14.400 | 28.374 | 3.670 | **1.991** | 11.263 | 20.909 |
| FlowBind | 14.826 | 24.270 | 3.477 | 2.166 | 17.211 | 30.905 |
| MUNITE | 15.939 | **21.511** | **3.059** | 2.135 | 37.409 | **57.592** |

### E.3.2 Component ablations

Table 14 ablates the training components of MUNITE on image–text–audio. *No target detach* lets reconstruction gradients flow through the target's key/value paths, *no contrastive* removes $\mathcal{L}_{\mathrm{con}}$, and *full-only distillation* applies the distillation loss only to fully observed examples, so partially observed examples no longer supervise the latent conditionals.

Every variant keeps one-to-many coherence within two points of the full model and above MUNI on all three routes (Table 2), so the one-to-many advantage does not depend on any single auxiliary component. Unconditional coherence depends more on these components. Target detaching matters most: without it, CLAP and AIS coherence drop by 5.6 and 9.6 points, and all three unconditional

Table 12: Image–text–audio one-to-one alignment. CoDi and OmniFlow scores are taken from Yeo et al. (2026, Table 9). Higher is better, and bold marks the best score in each column.

| | Text–image: CLIP | | Text–audio: CLAP | | Image–audio: AIS | |
|---|---|---|---|---|---|---|
| Method | $T \to I$ | $I \to T$ | $T \to A$ | $A \to T$ | $I \to A$ | $A \to I$ |
| CoDi | 29.70 | 26.05 | 16.72 | 32.72 | 59.45 | 84.47 |
| OmniFlow | **30.76** | 27.48 | 30.07 | **44.74** | 76.52 | 64.22 |
| MUNI | 29.718 | **28.373** | 30.596 | 40.102 | 90.390 | **94.773** |
| CFM | 27.710 | 24.918 | 25.127 | 32.094 | 82.888 | 80.842 |
| DFM | 27.321 | 24.353 | 24.883 | 32.262 | 79.903 | 79.933 |
| FlowBind | 29.202 | 26.004 | 28.011 | 35.393 | 87.621 | 92.445 |
| MUNITE | 29.007 | 27.526 | **30.695** | 40.335 | **90.655** | 94.254 |

Table 13: Image–text–audio many-to-one alignment on 975 triplets, with each generated target scored against both source modalities. CoDi and OmniFlow scores are taken from Yeo et al. (2026, Table 3). Higher is better, and bold marks the best score in each column.

| | $(I, A) \to T$ | | $(T, A) \to I$ | | $(T, I) \to A$ | |
|---|---|---|---|---|---|---|
| Method | CLIP | CLAP | CLIP | AIS | CLAP | AIS |
| CoDi | 24.05 | 33.72 | 24.98 | 85.52 | 11.06 | 65.31 |
| OmniFlow | 24.73 | 36.26 | 26.41 | 81.51 | 13.50 | 63.55 |
| MUNI | **27.922** | 36.695 | **27.056** | 85.168 | 30.524 | 82.621 |
| CFM | 25.221 | 31.889 | 25.712 | 80.305 | 24.907 | 76.098 |
| DFM | 24.855 | 33.259 | 25.398 | 80.347 | 25.141 | 75.846 |
| FlowBind | 25.541 | 31.624 | 25.841 | **88.089** | 28.748 | 81.878 |
| MUNITE | 27.289 | **38.969** | 26.721 | 86.745 | **30.653** | **82.677** |

scores fall below those of MUNI. Restricting distillation to fully observed examples lowers CLAP and AIS coherence by 2.8 and 4.6 points, since partially observed examples then no longer shape the latent distribution sampled in unconditional generation.

For single-target generation, removing target detaching or the contrastive loss degrades fidelity on all six one-to-one routes, most visibly image-to-audio FAD and audio-to-text CIDEr without the contrastive loss. Full-only distillation gives mixed results, slightly improving fidelity on the four routes that generate or condition on images and degrading text-to-audio FAD and audio-to-text CIDEr, so partial-data distillation mainly benefits coherence.

Table 14: Component ablations on image–text–audio. Control is the full MUNITE model of Tables 2 and 11–13. Each variant removes one component and otherwise keeps the control's architecture, optimization, and evaluation protocol. The $1 \to N$ and unconditional rows report coherence among jointly generated modalities. Best results in each row are in bold.

| Route | Input → output | Metric | Control | No target detach | No contrastive | Full-only distillation |
|---|---|---|---|---|---|---|
| $1 \to 1$ | $T \to I$ | FID ↓ | 15.939 | 16.700 | 17.308 | **15.230** |
| | | CLIP ↑ | 29.007 | 28.543 | 27.750 | **29.593** |
| | $I \to T$ | CIDEr ↑ | 37.409 | 35.414 | 35.741 | **39.667** |
| | | CLIP ↑ | 27.526 | 27.261 | 26.919 | **27.882** |
| | $T \to A$ | FAD ↓ | **3.059** | 3.145 | 3.321 | 3.234 |
| | | CLAP ↑ | 30.695 | 30.458 | **31.461** | 30.540 |
| | $A \to T$ | CIDEr ↑ | **57.592** | 53.462 | 51.927 | 53.302 |
| | | CLAP ↑ | 40.335 | 40.670 | 39.588 | **40.678** |
| | $I \to A$ | FAD ↓ | 2.135 | 2.146 | 2.622 | **2.094** |
| | | AIS ↑ | 90.655 | 90.768 | **91.952** | 90.157 |
| | $A \to I$ | FID ↓ | 21.511 | 21.923 | 22.466 | **20.823** |
| | | AIS ↑ | 94.254 | 93.932 | 93.565 | **94.260** |
| $N \to 1$ | $(I, A) \to T$ | CLIP ↑ | 27.289 | 27.191 | 27.148 | **27.441** |
| | | CLAP ↑ | **38.969** | 38.400 | 36.322 | 38.517 |
| | $(T, A) \to I$ | CLIP ↑ | 26.721 | 26.670 | 26.581 | **26.799** |
| | | AIS ↑ | **86.745** | 85.898 | 85.624 | 85.127 |
| | $(T, I) \to A$ | CLAP ↑ | 30.653 | 30.511 | **31.420** | 30.804 |
| | | AIS ↑ | 82.677 | 82.461 | **83.407** | 83.055 |
| $1 \to N$ | $T \to (I, A)$ | AIS ↑ | 83.336 | **84.174** | 83.771 | 82.487 |
| | $I \to (T, A)$ | CLAP ↑ | 34.762 | **34.844** | 32.852 | 34.392 |
| | $A \to (T, I)$ | CLIP ↑ | **26.334** | 26.261 | 26.024 | 26.287 |
| Uncond. | All three | CLIP $(T, I)$ ↑ | 27.398 | 26.713 | 26.639 | **27.539** |
| | | CLAP $(T, A)$ ↑ | **30.220** | 24.657 | 29.370 | 27.379 |
| | | AIS $(I, A)$ ↑ | **87.045** | 77.451 | 85.760 | 82.429 |

# F QUALITATIVE RESULTS

Qualitative examples complement the quantitative results of Section 4, particularly for joint generation, where they show whether the generated outputs agree on content that the input leaves unspecified. The examples are selected to illustrate each route, and the outputs of routes whose only output is audio are provided in the supplementary material.

## F.1 POLYMNIST-D-Q

In single-target generation (Figure 4), MUNI and MUNITE render the conditioned labels in every example, whereas CFM often produces illegible digits and DFM frequently omits the digit or renders a different one, in line with their accuracies in Table 1.

In unconditional co-generation (Figure 5), the three views and two labels of a sample must agree with one another. All four samples of MUNITE agree, whereas MUNI's views sometimes disagree, showing different digits or placing one digit in different quadrants.

In the one-to-many routes (Figure 6), the views must also agree on the attribute that the input leaves unspecified. MUNITE places a given digit in the same quadrant in all three views and draws the same digit for a given quadrant. MUNI's views are each consistent with the input but draw the missing attribute separately, so the digit 1 appears in three different quadrants and the lower-left quadrant holds the digits 0, 5, and 4. This is the gap between target-wise sufficiency and joint coherence discussed in Section 2.7.

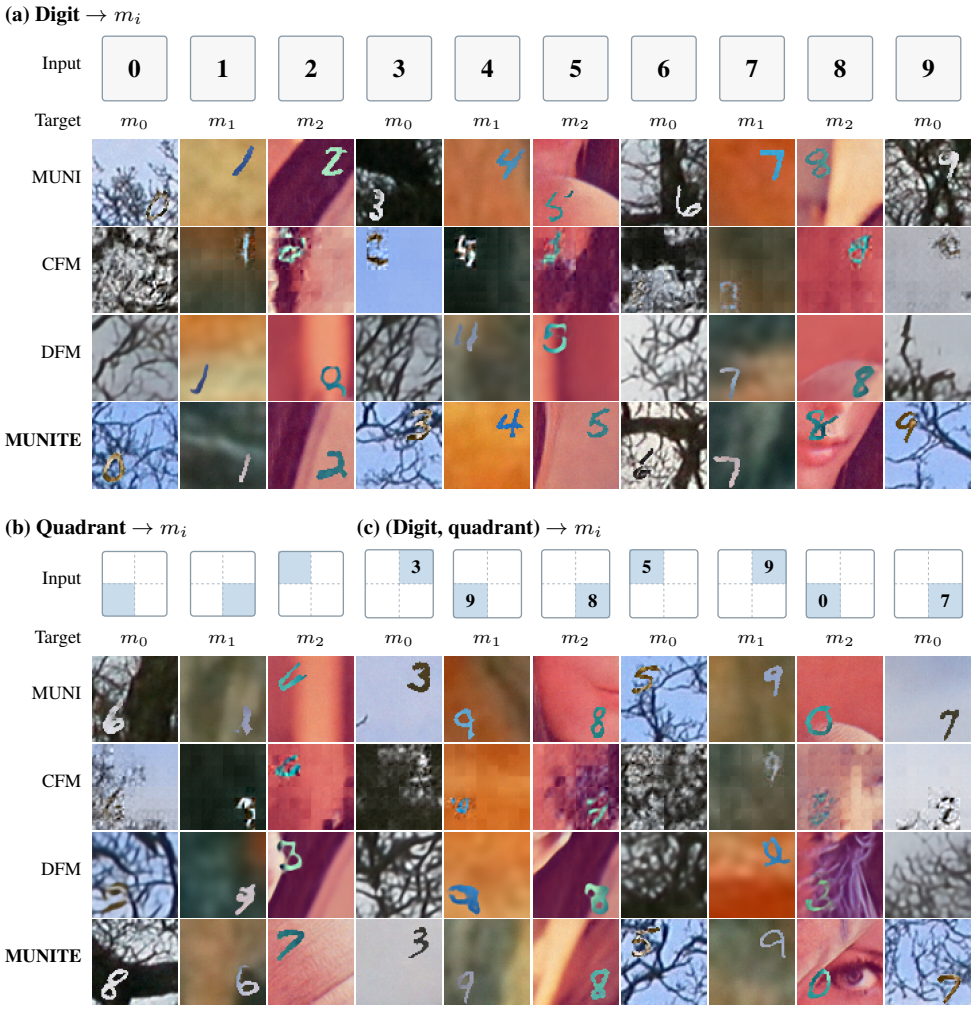



Figure 4: PolyMNIST-D-Q single-target generation. Icons show the observed labels: a digit, the conditioned quadrant shaded in a $2 \times 2$ grid, or the digit placed in that quadrant. Each column generates the view listed under its input.

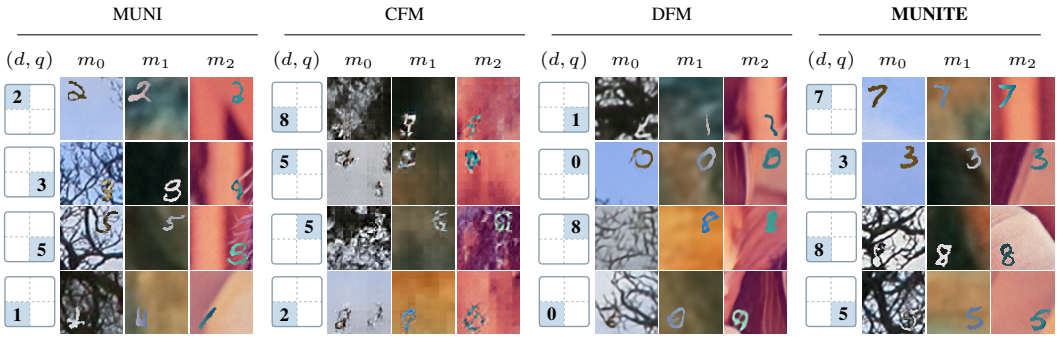



Figure 5: Unconditional PolyMNIST-D-Q co-generation. Each group is one joint sample of the three views and both labels, and its icon shows the generated digit in the generated quadrant.

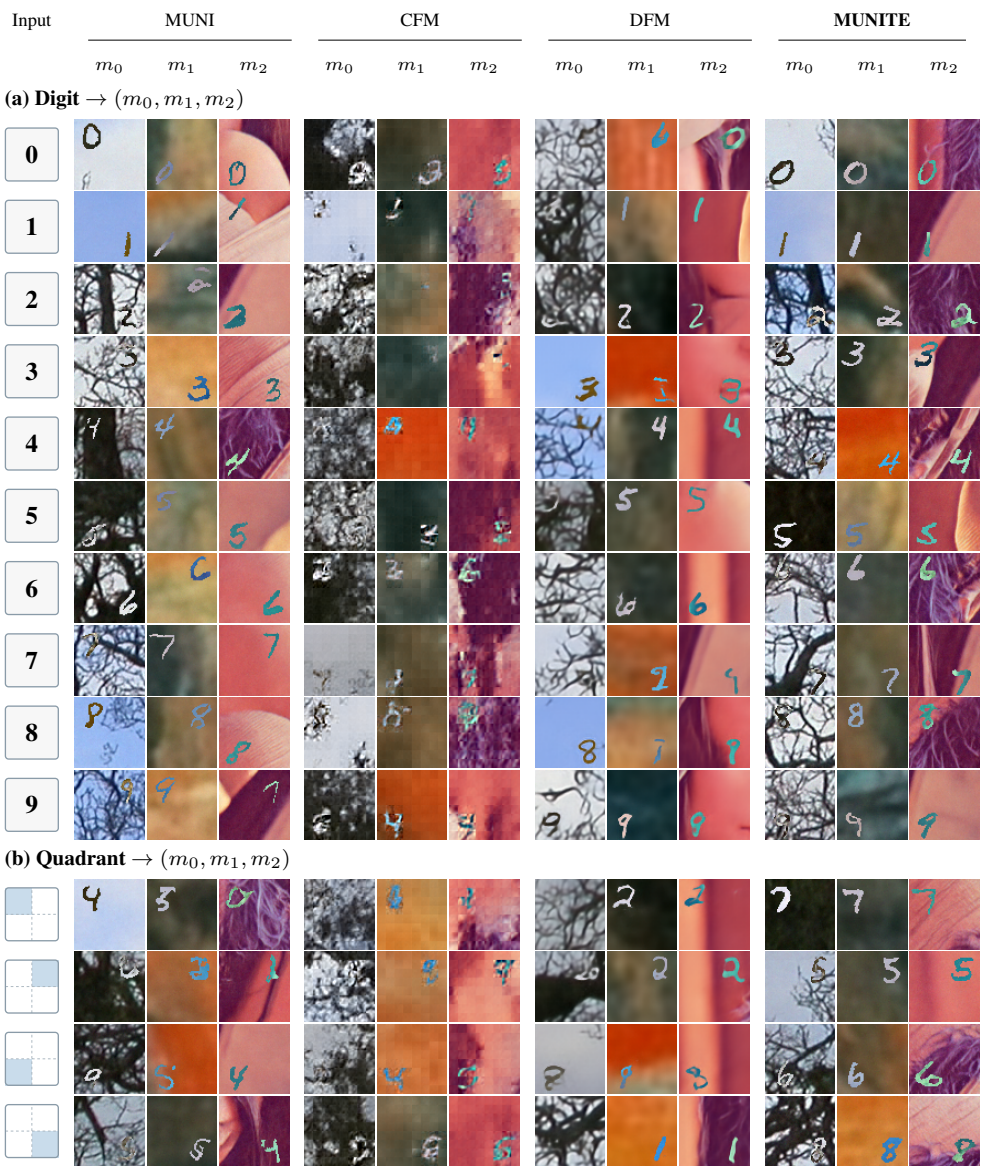



Figure 6: PolyMNIST-D-Q one-to-many generation. Each row generates all three views from a digit (a) or a quadrant (b), so the views must agree on the unobserved attribute.

### F.2 FFHQ64

In RGB generation from age and gender (Figure 7), MUNITE's faces follow the conditioned age from young children to older adults and match the conditioned gender. MUNI's faces also follow the attributes but are occasionally distorted, CFM's show strong artifacts, and DFM's are noticeably blurred.

In unconditional co-generation of all five modalities (Figure 8), MUNITE's generated labels agree with the apparent age and gender of its faces, and its segmentation and normals align with the RGB image. The man in the 50–69 sample, for instance, wears glasses in the RGB image, and the segmentation marks them as well.

In the many-to-many route (Figure 9), the segmentation and normals must describe the generated face. MUNITE's maps follow the head pose, hairline, and facial layout of its RGB image in every example. MUNI's maps sometimes describe a different face, as for the male 40–49 input, whose segmentation omits the glasses in the RGB image.

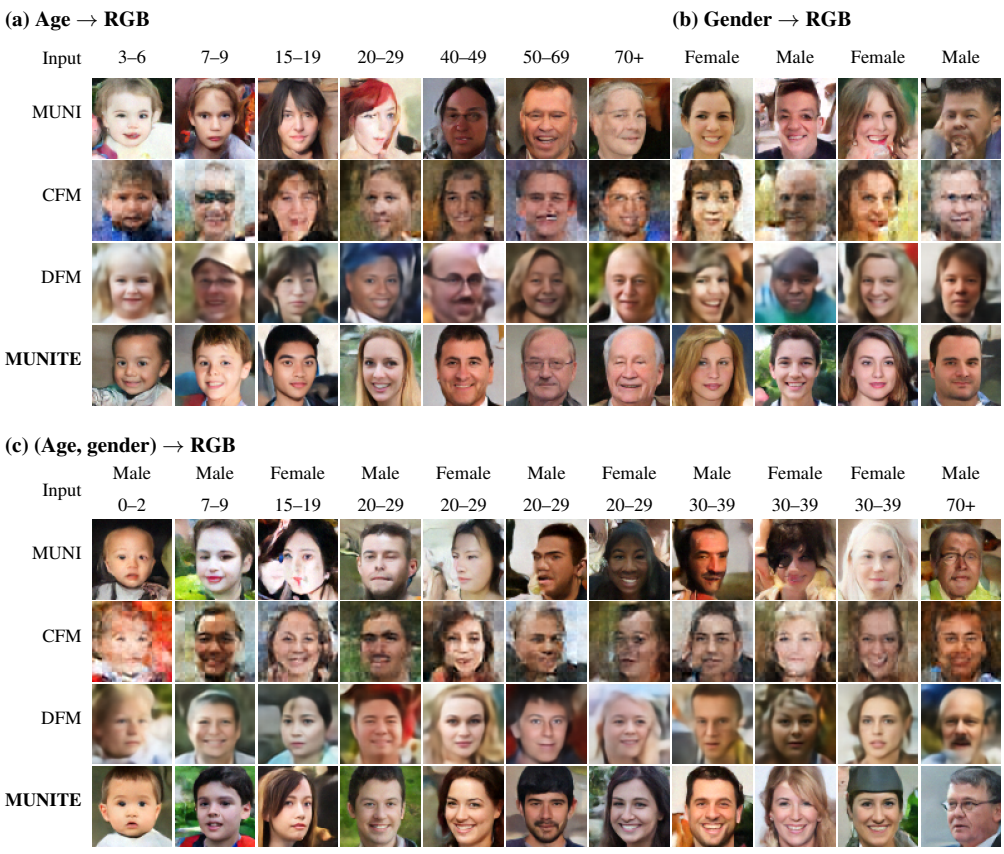



Figure 7: FFHQ64 RGB generation from age (a), gender (b), or both attributes (c).

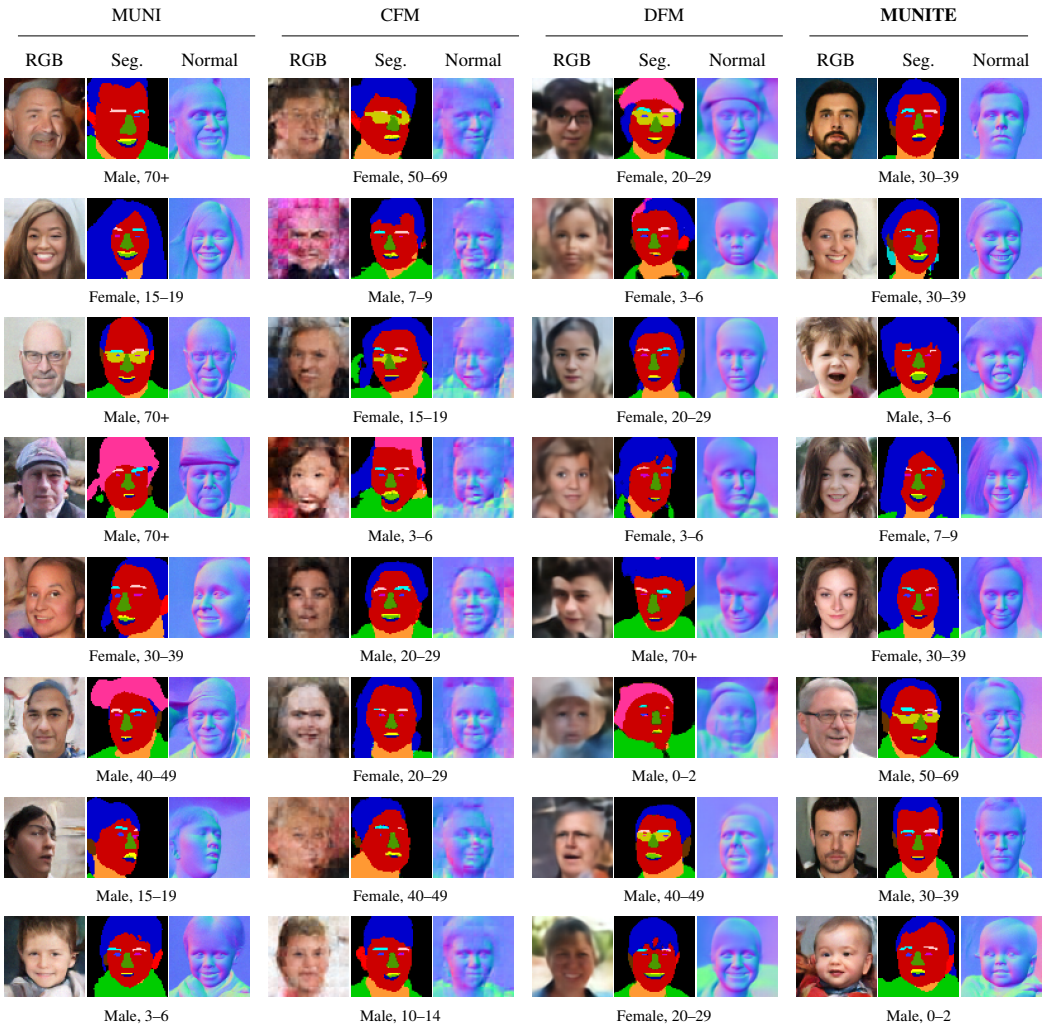



Figure 8: Unconditional FFHQ64 co-generation of all five modalities. The generated gender and age labels of each joint sample appear below its maps.

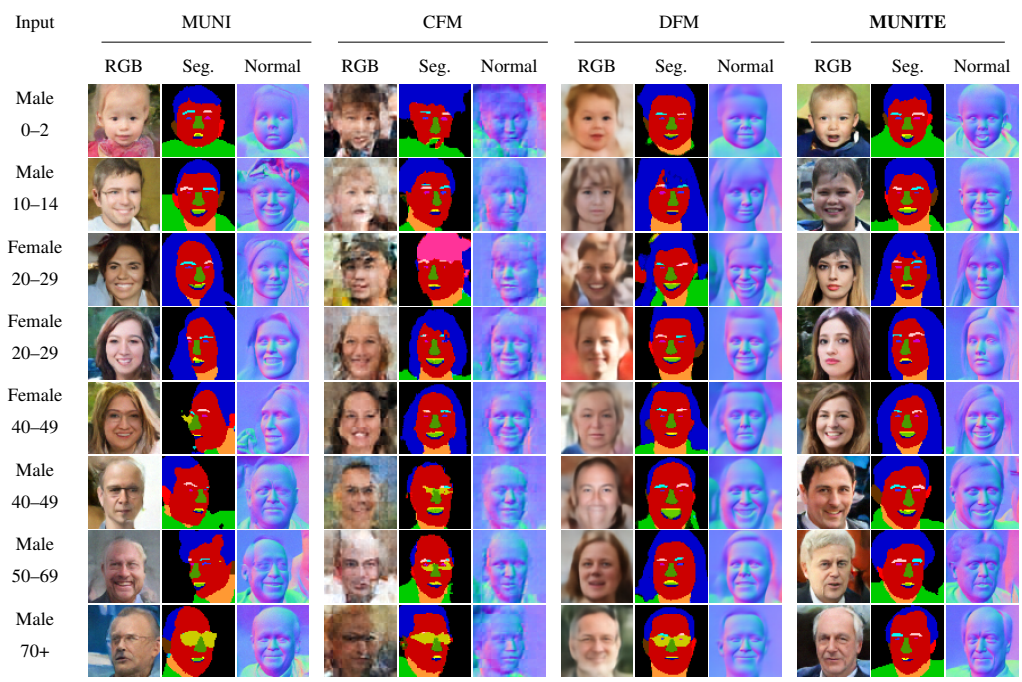



Figure 9: FFHQ64 many-to-many generation of RGB, segmentation, and surface normals from age and gender. The three maps of each group come from one joint sample.

### F.3 IMAGE–TEXT–AUDIO

Audio clips are provided in the supplementary material under the IDs printed below their waveforms. The bracketed descriptions are only a reading aid and summarize the top AudioSet labels predicted by PaSST (Koutini et al., 2022).

In one-to-one and many-to-one generation (Figures 10 and 11), MUNITE's outputs follow the specific content of their inputs. Its image for the roaring lions in A02 shows a lion where FlowBind shows a tiger, and its caption for the stop-sign image reads the sign. With two inputs, its captions draw on both. For the faucet image paired with running water and speech (A10), MUNITE mentions both the water and the speaker, whereas MUNI mentions only the water.

For joint generation, we show samples of MUNITE (Figures 12 and 13). An unconditional sample has no input, so whatever its outputs share comes from the latent sample. The three outputs of each column describe one scene, such as a coastal town at sunset with the sound of waves (A13). For the painting of a woman at a table, the image shows cups and a plate and the audio contains dishes and cutlery (A16), a detail that the caption leaves out.

In the one-to-many routes, the two outputs are decoded independently from one latent sample, so any detail they share beyond the input must also be carried by that sample. For the ducks and speech in A34, both outputs turn the speech into a crowd, with the caption describing people talking near quacking ducks and the image showing a crowd lining a canal full of ducks. Given the horse under lightning, the caption mentions the lightning and the audio adds thunder and rain (A29).

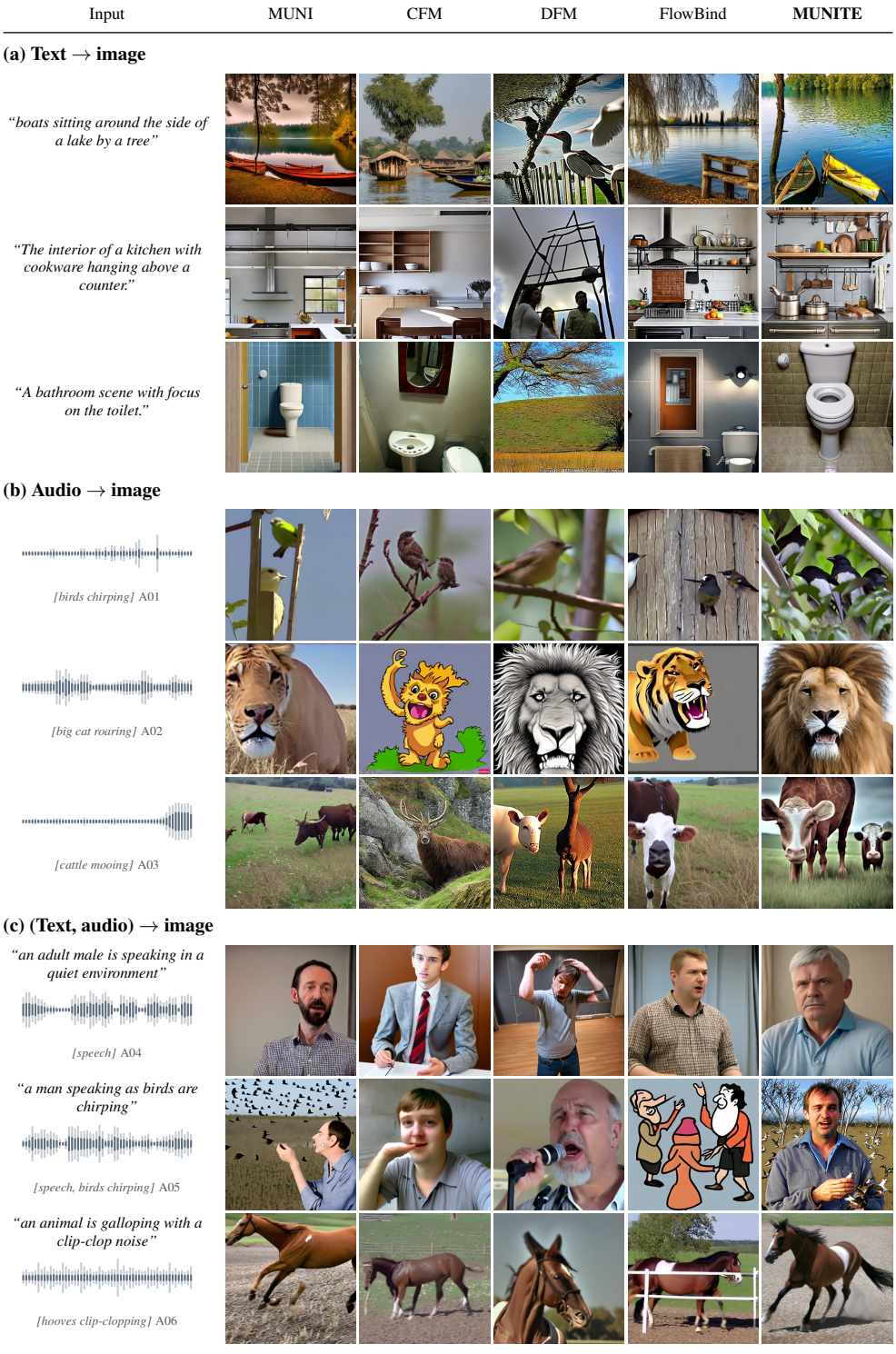



Figure 10: Image–text–audio one-to-one and many-to-one generation of images. Audio inputs are shown as waveforms.

| Input | MUNI | CFM | DFM | FlowBind | **MUNITE** |
|---|---|---|---|---|---|
| **(a) Image → text** | | | | | |
|  | A street light is covered in snow and has a green dome. | Many people in the picture are crossing the street light. | Night in the city, the night lights on snow covered mountains. | Side shot of the street with green rain. | traffic lights on the side of a road in winter |
|  | A street sign that says New York City Hall. | A long-sleeved photo shows the city with its large, stopped traffic. | Silverware on the side of a road with lights. | Closeup picture of a scene where a guy has stopped all the traffic signs | A sign on the side of a street that says, “Stop!” |
|  | Three teddy bears and two cups on a table. | Children, with plastic knee braces, are playing with a table of toddlers. | Purple roses are a great gesture for a female. | Silverware and tea bags are on the table with several teddy bears. | teddy bears sitting at a table with tea cups |
| **(b) Audio → text** | | | | | |
|  *[speech, motorboat, wind]* A07 | wind blows and a man and child speak | wind rustling, a child and an adult man speak in the distance | Asian men are talking, a man and a child talk after loud | wind rushes, a child and man speak, a child talks | a man and a child talk while wind blows |
|  *[speech, ducks quacking]* A08 | a man speaks as ducks quack | wind and rustling, people speaking, and ducks | windmills and people in the background with ducks | wind blows, and a man speaks, and ducks quack | ducks quack while a man talks |
|  *[train passing, horn]* A09 | a train horn blows as it passes | emergency sirens on a truck, with a loud opening of a train | wind, a train running, a horn blows | wind rustling, a train running on something, a horn | a train is blowing its horn as it passes by |
| **(c) (Image, audio) → text** | | | | | |
|  *[running water, speech]* A10 | water running from a faucet | wind rushes, a man talking, a water splash | wind rustles, a man speaks running water | heavy machinery, a man speaks, water running | water running while a man speaks |
|  *[glockenspiel melody]* A11 | bells ring out repeatedly | adult females speak, bells chime | tapping sounds, and some ringing of bells | animals are barking, and bells ring. | bells ring out playing a song |
|  *[woman speaking]* A12 | A woman sitting at a table with a microphone. | female speaking in front of a microphone | birds loudly chirping, a girl speaks | turquoise chairs, a woman speaking, | a woman is giving a speech |

Figure 11: Image–text–audio one-to-one and many-to-one generation of text. Audio inputs are shown as waveforms.




Figure 12: Unconditional image–text–audio co-generation by MUNITE. Each column decodes one latent sample into text, image, and audio.

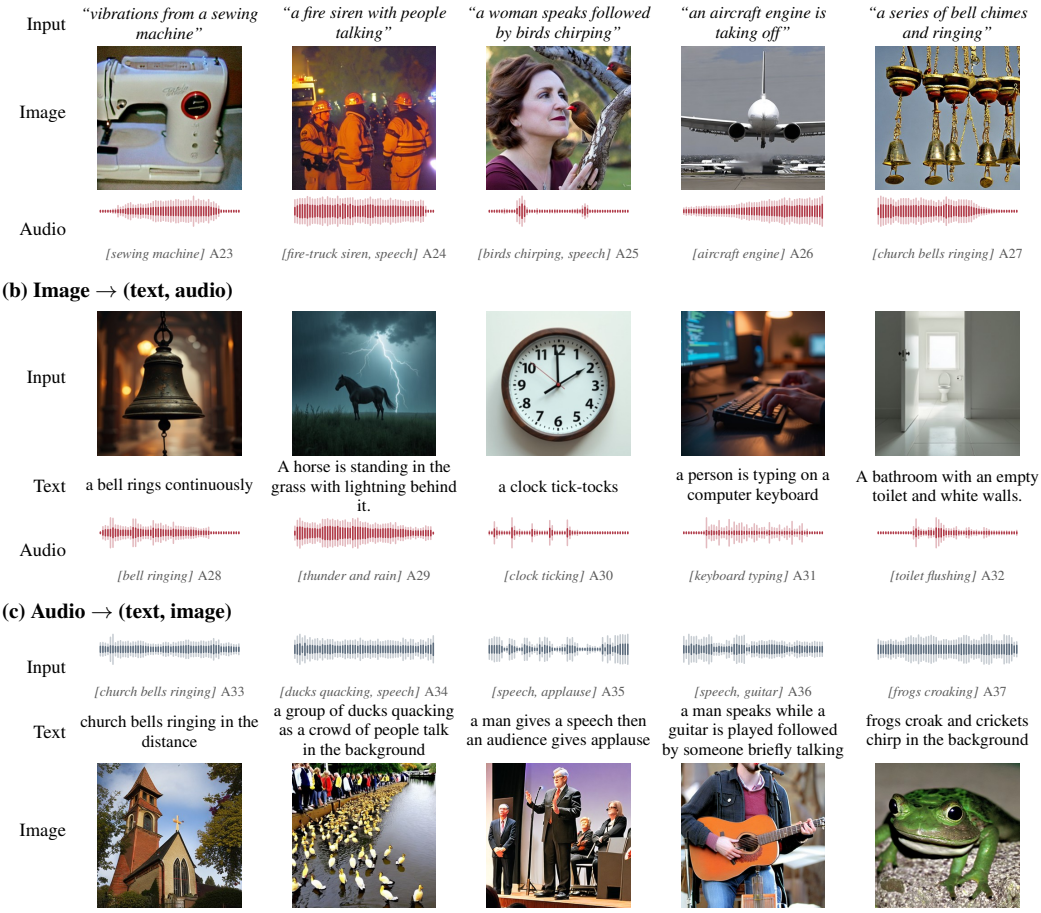



Figure 13: One-to-many image–text–audio generation by MUNITE. In each column, the two outputs are decoded independently from one latent sample inferred from the input.