Split and Bridge: Multimodal Generation via Diffusion Bridging

Ahmad Arrabi[1*] **Xiaohan Zhang**[1] **Xingyu Li**[2] **Safwan Wshah**[1*]

[1]Department of Computer Science, University of Vermont
[2]Center for Excellence in Brain Science and Intelligence Technology

Abstract

Visual data are often expressed through multiple complementary modalities (e.g., images and segmentation masks), each capturing different aspects of the same underlying structure. Most existing multimodal methods adopt a top-down paradigm, learning a shared latent variable through joint training from scratch, which limits its scalability. We propose Split and Bridge (SnB), a bottom-up framework for multimodal generation that couples pretrained unimodal diffusion models at sampling time. Rather than learning a unified latent space, SnB generates coherent multimodal samples by following a shared diffusion trajectory up to a split point and then branching into modality-specific trajectories guided by a bridge module. This design enables plug-and-play reuse of powerful pretrained models without joint training, making it scalable and easily adaptable. Across standard benchmarks (PolyMNIST, CelebAMask-HQ) and more complex settings (PIE-Bench++, ROCO), SnB achieves a strong balance between sample quality and coherence. We show that SnB naturally extends to downstream zero-shot tasks such as interleaved co-generation, image editing, and style coupling/decoupling. Our results demonstrate that bottom-up generation provides a practical framework that complements existing top-down multimodal approaches. Code is available at `https://github.com/AhmadArrabi/SnB`

1 Introduction

Different visual modalities capture multiple complementary aspects of the same underlying scene. For example, a segmentation mask highlights geometric shapes, whereas the image itself captures details about the scene. This variety of visual representations increases the demand for models that capture their shared underlying structure as well as modality-specific information. A particularly powerful formulation of this goal is multimodal generation, which aims to learn a joint distribution that enables the simultaneous sampling of coherent and high-fidelity samples of multiple modalities [17, 47, 45]. Accurate modeling of this joint distribution implies capturing the underlying mechanism that generated these modalities. This capability benefits multiple downstream applications such as representation learning [54], interleaved generation [50, 5], story telling [1], image editing, and medical imaging [55].

Existing multimodal generation methods rely on a *top-down* approach that learns a hidden latent variable z that encompasses all modalities through joint training from scratch [17, 41, 29, 45, 23, 44]. However, empirical evidence reveals several critical limitations of top-down approaches: **1)** They fail to scale to complex, high-resolution visual data (e.g., 512×512 images) due to expensive latent bottleneck design (see Section 4.1), **2)** They depend on *restrictive* joint-posterior assumptions, such as the Product of Experts (PoE) or Mixture of Experts (MoE) [53, 41, 45], **3)** The unified latent design

*{ahmad.arrabi, safwan.wshah}@uvm.edu

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

imposes rigidity, limiting adaptability to diverse downstream tasks (see Section 4.2), and **4)** They suffer from the coherence-quality trade-off, mainly in VAE-based formulations [7] (see Section 4.1). Recent diffusion-based extensions (e.g., MM-Diffusion, MLD [37, 4]) partially alleviate fidelity issues. However, they still inherit key limitations as they rely on modality-specific architectures and require large scale training from scratch with paired data. For all those reasons, the dependence on top-down modeling for multimodal generation limits scalability in practice.

In contrast, a *bottom-up* approach adopts a compositional design, where modality-specific models are developed independently and later coupled for cross-modal interactions. This paradigm offers greater flexibility and scalability by utilizing strong pretrained unimodal models. Similar organizational principles are observed in biological systems, where distinct sensory pathways process different modalities while interacting before higher-level integration [27, 9, 35, 10, 43, 18].

Motivated by this perspective, we propose **S**plit a**n**d **B**ridge (SnB), a bottom-up framework for multimodal generation. SnB leverages pretrained unimodal diffusion models (e.g., LDM [36], SDXL [21], and FLUX [21]) and introduces a bridge module that connects modalities during sampling. This formulation constructs the joint generative process through coupling modality-specific information. As a result, it relaxes restrictive joint-posterior assumptions while retaining the scalability of modern diffusion models. Given two (or more) diffusion models, we share a common reverse trajectory up to a split time step, then diverge into modality-specific branches via a bridge module. Since all trajectories originate from a common intermediate state, we hypothesize that the resulting samples share structure while preserving modality-specific information. SnB is a general framework, and we propose multiple bridge designs based on dataset complexity and the required task. This modular design makes SnB flexible and easier to scale while gaining competitive performance in multimodal generation.

To illustrate SnB's flexibility and performance, we evaluate on standard multimodal generation benchmarks, including PolyMNIST [45] and CelebAMask-HQ [22]. SnB consistently produces coherent multimodal samples while maintaining high generation fidelity. SnB supports both unconditional joint generation and conditional generation within a unified framework while being competitive with top-down approaches. Beyond benchmark evaluation, we further demonstrate the practical flexibility of SnB through downstream tasks on PIE-Bench++ [16], including interleaved co-generation and image editing. We also apply SnB on a practical use case, multimodal MRI generation using the ROCO [32] radiology dataset. In summary, our contributions are as follows:

- We propose Split and Bridge (SnB), a bottom-up framework for multimodal generation that composes pretrained unimodal diffusion models without requiring joint retraining. SnB avoids the scalability limitations and restrictive assumptions of top-down approaches.
- We introduce bridge mechanisms that enable controlled coupling of modality trajectories, supporting different levels of modality interaction.
- We demonstrate that SnB achieves superior performance compared to top-down multimodal methods in both sample quality and coherence on CelebAMask-HQ, while providing competitive performance on PolyMNIST.
- We show that SnB can be extended to downstream tasks, including interleaved co-generation, image editing, and MRI synthesis, highlighting its flexibility and practical applicability.

2 Related Work

VAE-based methods Multimodal VAE methods were initially introduced to address the multimodal generation task through a top-down hidden variable approach [47, 53, 51]. Early works extended the VAE framework by training multiple encoders, one for each modality, to infer a shared latent variable $\mathbf{z}$ [47]. However, these models neither explored the structure of the latent representation nor scaled effectively as the number of modalities increased [29]. To tackle this limitation, more theoretically grounded approaches [53, 41] have been proposed to explicitly model the multimodal posterior $p(\mathbf{z}|\mathbf{x}_1, \ldots, \mathbf{x}_m)$, where m is the number of modalities. These works introduced different assumptions on the joint posterior, such as the Product of Experts (PoE) [53], Mixture of Experts (MoE) [41], and Mixture of Product of Experts (MoPoE) [45], leading to enhanced modeling. Further refinements addressed specific weaknesses of multimodal VAEs like proposing improved training objectives [44, 17, 29, 30, 52, 46] and learning disentangled representations to separate modality-

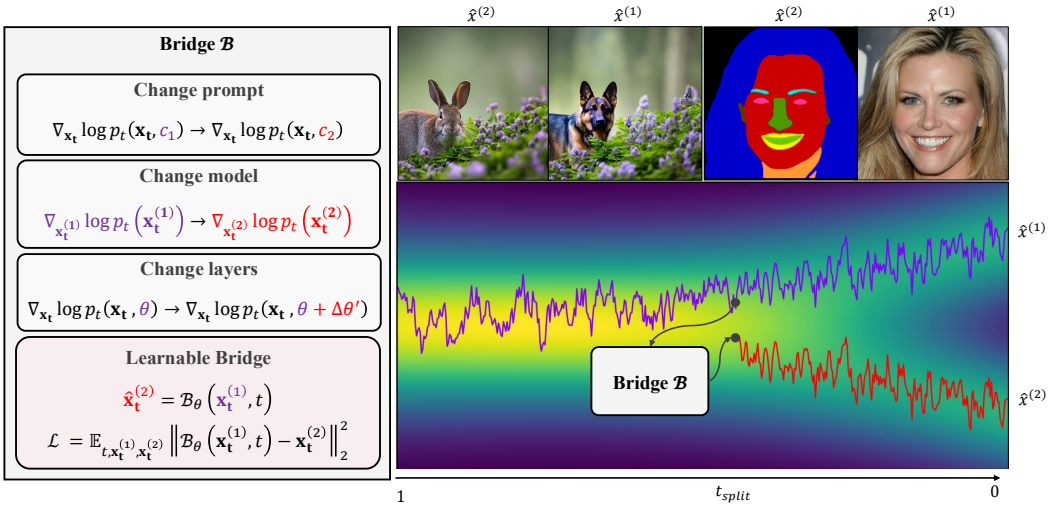

Figure 1: SnB leverages pretrained diffusion models by coupling reverse trajectories during sampling. We start from a common state sampled from a prior, then at t_{split}, a bridge module is used to implement a score intervention on the reverse process. This intervention can be through modifying conditioning signals (e.g., prompts), changing diffusion models that estimate the score functions, or changing certain layers in the diffusion models. This flexibility applies SnB in various tasks, like interleaved co-generation (top left pair) and multimodal generation (top right pair). Learnable bridges allows stronger trajectory manipulation in mapping one modality to another.

private and -shared information [23, 29]. Recent work in [6] has shown promising multimodal generation capabilities, but still faces top-down limitations like controllability and flexibility issues.

Diffusion-based methods Recent advances in diffusion models have inspired many works to apply them in multimodal problems. MM-Diffusion [37] introduced the first diffusion-based model for joint audio–video generation. It assumes two independent forward processes per modality and models their correlation through a single shared reverse process implemented via a U-Net with multimodal blocks. Although MM-Diffusion demonstrates impressive results in the audio–video domain, its architecture is highly specialized and tailored to this specific task, limiting its generalization to visual modalities. Similarly, MLD [4] proposed a diffusion-based joint generation framework, but follows a top-down paradigm in which the diffusion process is applied on a concatenation of all latent representations encoded through pretrained encoders. This design leads to an expensive multi-stage training procedure that restricts scalability. Another family of diffusion models are diffusion bridges, which learn a stochastic transport process between modalities, mapping samples from one domain to another [19, 25, 59]. However, a key difference between multimodal models and diffusion bridges is that the latter are *not* designed for joint generation. Instead, they address a transport problem between existing modalities, typically assuming a predefined source-target relationship.

Connection to Conditional Generation Multimodal generation has advanced significantly in conditional generation, where models learn a directed mapping from a source modality to a target [56, 24, 28, 57]. Although effective for tasks like text-to-image [57] or image-to-image translation [58], these methods rely on strong supervision (e.g., prior image) and thus *fall outside the scope of joint generation*. More recently, large foundation diffusion models such as CoDI and CoDI-2 [49, 48] have enabled any-to-any generation across modalities. However, they remain inherently conditional, rather than jointly sampling from a true multimodal distribution (co-generation without priors).

3 Methodology

Diffusion Models Preliminaries Diffusion models learn a data distribution $p_{\text{data}}(\mathbf{x})$ through a continuous-time stochastic process that progressively perturbs data into noise and reverses this process for generation [42]. The forward process is defined by the stochastic differential equation (SDE),

$$d\mathbf{x}_t = f(\mathbf{x}_t, t)\, dt + g(t)\, d\mathbf{w}_t \tag{1}$$

where $f(\mathbf{x}_t, t)\colon \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is the drift coefficient, $g(t)\colon [0,1] \to \mathbb{R}$ is the scalar diffusion coefficient, and $\mathbf{w}_t \in \mathbb{R}^d$ is a standard Wiener process. The process is defined over $t \in [0,1]$, where $\mathbf{x}_t$ evolves from clean data $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$, and as $t \to 1$, $\mathbf{x}_t$ converges to a tractable Gaussian prior $p_1(\mathbf{x}_1) \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$. [42] and [2] showed that Equation (1) admits a corresponding reverse-time SDE that reconstructs data by inverting the noising dynamics,

$$d\mathbf{x}_t = \Big[f(\mathbf{x}_t, t) - g(t)^2 \, \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) \Big] dt + g(t) \, d\bar{\mathbf{w}}_t \tag{2}$$

where $p_t(\mathbf{x}_t)$ is the marginal density of $\mathbf{x}_t$ under the forward process, and $\bar{\mathbf{w}}_t$ is a standard Wiener process running in reverse time (i.e., as t decreases from 1 to 0). The term $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is the Stein score function of the marginal distribution at time t. By approximating it, one can simulate the reverse-time SDE starting from $\mathbf{x}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to generate samples approximately distributed according to $p_{\text{data}}(\mathbf{x})$.

Problem Formulation Let $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ denote a tuple of m modalities drawn from an unknown joint distribution $p_{\text{data}}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)})$. The goal of multimodal generation is to generate samples $\hat{\mathbf{X}}$ that show unimodal fidelity and cross-modal coherence. Here, coherence means that all generated modalities share semantic factors (e.g., object geometry, facial attributes, or digit identity).

Multimodal generation is often tackled from a top-down approach in learning a hidden variable $\mathbf{z}$ by approximating its posterior in the following general form [45],

$$q_\phi(\mathbf{z} \mid \mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(m)}) = \mathcal{C}\left(q_{\phi_1}(\mathbf{z} \mid \mathbf{x}^{(1)}), \cdots, q_{\phi_m}(\mathbf{z} \mid \mathbf{x}^{(m)}) \right) \tag{3}$$

where $\mathcal{C}$ is an aggregation function that combines all individual posteriors per modality, $\phi = \{\phi\}_{j=1}^m$ is the model's parameters. For example, in [53] $\mathcal{C}$ was defined as a product function, which led to the PoE formulation, [41] used a summation, leading to the MoE formulation.

Split and Bridge (SnB) We don't limit multimodal generation to restrictions over the posterior's functional form. We consider the problem alternatively from a *bottom-up* approach. Rather than explicitly learning posteriors, or training large multi-branch models [4, 49, 37], we leverage pretrained unimodal diffusion models, and learn to *bridge* their trajectories *during sampling* to achieve joint multimodal generation. We consider each modality modeled by its reverse SDE following Equation (2), with a drift learned through training its individual diffusion model.

Algorithm 1 SnB

Input: Diffusion models $\mathcal{M}_1, \mathcal{M}_2$, bridge $\mathcal{B}$, t_{split}
Sample $\mathbf{x}_1^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = 1 \to 0$ **do**
  if $t = t_{\text{split}}$ **then** ▷ split
    $\mathbf{x}_t^{(2)} \leftarrow \mathcal{B}(\mathbf{x}_t^{(1)}, t)$ ▷ bridge
  end if
  $\hat{\mathbf{y}}_t^{(1)} \leftarrow \mathcal{M}_1(\mathbf{x}_t^{(1)}, t)$ ▷ noise / velocity / score
  $\mathbf{x}_{t-\Delta t}^{(1)} \leftarrow \text{Denoise}(\mathbf{x}_t^{(1)}, \hat{\mathbf{y}}_t^{(1)})$
  if $t \leq t_{\text{split}}$ **then**
    $\hat{\mathbf{y}}_t^{(2)} \leftarrow \mathcal{M}_2(\mathbf{x}_t^{(2)}, t)$
    $\mathbf{x}_{t-\Delta t}^{(2)} \leftarrow \text{Denoise}(\mathbf{x}_t^{(2)}, \hat{\mathbf{y}}_t^{(2)})$
  end if
end for
Output: $\{\mathbf{x}_0^{(1)}, \mathbf{x}_0^{(2)}\}$

Multimodal generation requires joint sampling of a modality set $\mathbf{X}$. Thus, we propose a new sampling framework, as illustrated in Figure 1, where we first sample a single diffusion trajectory (any chosen initial modality) backward in time starting at $t = 1$ to an intermediate step t_{split}. Then, we divide the denoising process into m independent processes, one per modality. We hypothesize that this split generates a latent noisy representation that contains the initial structure of the synthesized object (e.g., a noisy face of a person), which incentivizes correlated trajectories. At t_{split}, we introduce a bridge module $\mathcal{B} : \mathbb{R}^d \to \mathbb{R}^k$ that transports one modality closer to the other through a transport operation, where d and k are the dimensions of the corresponding modalities. After t_{split}, each trajectory would be determined by its own reverse process that is driven by its modality-specific diffusion model. More formally, we can define the joint sampling process as multiple reverse SDEs as shown in Equation (4). For clarity and notation simplicity, we present the case of $m = 2$ but the formulation extends to m modalities.

$$
\begin{aligned}
d\mathbf{x}_t^{(1)} &= \left[f_1(\mathbf{x}_t^{(1)}, t) - g_1(t)^2 \nabla_{\mathbf{x}_t^{(1)}} \log p_t(\mathbf{x}_t^{(1)}) \right] dt + g_1(t) d\mathbf{w}_t \\
d\mathbf{x}_t^{(2)} &= \begin{cases} d\mathbf{x}_t^{(1)} & : t > t_{split} \\ \mathcal{B}(\mathbf{x}_t^{(1)}, t) & : t = t_{split} \\ \left[f_2(\mathbf{x}_t^{(2)}, t) - g_2(t)^2 \nabla_{\mathbf{x}_t^{(2)}} \log p_t(\mathbf{x}_t^{(2)}) \right] dt + g_2(t) d\mathbf{w}_t & : t < t_{split} \end{cases}
\end{aligned} \tag{4}
$$

Algorithm 1 implements SnB in Equation (4), where $\mathcal{M}_1$ and $\mathcal{M}_2$ are diffusion models pretrained for each modality. SnB is agnostic to the diffusion parameterization (e.g., score-based [42], noise-prediction [14], or velocity-prediction [38]). Sampling begins by drawing $\mathbf{x}_1^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and denoising it along the reverse trajectory defined by $\mathcal{M}_1$ until the split timestep t_{split}. At t_{split}, a second trajectory is initialized via the bridge module, $\hat{\mathbf{x}}_{t_{split}}^{(2)} = \mathcal{B}\big(\mathbf{x}_{t_{split}}^{(1)}, t_{split}\big)$, which maps the state from the first modality to the second. Finally, the two representations $\mathbf{x}_t^{(1)}$ and $\mathbf{x}_t^{(2)}$ are denoised in parallel by their respective models to produce the coherent output set $\{\mathbf{x}_0^{(1)}, \mathbf{x}_0^{(2)}\}$.

Bridge Module $\mathcal{B}$ The transport strength needed at t_{split} depends largely on the discrepancy between modalities and the complexity of the generated data. We therefore introduce different bridges that can be applied to different data complexity levels and tasks. This flexibility is mainly a result of the bottom-up formulation of SnB, as it allows leveraging strong modality-specific models.

i. Identity Bridge We first define $\mathcal{B}$ as an identity mapping as shown in Equation (5). Despite its simplicity, the identity bridge enables multimodal generation by switching the score function that governs the reverse process after t_{split}. Concretely, a shared latent state is continued under a new score function, effectively transitioning from one data distribution to another without explicit transport: $\nabla_{\mathbf{x}_t^{(1)}} \log p_t(\mathbf{x}_t^{(1)}) \to \nabla_{\mathbf{x}_t^{(2)}} \log p_t(\mathbf{x}_t^{(2)})$. This perspective shows that multimodal generation can be achieved through interventions on the score function during sampling, rather than via joint training. The identity bridge therefore represents the minimal mechanism for coupling modalities.

$$
\hat{\mathbf{x}}_{t_{split}}^{(2)} = \mathcal{B}(\mathbf{x}_{t_{split}}^{(1)}, t_{split}) = \mathbf{x}_{t_{split}}^{(1)} \tag{5}
$$

We consider several ways in which a score intervention can happen, all unified under the same principle: 1) *Model switch*, where the reverse process transitions between pretrained diffusion models corresponding to different modalities. This case is effective in simple data structures where a model per modality is feasible. 2) *Condition switch*, where the score is modified via conditioning signals (e.g., prompt changes in text-guided diffusion models), enabling controlled semantic variation, and 3) *Parameter switch*, where lightweight adaptations (e.g., LoRA[15]) alter the score function without replacing the full model allowing zero-shot compositional generation (see Appendix D).

Different kinds of score interventions allow SnB to integrate with pretrained and foundation diffusion models. For instance, in text-conditioned models, a simple prompt change at t_{split} acts as an effective bridge, preserving global structure while inducing targeted variations aligned with the new condition. The identity bridge is most effective when modalities share similarity, such that the intermediate state provides a meaningful initialization across branches. In larger modality gaps, stronger bridge designs that explicitly transform the latent state would be needed (see Section 4.1).

ii. Learnable Bridge When the domain gap between modalities is large (e.g., faces and segmentation maps), directly switching the drift from f_1 to f_2 at t_{split} leads to poor generation (as will be shown in Section 4.1). The intermediate sample $\mathbf{x}_{t_{split}}^{(1)}$ results in an improper initialization for the new trajectory. To address this, we introduce a learnable bridge network $\mathcal{B}_\theta$ that maps samples across modalities at the splitting point. The bridge acts as a drift correction mechanism, projecting $\mathbf{x}_{t_{split}}^{(1)}$ onto a suitable region of the target modality's latent space. Formally, $\mathcal{B}_\theta$ is defined as follows,

$$
\hat{\mathbf{x}}_{t_{split}}^{(2)} = \mathcal{B}_\theta(\mathbf{x}_{t_{split}}^{(1)}, t_{split}) \tag{6}
$$

After bridging, each modality continues its own reverse diffusion trajectory driven by its diffusion model that controls the drift f_i, where i is the modality index. Because the transition occurs under shared noise, the generated modalities remain stochastically aligned. We implement $\mathcal{B}_\theta$ as a time-conditioned U-Net that takes a noisy sample $\mathbf{x}_{t_{split}}^{(1)}$ as input and outputs its aligned counterpart $\mathbf{x}_{t_{split}}^{(2)}$.

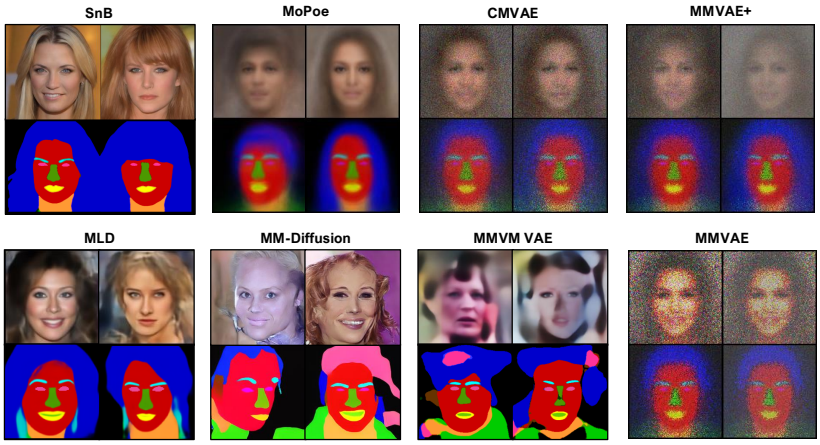

Figure 3: Qualitative comparison of SnB and previous top-down multimodal approaches on CelebAMask-HQ. SnB balances between coherence and quality of generated modality pairs.

During training, we sample time steps $t_{split} \sim \mathcal{U}(0, 1)$ uniformly to ensure $\mathcal{B}_\theta$ can operate anywhere on the diffusion trajectory $[0, 1]$. The bridge is trained using the following reconstruction loss,

$$\mathcal{L}_{\text{bridge}} = \mathbb{E}_{t,\mathbf{x}_t^{(1)},\mathbf{x}_t^{(2)}} \|\mathcal{B}_\theta(\mathbf{x}_t^{(1)}, t) - \mathbf{x}_t^{(2)}\|_2^2 \tag{7}$$

A Mean Squared Error (MSE) objective encourages $\mathcal{B}_\theta$ to approximate the conditional expectation, which can reduce stochastic variation in $\hat{\mathbf{x}}_t^{(2)}$. As a result, the predicted state may become overly smooth and misaligned with the noise level expected by the diffusion process at t_{split}. Inspired by [26], we reintroduce noise into $\hat{\mathbf{x}}_t^{(2)}$ using the diffusion scheduler (derived from the coefficient $g_2(t)$ in Equation (1)) to ensure consistency with the reverse trajectory. We provide additional details and a formal derivation in Appendix A.

4 Experiments

Implementation Details We evaluate SnB on PolyMNIST [45], CelebAMask-HQ [22], PIE-Bench++ [16], and ROCO [32]. We compare against top-down baselines, including VAE-based methods (MVAE [53], MMVAE, MMVAE+, MoPoE, mmJSD [44], MMVM VAE[46]) and diffusion-based approaches (MLD[4] and MM-Diffusion variants with both Stable Diffusion [36] and ResNet [11] U-Net backbones for a fair comparison). Image quality is measured using FID, and segmentation quality using F1 score when applicable. For coherence, we follow prior work on PolyMNIST using pretrained classifiers [45]. We also introduce a learned metric for CelebAMask-HQ that measures the cosine similarity between ℓ_2-normalized embeddings of generated modalities. The full implementation details are provided in Appendix B.

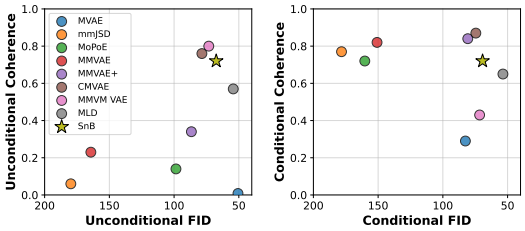

Figure 2: PolyMNIST results using the identity bridge with SnB. SnB strikes a balance between coherence and quality across the 5 modalities.

4.1 Multimodal Generation

Identity Bridge Following prior works, we first evaluate the identity Bridge on PolyMNIST. This benchmark provides a simple multimodal case suitable for evaluating the identity bridge with a direct score switch. Figure 2 compares SnB against prior approaches in both conditional and unconditional generation. SnB prevailed in unconditional generation, striking a good balance between coherence and quality. In the conditional setting, it remains competitive with prior methods while attaining the highest image quality (lowest FID) with minimal loss in coherence. These results suggest that even a minimal bridge design is sufficient to capture shared structure in simple multimodal cases. Qualitative results further support these findings and are provided in Appendix C.

Learnable Bridge We evaluate SnB with a learnable bridge on CelebAMask-HQ. *Note that* here, t_{split} denotes the percentage of the trajectory shared before the split. For instance, $t_{split} = 30\%$ means that the first 30% of the diffusion process is shared across modalities. As shown in Table 1, SnB achieves the best image quality (FID) across all t_{split} values while maintaining strong coherence, consistently outperforming both VAE- and diffusion-based baselines. In contrast, VAE methods achieve higher coherence at the cost of image quality, while MM-Diffusion improves fidelity but struggles to maintain alignment. In conditional generation (face→mask and mask→face), SnB achieves comparable performance, achieving high-quality synthesis with accurate structural alignment, as reflected by competitive F1 scores for mask prediction. Qualitatively (Figures 3 and 4), SnB produces realistic samples with clear correspondence between faces and masks, significantly outperforming top-down baselines. Additional examples are provided in Appendix C.

Table 1: Results of multimodal generation on CelebAMask-HQ. SnB achieved the highest coherence without losing generation quality in conditional and unconditional generation. † and * refer to the used backbones: SD and ours, respectively.

Method	Joint		M→F		F→M	
	Coherence ↑	FID ↓	Coherence ↑	FID ↓	Coherence ↑	F1 ↑
MoPoE [45]	0.59	151.23	0.50	163.83	0.51	0.72
CMVAE [30]	0.52	303.50	0.48	233.04	0.47	0.56
MMVAE [41]	0.58	358.21	0.46	375.79	0.47	0.58
MMVAE+ [29]	0.56	345.29	0.48	202.86	0.47	0.57
MMVM VAE [46]	0.51	178.87	**0.65**	218.69	0.67	0.73
MM-Diffusion† [37]	0.49	91.08	0.57	353.10	0.56	0.15
MM-Diffusion*	0.50	137.75	0.51	230.04	0.56	0.15
MLD [4]	0.68	121.69	0.64	227.00	0.71	0.76
SnB ($t_{split} = 10\%$)	0.49	60.24	0.49	200.84	0.49	0.54
SnB ($t_{split} = 20\%$)	0.52	**59.22**	0.49	201.62	0.54	0.62
SnB ($t_{split} = 30\%$)	0.60	59.73	0.50	**65.50**	0.62	0.76
SnB ($t_{split} = 40\%$)	0.67	60.30	0.55	66.83	0.68	0.82
SnB ($t_{split} = 50\%$)	0.73	60.84	0.60	107.22	0.72	0.85
SnB ($t_{split} = 60\%$)	0.76	60.20	0.59	155.90	0.75	0.87
SnB ($t_{split} = 70\%$)	0.78	59.47	0.55	208.05	0.76	**0.88**
SnB ($t_{split} = 80\%$)	**0.79**	59.23	0.55	180.58	**0.77**	0.82
SnB ($t_{split} = 90\%$)	0.78	59.41	0.56	195.57	0.76	0.81

Ablation Studies We ablate two main components in the learnable bridge: the bridge network $\mathcal{B}_\theta$ and the noise injector. The results are shown in Table 2. While the FID score for unconditional generation remains stable (around 60) across all ablations, the coherence changes significantly. Removing the bridge network (identity bridge) causes a substantial decrease in coherence, indicating that without it, the generated samples are not well-correlated even if they are realistic; this finding is illustrated in Figure 5. These results clearly show the need for learnable bridges in complex data structures compared to a simple change of score function at t_{split}. The noise injector primarily impacts conditional generation, as the FID increases when it is omitted in the mask→face case. We further ablate the capacity of $\mathcal{B}_\theta$ across small, medium, and large variants in Table 3 by increasing and depth and base channels of the U-net. Increasing the size of the bridge consistently improves coherence while maintaining strong FID, suggesting that a larger capacity places the trajectory in a better initialization for the second model. These findings show that bridge design choices significantly affect performance.



Figure 4: conditional generation of SnB on CelebAMask-HQ.

Sensitivity Analysis on t_{split} A natural question that arises in SnB is: *when should the diffusion trajectories be split, and how does the choice of t_{split} affect generation quality and coherence?* We answer these questions by conducting a sensitivity analysis by varying the split point t_{split} across the reverse process on both PolyMNIST and CelebAMask-HQ. The corresponding results are shown in Table 1 and Figure 6. The coherence shows a clear trend: *later splits (higher t_{split}) lead to higher coherence, meaning that the generated representations are more correlated with each other*. This behavior aligns with our core hypothesis regarding shared latent representations. A larger shared segment of the reverse trajectory (e.g., $t_{split} = 70\%$) allows both modalities to co-evolve for a longer duration, enhancing their structural alignment. Conversely, earlier splits (e.g., $t_{split} = 30\%$) reduce the overlap in shared latent evolution, leading to weaker correspondence. Moreover, the coherence drops after a certain point (around $t_{split} = 70\%$) given that more steps are needed to move the trajectory to the new modality distribution.

Table 2: Ablation study of SnB. Br. and N.I. denote the bridge and noise injector. Coh. is coherence.

Ablation		Joint		M→F		F→M	
Br.	N.I.	Coh. ↑	FID ↓	Coh. ↑	FID ↓	Coh. ↑	F1 ↑
✓	✓	0.73	60.84	**0.60**	107.22	**0.72**	**0.85**
✓	×	**0.76**	**59.89**	0.50	216.39	0.70	0.78
×	✓	0.48	60.82	0.50	**52.03**	0.48	0.53
×	×	0.48	59.90	0.50	292.58	0.49	0.02

Table 4: Quantitative evaluation on PIE-Bench for interleaved co-generation and image editing tasks.

Model	SnB Variant	Interleaved Co-Generation								Image Editing				
		Target		Source		Consistency				Target				
		FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	LPIPS ↓	MSE ↓	PSNR ↑	FID ↓	LPIPS ↓	MSE ↓	PSNR ↑	CLIP ↑
Flux [21]	Independent	76.60	32.58	75.27	32.18	61.93	0.67	0.10	10.00	-	-	-	-	-
	shared noise	76.09	32.38	74.74	32.18	56.32	0.48	0.03	14.76	78.24	0.76	0.15	8.25	32.52
	SnB	**75.95**	31.79	74.86	32.23	**51.69**	**0.39**	**0.02**	**17.03**	76.94	0.70	0.09	10.55	32.45
SDXL [33]	Independent	77.61	**33.59**	**74.14**	**32.63**	70.96	0.69	0.08	11.03	-	-	-	-	-
	shared noise	79.03	33.39	74.59	32.57	68.16	0.50	0.03	15.02	110.41	0.71	0.11	9.77	31.30
	SnB	77.49	33.32	74.49	32.54	65.53	0.45	**0.02**	16.59	114.47	0.69	0.09	10.46	30.96
SD 2.1 [36]	Independent	81.82	33.27	75.61	32.54	75.36	0.77	0.10	9.94	-	-	-	-	-
	shared noise	82.07	33.13	76.13	32.39	71.47	0.54	0.04	13.99	86.94	0.69	0.09	10.65	31.69
	SnB	79.93	32.70	76.89	32.45	67.00	0.48	0.03	15.10	88.57	0.66	0.07	11.44	31.53
SD 1.5 [36]	Independent	79.88	33.05	76.69	32.42	72.84	0.78	0.13	8.85	-	-	-	-	-
	shared noise	80.22	32.91	75.54	32.28	68.98	0.55	0.05	12.88	77.19	0.67	0.08	11.07	**32.73**
	SnB	78.77	32.66	77.00	32.46	67.28	0.49	0.04	13.93	**75.54**	**0.63**	**0.06**	**11.99**	32.50

4.2 Scalability and Flexibility of Bottom-up SnB

We analyze SnB's scalability through leveraging large foundation diffusion models [21, 33, 36] on downstream tasks. We evaluate on PIE-Bench++ using the identity bridge with a prompt swap at t_{split} as the score intervention. We consider two tasks, *interleaved co-generation*, where paired images are generated from two prompts with shared structure, and *image editing*, where an input image is noised to t_{split} and then denoised under a new prompt. We compare against two baselines, independent diffusion models per prompt (no shared noise) and shared-noise diffusion models (could be considered SnB with $t_{split} = 0\%$). Table 4 shows that across multiple backbones, all methods achieve comparable FID and CLIP [12] scores, indicating similar image quality and prompt alignment. However, SnB achieves the highest pair consistency, producing semantically aligned pairs while preserving prompt-specific variations. This advantage is clearly visualized in Figures 7 and 8, where the generated images are aligned and representing the prompts. These results highlight that SnB enables coherent multi-output generation at scale, acting as a plug-and-play framework for foundation models. Additional details and visualizations are provided in Appendix D.

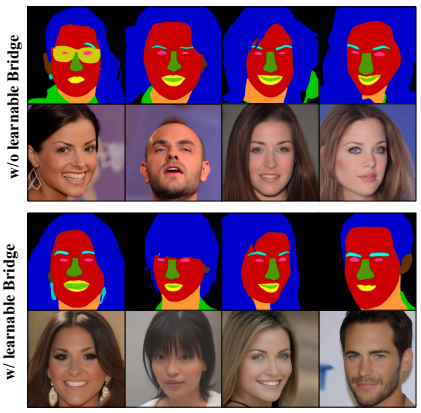

Figure 5: Images generated with a learnable bridge show superior results in terms of coherence over identity bridge in complex datasets.

Another form of score intervention is through parameter modification (Section 3). Leveraging the availability of task-specific fine-tuned variants of diffusion models (e.g., LoRA [15]), we demonstrate the compositionality of SnB on the style transfer subset of PIE-Bench++. We apply two distinct LoRA styles[2][3] at different timesteps, introducing two intervention points. Starting from the base model, we activate style 1 at $t^{(1)}_{style}$ and subsequently style 2 at $t^{(2)}_{style}$. This type of score modification results in blended images that exhibit both styles. Remarkably, we demonstrate that this process is reversible, where individual styles can be decoupled from their combination, as shown in Figure 9. These results highlight the ability of SnB to compose and control multiple model adaptations within a unified sampling process. Additional results and details are provided in Appendix D.

Table 3: Ablation study on the learnable bridge ($\mathcal{B}_\theta$) capacity on CelebAMask-HQ at $t_{\text{split}} = 50\%$. Medium denotes the default configuration used in all other experiments.

Bridge Variant	Depth	Base Channels	Params (M)	Mem. (MB)	Coh. ↑	FID ↓
Small	2	16	2.1	8.37	0.68	**60.14**
Medium	4	32	15.8	60.29	0.73	60.84
Large	6	64	233.3	890.18	**0.76**	60.72

[2]https://huggingface.co/multimodalart/flux-tarot-v1
[3]https://huggingface.co/SouthbayJay/t-shirt-vector-style-flux

Table 5: Quantitative results across three medical imaging categories of ROCO. Modalities not present in a specific anatomy are marked with a dash (-)

Anatomy	Split Point	PubMedCLIP [8] ↑					FID ↓					Consistency		
		T1	T2	FLAIR	DWI	PD	T1	T2	FLAIR	DWI	PD	MSE ↓	PSNR ↑	LPIPS ↓
Brain	Independent	28.87	**29.42**	30.58	**27.43**	-	137.37	152.98	145.25	134.90	-	0.13	8.79	0.60
	Shared Noise	**28.95**	29.28	**30.70**	27.38	-	137.57	154.88	138.01	132.59	-	0.04	14.06	0.38
	SnB	28.85	29.17	30.24	27.08	-	**134.67**	**147.44**	**134.37**	**132.46**	-	**0.02**	**17.90**	**0.27**
Hip/Shld/Elbow	Independent	30.96	31.15	-	-	30.13	258.82	256.06	-	-	268.05	0.16	7.93	0.61
	Shared Noise	**31.38**	**31.34**	-	-	30.04	**247.53**	**255.18**	-	-	**254.40**	0.04	14.60	0.34
	SnB	30.81	31.10	-	-	**30.19**	258.76	262.59	-	-	258.97	**0.02**	**18.09**	**0.24**
Spine	Independent	31.18	**31.17**	-	-	-	264.25	269.55	-	-	-	0.14	8.64	0.62
	Shared Noise	31.09	30.88	-	-	-	**263.34**	**263.08**	-	-	-	0.02	17.77	0.24
	SnB	**31.25**	31.02	-	-	-	267.65	270.08	-	-	-	**0.01**	**21.36**	**0.16**

Downstream Application: Zero-Shot Multimodal MRI Synthesis Medical imaging data is inherently multimodal, for example, MRI scans provide complementary diagnostic views across multiple contrast levels. Multimodal MRI synthesis is a challenging problem due to the difficulty of collecting paired multi-contrast scans at scale. Existing top-down multimodal generation methods typically require jointly aligned training data across all modalities, which is expensive and often impractical in real-world medical settings. At the same time, recent medical foundation diffusion models have demonstrated strong capabilities in generating high-quality MRI images from text prompts, motivating the question of *whether such pretrained unimodal generators can be composed into a multimodal generation framework without retraining*.

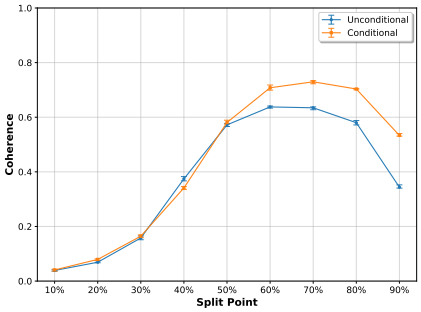

Figure 6: Sensitivity analysis on t_{split} on PolyMNIST

To study this setting, we apply SnB in a zero-shot manner for multimodal MRI synthesis using ROCO and the pretrained medical diffusion model Prompt2MedImage[4]. We consider four common MRI contrasts: T1, T2, FLAIR, and DWI, each capturing complementary physiological properties of the same anatomy. Using the identity bridge with prompt-based score intervention, SnB generates different MRI contrasts from a partially shared diffusion trajectory while preserving the underlying anatomical structure.

We evaluate on a test subset spanning brain, joint, and spine MRI scans. This setting directly tests whether SnB can coordinate independently generated modalities without paired multimodal training. As shown in Table 5, all methods achieve similar FID and PubMedCLIP [8] scores due to the shared pretrained backbone. However, SnB consistently achieves stronger modality consistency, indicating improved structural alignment across generated contrasts. Qualitative examples in Figure 10 further demonstrate that SnB preserves consistent anatomy while producing clear contrast-specific appearance changes across modalities. For example, the prompt 'T2 axial brain magnetic resonance imaging scan showing hypointense rim with hyperintense heterogenous core in the left frontoparietal region" can be transformed into different modalities by replacing 'T2" with 'T1", 'FLAIR", or "DWI".

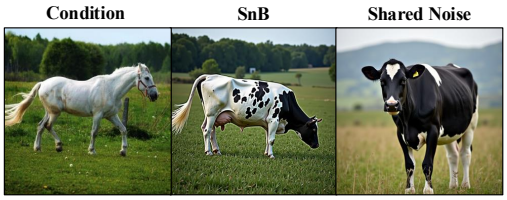

Figure 7: Editing on PIE-Bench++. a gray horse in the field→ a black cow with white patterns in the field.

5 Conclusion

We introduced SnB, a bottom-up framework for multimodal generation that composes pretrained unimodal diffusion models during sampling. Unlike prior top-down approaches that rely on learning a shared latent representation through joint training, SnB couples modality-specific trajectories. This design scales with foundation diffusion models and enables direct reuse of strong pretrained experts.

[4]https://huggingface.co/Nihirc/Prompt2MedImage

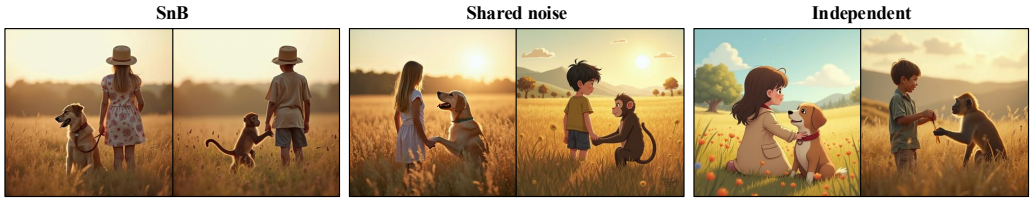

Figure 8: Interleaved co-generation on PIE-Bench++. SnB shows the most correlated co-generated pairs. Prompt 1: a girl and her dog in a field. Prompt 2: a boy and his monkey in a field.

SnB empirically achieves a favorable balance between coherence and sample quality. Beyond standard multimodal generation, we demonstrate its flexibility through zero-shot extensions to downstream tasks by leveraging foundation models, highlighting a level of scalability and practicality not accessible to traditional approaches. Our results in medical imaging further illustrate the potential of this paradigm in real-world settings, suggesting that effective multimodal generation benefits from composable and scalable designs.



Figure 9: Qualitative PIE-Bench++ style results. SnB demonstrates compositional abilities in combining and separating modalities. Prompt: a bird standing on tree branch.

Limitations Despite its flexibility, SnB introduces new design choices. In particular, selecting an appropriate bridge mechanism and t_{split} may require empirical tuning depending on the modality gap and data complexity. Additionally, performance can be sensitive to the choice of the initial trajectory, as some modalities provide more suitable shared representations than others (see Appendix E). Finally, SnB assumes access to pretrained unimodal models for each modality, which may be costly in some settings. However, this requirement is mitigated in practice by the growing availability of large pretrained and adaptable foundation models (see Section 4.2).

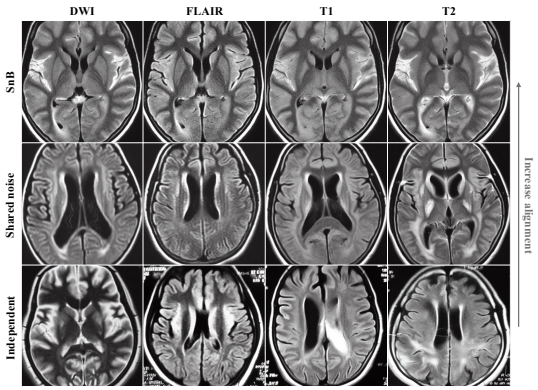

Figure 10: Independent models lead to completely different structures, sharing the noise component increases alignment, while SnB is the most aligned.

Future Work Future directions include extending SnB beyond visual modalities, handling larger semantic gaps between modalities, and developing robust strategies for selecting split points and bridge designs. More broadly, we view the bridge as a modular component, and further exploration of its design and its interaction with diffusion trajectories is a promising avenue. We hope this work encourages a shift toward bottom-up composable approaches for scalable multimodal generation.

6 acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Grant Nos. 2218063 and 2551626. Computations were supported through the AMD University Program AI & HPC Cluster.

References

[1] J. An, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, L. Wang, and J. Luo. Openleaf: Open-domain interleaved image-text generation and evaluation. *arXiv preprint arXiv:2310.07749*, 2023.

[2] B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[3] G. Balakrishnan, Y. Xiong, W. Xia, and P. Perona. Towards causal benchmarking of bias in face analysis algorithms. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, pages 547–563, Cham, 2020. Springer International Publishing.

[4] M. Bounoua, G. Franzese, and P. Michiardi. Multi-modal latent diffusion. *Entropy*, 26(4):320, 2024.

[5] W. Chen, L. Li, Y. Yang, B. Wen, F. Yang, T. Gao, Y. Wu, and L. Chen. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8073–8082, June 2025.

[6] J. Cui, Y.-Y. Chen, Y. Zhang, and M. Klenk. Shala: Multimodal shared latent generative modelling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(25):20658–20666, Mar. 2026.

[7] I. Daunhawer, T. M. Sutter, K. Chin-Cheong, E. Palumbo, and J. E. Vogt. On the Limitations of Multimodal VAEs, Apr. 2022. arXiv:2110.04121 [cs].

[8] S. Eslami, C. Meinel, and G. De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1151–1163, 2023.

[9] A. Falchier, S. Clavagnier, P. Barone, and H. Kennedy. Anatomical Evidence of Multimodal Integration in Primate Striate Cortex. *Journal of Neuroscience*, 22(13):5749–5759, July 2002. Publisher: Society for Neuroscience Section: ARTICLE.

[10] A. A. Ghazanfar and C. E. Schroeder. Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, 10(6):278–285, June 2006.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi. CLIPScore: A reference-free evaluation metric for image captioning. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[13] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[14] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[15] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[16] M. Huang, J. Cai, S. Jia, V. S. Lokhande, and S. Lyu. Paralleledits: Efficient multi-aspect text-driven image editing with attention grouping. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 22569–22595. Curran Associates, Inc., 2024.

[17] H. Hwang, G.-H. Kim, S. Hong, and K.-E. Kim. Multi-View Representation Learning via Total Correlation Objective. In *Advances in Neural Information Processing Systems*, volume 34, pages 12194–12207. Curran Associates, Inc., 2021.

[18] I. Kanemura and K. Kitano. Association between different sensory modalities based on concurrent time series data obtained by a collaborative reservoir computing model. *Scientific Reports*, 13(1):173, Jan 2023.

[19] B. Kim, G. Kwon, K. Kim, and J. C. Ye. Unpaired image-to-image translation via neural schrödinger bridge. In *International Conference on Learning Representations*, volume 2024, pages 19312–19331, 2024.

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] B. F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

[22] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[23] M. Lee and V. Pavlovic. Private-Shared Disentangled Multimodal VAE for Learning of Hybrid Latent Representations, Dec. 2020. arXiv:2012.13024 [cs].

[24] M. Li, T. Yang, H. Kuang, J. Wu, Z. Wang, X. Xiao, and C. Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, editors, *Computer Vision – ECCV 2024*, pages 129–147, Cham, 2025. Springer Nature Switzerland.

[25] J. Liu, F. Petersen, Y. Gao, Y. Zhang, H. Kim, A. S. Chaudhari, Y. Sun, S. Ermon, and S. Gatidis. Unpaired image-to-image translation via a self-supervised semantic bridge. *arXiv preprint arXiv:2602.16664*, 2026.

[26] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

[27] S. Molholm, W. Ritter, M. M. Murray, D. C. Javitt, C. E. Schroeder, and J. J. Foxe. Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Brain Research. Cognitive Brain Research*, 14(1):115–128, June 2002.

[28] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024.

[29] E. Palumbo, I. Daunhawer, and J. E. Vogt. MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises. In *The Eleventh International Conference on Learning Representations*. OpenReview, 2023. Accepted: 2023-10-23T07:38:19Z.

[30] E. Palumbo, L. Manduchi, S. Laguna, D. Chopard, and J. E. Vogt. Deep generative clustering with multimodal diffusion variational autoencoders. In *International Conference on Learning Representations*, volume 2024, pages 24286–24310, 2024.

[31] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[32] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In D. Stoyanov, Z. Taylor, S. Balocco, R. Sznitman, A. Martel, L. Maier-Hein, L. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S.-L. Lee, S. Moriconi, V. Cheplygina, D. Mateus, E. Trucco, E. Granger, and P. Jannin, editors, *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 180–189, Cham, 2018. Springer International Publishing.

[33] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pages 1862–1874, 2024.

[34] V. V. Ramaswamy, S. S. Kim, and O. Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9301–9310, 2021.

[35] R. P. N. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, Jan. 1999. Publisher: Nature Publishing Group.

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

[37] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10219–10228. IEEE, 2023.

[38] T. Salimans and J. Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

[39] A. Senellart, C. Chadebec, and S. Allassonnière. Multivae: A python package for multimodal variational autoencoders on partial datasets. *Journal of Open Source Software*, 10(110):7996, 2025.

[40] A. Serianni, T. Zhu, O. Russakovsky, and V. V. Ramaswamy. Attention iou: Examining biases in celeba using attention maps. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4386–4397. IEEE, 2025.

[41] Y. Shi, S. N, B. Paige, and P. Torr. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[42] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-Based Generative Modeling through Stochastic Differential Equations, Feb. 2021. arXiv:2011.13456.

[43] B. E. Stein, T. R. Stanford, and B. A. Rowland. Multisensory integration and the society for neuroscience: Then and now. *Journal of Neuroscience*, 40(1):3–11, 2020.

[44] T. Sutter, I. Daunhawer, and J. Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in neural information processing systems*, 33:6100–6110, 2020.

[45] T. M. Sutter, I. Daunhawer, and J. E. Vogt. Generalized multimodal ELBO. In *International Conference on Learning Representations*, 2021.

[46] T. M. Sutter, Y. Meng, A. Agostini, D. Chopard, N. Fortin, J. E. Vogt, B. Shahbaba, and S. Mandt. Unity by diversity: improved representation learning for multimodal vaes. *Advances in Neural Information Processing Systems*, 37:74262–74297, 2024.

[47] M. Suzuki, K. Nakayama, and Y. Matsuo. Joint Multimodal Learning with Deep Generative Models, Nov. 2016. arXiv:1611.01891.

[48] Z. Tang, Z. Yang, M. Khademi, Y. Liu, C. Zhu, and M. Bansal. Codi-2: In-context, interleaved, and interactive any-to-any generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27415–27424. IEEE, 2024.

[49] Z. Tang, Z. Yang, C. Zhu, M. Zeng, and M. Bansal. Any-to-Any Generation via Composable Diffusion, May 2023. arXiv:2305.11846 [cs].

[50] C. Tian, X. Zhu, Y. Xiong, W. Wang, Z. Chen, W. Wang, Y. Chen, L. Lu, T. Lu, J. Zhou, et al. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *arXiv preprint arXiv:2401.10208*, 2024.

[51] R. Vedantam, I. Fischer, J. Huang, and K. Murphy. Generative models of visually grounded imagination. In *International Conference on Learning Representations*, 2018.

[52] D. Wesego and P. Rooshenas. Score-based multimodal autoencoder. *Transactions on Machine Learning Research*, 2024.

[53] M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018.

[54] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6995–7004, June 2021.

[55] C. Zhan, Y. Lin, G. Wang, H. Wang, and J. Wu. Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11502–11512, June 2024.

[56] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824. IEEE, 2023.

[57] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:11127–11150, 2023.

[58] L. Zhong, C. Guo, Y. Xie, J. Wang, and C. Li. Sketch2anim: Towards transferring sketch storyboards into 3d animation. *ACM Transactions on Graphics (TOG)*, 44(4):1–15, 2025.

[59] L. Zhou, A. Lou, S. Khanna, and S. Ermon. Denoising diffusion bridge models. In *International Conference on Learning Representations*, volume 2024, pages 8160–8171, 2024.

A Noise Injector in $\mathcal{B}_\theta$

A Mean Squared Error (MSE) objective essentially cannot model stochastic variations, as MSE minimization leads to the conditional expectation $\mathbb{E}[\mathcal{B}_\theta(\mathbf{x}^{(1)}_{t_{split}}) \mid \hat{\mathbf{x}}^{(1)}_{t_{split}}]$ as the optimal predictor. Given that $\mathbb{E}[\epsilon_{t_{split}}] = \mathbf{0}$, the model would never learn noisy latents across the reverse process, resulting in smooth deterministic outputs. To map $\mathbf{x}^{(1)}_{t_{split}}$ to the correct corresponding state $\mathbf{x}^{(2)}_{t_{split}}$ after bridging, we introduce a **noise injector** that adds noise back into $\hat{\mathbf{x}}^{(2)}_t$. The amount of noise added is determined by the diffusion coefficient $g_2(t)$ of the forward process in Equation (1). Practically, this is controlled by the scheduler of the diffusion model. In this section, we provide a formal derivation showing why an MSE loss fails to learn the stochastic mapping between noisy latents, causing the model to converge to the conditional mean predictor and produce non-noisy results.

Let $\mathbf{x}^{(1)}_t$ and $\mathbf{x}^{(2)}_t$ be the noisy latents of their respective modalities at timestep t. For notation simplicity, we consider the noise added by the scheduler at t as ϵ_t, where $\mathbf{x}^{(2)}_t = \gamma_t \mathbf{x}^{(2)}_0 + \epsilon_t$, γ_t and ϵ_t are determined by the scheduler. The standard MSE objective is given as follows,

$$\begin{aligned}\mathcal{L}_{\text{MSE}} &= \mathbb{E}_{t,\mathbf{x}^{(1)}_t,\mathbf{x}^{(2)}_t} \|\mathcal{B}_\theta(\mathbf{x}^{(1)}_t, t) - \mathbf{x}^{(2)}_t\|_2^2 \\ &= \mathbb{E}_{t,\mathbf{x}^{(1)}_t}\left[\mathbb{E}_{\mathbf{x}^{(2)}_0}\left[\|\mathcal{B}_\theta(\mathbf{x}^{(1)}_t, t) - (\gamma_t \mathbf{x}^{(2)}_0 + \epsilon_t)\|_2^2\right]\right]\end{aligned} \tag{8}$$

For any fixed input $(\mathbf{x}^{(1)}_t, t)$, the model predicts a deterministic output $\hat{\mathbf{x}_t}^{(2)} = \mathcal{B}_\theta(\mathbf{x}^{(1)}_t, t)$, we can extend the inner expectation as follows,

$$\begin{aligned}&\mathbb{E}_{\mathbf{x}^{(2)}_0}\left[\|\hat{\mathbf{x}}^{(2)}_t - (\gamma_t \mathbf{x}^{(2)}_0 + \epsilon_t)\|_2^2\right] \\ &= \mathbb{E}_{\mathbf{x}^{(2)}_0}\left[\|(\hat{\mathbf{x}}^{(2)}_t - \gamma_t \mathbf{x}^{(2)}_0) - \epsilon_t\|_2^2\right] \\ &= \mathbb{E}_{\mathbf{x}^{(2)}_0}\left[\|\hat{\mathbf{x}}^{(2)}_t - \gamma_t x^{(2)}_0\|_2^2 - 2(\hat{\mathbf{x}}^{(2)}_t - \gamma_t \mathbf{x}^{(2)}_0)^\top \epsilon_t + \|\epsilon_t\|_2^2\right] \\ &= \|\hat{\mathbf{x}}^{(2)}_t - \gamma_t \mathbf{x}^{(2)}_0\|_2^2 - 2(\hat{\mathbf{x}}^{(2)}_t - \gamma_t \mathbf{x}^{(2)}_0)^\top \mathbb{E}_{\mathbf{x}^{(2)}_0}[\epsilon_t] + \mathbb{E}_{\mathbf{x}^{(2)}_0}[\|\epsilon_t\|_2^2] \\ &= \|\hat{\mathbf{x}}^{(2)}_t - \gamma_t x^{(2)}_0\|_2^2 + \sigma_t^2\end{aligned} \tag{9}$$

Where $\mathbb{E}[\epsilon_t] = 0$ and $\mathbb{E}[\epsilon_t^\top \epsilon_t] = \sigma_t^2$ (σ_t^2 is the variance of the added Gaussian distribution at t). Since σ_t^2 and γ_t are constants with respect to $\hat{\mathbf{x}}^{(2)}_t$, the loss is minimized when $\|\hat{\mathbf{x}}^{(2)}_t - \mathbf{x}^{(2)}_0\|_2^2$ is minimized. Therefore, for any input $\mathbf{x}^{(1)}_t$, the optimal prediction under the MSE loss is the clean latent $\hat{\mathbf{x}_t}^{*(2)} = \mathbf{x}^{(2)}_0$. This proves that the MSE-optimal bridge network ignores all noise terms added and predicts the conditional expectation $\mathbb{E}[\mathbf{x}^{(2)}_t \mid \mathbf{x}^{(1)}_t] = \mathbf{x}^{(2)}_0$, failing to learn the true noisy components in the reverse process.

In practice, we treat the noise injection strength as a hyperparameter and select it empirically. It is typically set to correspond to reverting the reverse process by approximately 10% to 30% of the diffusion trajectory from t_{split}. We note that it had higher impact on conditional generation as can be shown in Table 2.

B Implementation Details

B.1 Datasets

We report multimodal experimental results on multiple benchmarks. PolyMNIST and CelebAMask-HQ are two well-established datasets for multimodal generation, allowing both unconditional joint and conditional generation. PolyMNIST is one of the first datasets proposed in this field, and it consists of five visual modalities of digits with varying backgrounds. Generally, this benchmark is low in complexity due to the lower resolution of images; however, it serves as a great way to measure a model's capability in joint generation across a relatively high number of modalities, five.

CelebAMask-HQ, on the other hand, is much more complex and higher in resolution. It contains $30,000$ paired faces and segmentation masks at 512×512 resolution, we use 15-class masks, extending

beyond the binary masks used in prior works [4, 53]. Figure 11 shows a batch of samples from PolyMNIST alongside the binary-mask variant from [4, 53] for comparison.



Figure 11: A sample of PolyMNIST. Unlike prior works that use binary segmentation masks (right samples), we use a more comprehensive mask with 15 classes (center samples) [4, 53].

PIE-Bench++ was introduced as an image editing benchmark where a source image is given along with a source and target prompts. The goal is to generate corresponding target images that align with both the source image and target prompt (only change what is mentioned in text). We view this dataset as a valuable asset that can be used in a direct application of multimodal generation, interleaved co-generation. Given the high-quality annotations, we can use only the source and target prompts for joint generation. The generated pairs are required to be both correlated and faithful to their text descriptions. A strong multimodal model can be used for interleaved co-generation which facilitates data augmentation and large scale dataset generation for such tasks and benchmarks. A sample of PIE-Bench++ is shown in Figure 12. The dataset included 9 different image editing types, e.g., object, colour, and background changes. In total, the benchmark consists of 700 samples.



Source Prompt	a slanted mountain bicycle on the road in front of a building	a girl and her dog in a field
Target Prompt	a slanted rusty mountain motorcycle in front of a fence	a boy and his monkey in a field

Figure 12: A sample of PIE-Bench++. The prompts can be leveraged for interleaved co-generation.

ROCO is a large-scale medical imaging benchmark consisting of paired radiology images and textual descriptions, commonly used for tasks such as image captioning and visual question answering. Its diversity in anatomical regions, imaging modalities, and clinical conditions also makes it suitable for text-to-image generation. For our experiments, we use a curated subset of the test split, focusing on MRI data from three anatomical groups: brain, joints (hips, shoulders, and elbows), and spine. The final evaluation set consists of 113 brain, 47 joint, and 35 spine MRI scans. All samples were manually verified for quality and relevance.

B.2 Baselines

We compare against top-down VAE-based methods (MVAE [53], mmJSD [44], MoPoE [45], MMVAE [41], MMVAE+ [29], MMVM VAE [46], and CMVAE [30]). For all multimodal VAE baselines, we use the publicly available implementation from [39] ($\beta = 1$). For diffusion-based baselines, we follow the standard top-down design used in multimodal diffusion models, where all modality

streams are coupled within a shared U-Net denoiser and processed in a single reverse diffusion process. A representative example is MM-Diffusion [37], for which we report two variants for fair comparison: one based on a Stable-Diffusion-style U-Net matching our backbone, and another using a ResNet-based U-Net. We note that MVAE and mmJSD fail to produce meaningful high-resolution samples on CelebAMask-HQ, and are therefore evaluated only on PolyMNIST. We also acknowledge the recent work in [6], but do not include quantitative results due to the lack of publicly available code at the time of submission. In our scalability analysis (Section 4), we focus on foundation model-based baselines, given the nature of the task and to ensure a fair comparison. We further observe that standard top-down baselines are not directly applicable to text-to-image settings due to their reliance on jointly modeled modalities during training.

B.3 Evaluation Metrics

B.3.1 Coherence

Measuring coherence, i.e., how well two generated representations correspond to one another, is not a direct task. Prior work typically relies on pretrained classifiers to assess cross-modality agreement; if all classifiers predict the same label for each modality, the sample is considered coherent [45, 53, 4, 41, 29, 7, 44, 17]. While effective for simple datasets such as PolyMNIST, such classifier-based metrics fail in more complex settings like face–mask joint generation. Moreover, previous benchmarks on CelebAMask-HQ evaluate coherence only in conditional settings where a ground truth mask exists, using metrics such as the F1 score between generated and true masks [4]. These limitations motivate the need for a coherence metric for joint generation on PolyMNIST.

Consequently, We develop a learned coherence metric based on a shared embedding space. We train encoders that project the two representations into a joint embedding space, where positive pairs (segmentation masks and faces of the same person) are moved closer, and negative pairs are pushed apart. These encoders are trained using a soft-margin triplet loss of the following form,

$$\mathcal{L}_{triplet} = \log\left(1 + \exp^{\alpha\left(||h_m^{(1)} - h_m^{(2)}||_2^2 - ||h_m^{(1)} - h_n^{(2)}||_2^2\right)}\right) \tag{10}$$

where α is a scaling hyperparameter set to 10, and $m, n \in \{1, \ldots, N\},\ m \neq n$. Here, $h_m^{(1)}$ and $h_m^{(2)}$ denote the embeddings of the positive mask–face pair, $||\cdot||_2^2$ is the squared Euclidean distance, and ResNet-34 backbones were used to extract $h^{(1)}$ and $h^{(2)}$ [11]. A schematic of the contrastive learning setup is shown in Figure 13.

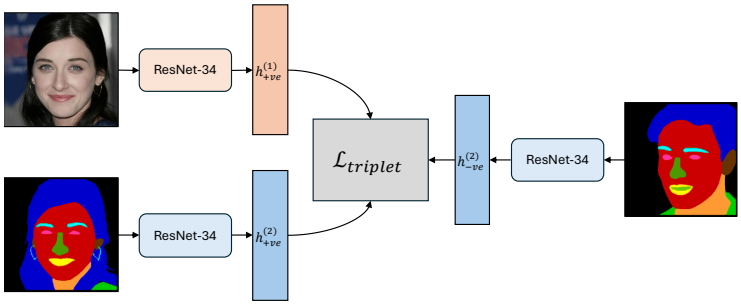

Figure 13: Contrastive learning framework used to learn the embedding space for coherence measurement. Each representation has its own ResNet-34 model, sharing the architecture not the weights.

After training the ResNet-34 encoders, coherence between any mask–face pair is computed using the distance between their normalized embeddings, where $||h^{(1)}||_2 = ||h^{(2)}||_2 = 1$. The squared Euclidean distance can be simplified to the following,

$$\begin{aligned} ||h^{(1)} - h^{(2)}||_2^2 &= ||h^{(1)}||_2^2 - 2(h^{(1)} \cdot h^{(2)}) + ||h^{(2)}||_2^2 \\ &= 2 - 2(h^{(1)} \cdot h^{(2)}) \end{aligned} \tag{11}$$

which lies in the range $[0, 4]$ due to the ℓ_2 normalization. To map this distance to a coherence score in $[0, 1]$, we compute the following,

$$\text{coherence} = 1 - \frac{||h^{(1)} - h^{(2)}||_2^2}{4} = 1 - \frac{2 - 2(h^{(1)} \cdot h^{(2)})}{4} \tag{12}$$

Thus, higher scores indicate better alignment between the generated representations in the learned space.

B.3.2 Image Quality Metrics

We evaluate image fidelity using Fréchet Inception Distance (FID) [13], which measures the distributional similarity between generated and real images using features from a pretrained Inception network. For modalities involving segmentation masks, we additionally report the F1 score to assess segmentation accuracy. In Section 4.2, to evaluate pairwise consistency, we use LPIPS, MSE, and PSNR. These metrics are applied both to measure agreement between generated modality pairs and, in image editing settings (Section 4.2), to compare edited outputs with their corresponding source images. When generating multiple modalities (e.g., 2-4 in ROCO), we report the average consistency across all modality pairs. We further assess text-image alignment using CLIP and PubMedCLIP, depending on the domain.

B.4 Training and Hardware Details

All models in the main manuscript were implemented in PyTorch [31]. For PolyMNIST, we train a separate DDPM for each modality. We use a U-Net backbone with attention layers at 32×32 resolution and train using a standard noise prediction objective with an MSE loss. The forward process follows a DDPM scheduler with 1000 timesteps. Models are trained with AdamW (learning rate 2×10^{-4}) and a cosine learning rate schedule with warmup, using a batch size of 128. We maintain an exponential moving average (EMA) of model weights for stable sampling. Training was done for 500 epochs.

For CelebAMask-HQ, we used the Adam [20] optimizer with $\beta = (0.9, 0.999)$, a learning rate of 1×10^{-4}, and a batch size of 32. The face model was a fine-tuned Stable Diffusion v1.5 backbone, while the mask model followed a DDPM-style architecture with five ResNet downsample blocks. The learnable bridge network was implemented as a U-Net that connects the diffusion models of the two modalities. Thus, its input channels depend on the underlying modality (e.g., four channels for the face model due to Stable Diffusion's latent representation). The U-Net bridge had four downsampling and upsampling layers leading to a depth of 4, with SiLU activations throughout and a base channel of size 32. All training and fine-tuning experiments on CelebAMask-HQ were done for 50 epochs, and were conducted on a single AMD MI210 GPU.

C More Qualitative Results

In this section, we present additional qualitative results of SnB. More unconditional generations are shown in Figures 14 to 16, where we include randomly selected batches. In PolyMNIST, it is clear that the generated digits are the same while only the background changes. Regarding CelebAMask-HQ, we observe a slight tendency toward generating blonde female faces, a known bias previously reported in works using the same dataset [3, 40, 34]. SnB consistently produces coherent face–mask pairs that remain faithful to each other without sacrificing image quality. A zoomed-in example highlighting finer details of the unconditional generation is provided in Figure 17. We note these samples were generated with setting $t_{split} = 50\%$.

For conditional generation, we visualize results in both directions: mask→face and face→mask. Figure 18 shows the face → mask direction, where the first two rows contain the input faces. The generated masks align well with their corresponding faces, particularly in head shape and pose. Conversely, Figure 19 presents the mask→face direction. This task is inherently more challenging, as faces contain substantially richer information than segmentation masks. Transporting information from a low-detail modality to a high-detail one requires learning a more complex mapping. Nevertheless, the generated faces in Figure 19 exhibit strong structural alignment with the input masks and preserve fine facial details.

C.1 Failure Cases

We report some failure cases in our generated outputs, particularly in the conditional setting on CelebAMask-HQ. As shown in Figure 20, some generated faces deviate from their conditioning masks. For example, exhibiting mismatched hair length, altered facial expressions, or missing accessories. We consistently found that earrings and hats were difficult to reproduce, which we attribute to the inherent difficulty of transporting low-detail segmentation masks to high-detail face images. This limitation suggests that the mask→face direction may require a stronger or more expressive bridge network.

D Scalability Experiments Details

D.1 Interleaved Co-Generation

For experiments on PIE-Bench++, we use $t_{split} = 5\%$ unless stated otherwise. As illustrated in Figures 21 and 22, sharing a larger portion of the initial diffusion trajectory leads to more correlated outputs across modalities. We observe that, in large diffusion models, most structural and semantic attributes are established in early timesteps. Consequently, increasing t_{split} improves alignment (e.g., matching pose and orientation), as seen in the later rows of Figure 21. However, this introduces a trade-off: smaller t_{split} values (e.g., 0%–5%) better preserve prompt-specific details, while larger values favor stronger cross-modal consistency.

D.2 Style Coupling and Decoupling

We further demonstrate compositionality through parameter-based score interventions using LoRA. As shown in Figures 23 and 24, multiple styles can be combined by sequentially applying LoRA adaptations during sampling. Starting from the base model parameters θ_{base}, we first inject a LoRA module at timestep $t^{(1)}_{style}$, making it $\theta_{\text{base}} + \Delta\theta_1$, followed by a second LoRA intervention at $t^{(2)}_{style}$, leading to $\theta_{\text{base}} + \Delta\theta_1 + \Delta\theta_2$. Intuitively, this process can also be decoupled by adding noise back to a composed sample to an intermediate timestep t_{decouple} and continuing sampling with a different style adaptation, enabling partial recovery or modification of individual styles from the combined output. In our experiments, we use $t^{(1)}_{style} = 10\%$ and $t^{(2)}_{style} = 10\%$. Prompts are taken from the style-change subset of PIE-Bench++.

D.3 Multimodal MRI Generation

Additional qualitative results for multimodal MRI generation on ROCO are shown in Figure 25. We include reference samples for each modality for clarity. SnB produces consistent anatomical structures across modalities while varying only the contrast (e.g., T1, T2), as observed for both $t_{split} = 5\%$ and 10%. The quantitative results in Table 5 correspond to $t_{split} = 5\%$, which provides a good balance between coherence and diversity. While these results demonstrate the practical potential of SnB, especially in data-scarce domains such as medical imaging, further evaluation with clinically validated metrics and expert assessment is required for real-world deployment.

E Initial Modality Choice

Does the direction of the bridge matter? How does the initial modality choice before t_{split} influence performance? For example, in face–mask joint generation, should the trajectory originate from the face and then branch to the mask, or should it be the other way around? To answer this, we report the performance of both variants in Table 6. We evaluate two settings: one that begins with the face trajectory and splits at $t_{split} = 50\%$, and another that starts from the mask and then diverges. As shown, in joint generation, starting from the face produces higher coherence and better FID scores. Although the mask-first variant still functions, its performance lags behind the face-first variant.

We attribute this gap to the inherent domain gap between modalities. Mapping a low-information representation (mask) to a high-information one (face) is fundamentally more challenging than the reverse. This difficulty likely drives the observed performance discrepancy. Therefore, in our

main experiments, we adopt the face-first direction as the default bridging strategy (unless in the mask→face conditional case).

Table 6: Results of multimodal generation on CelebAMask-HQ at $t_{split} = 50\%$ with different initial trajectories. Starting with the face modality leads to more coherent samples.

Initial trajectory	**joint**		**M→F**		**F→M**	
	coherence ↑	FID ↓	coherence ↑	FID ↓	coherence ↑	F1 ↑
face-first	**0.73**	**60.84**	-	-	**0.72**	**0.85**
mask-first	0.61	97.20	**0.60**	**107.22**	-	-

F Societal Impact

Recent advances in generative modeling have led to widespread adoption across many applications. Our framework has the potential to support positive downstream uses, such as controllable content creation, personalized user experiences, and new artistic workflows, as demonstrated by our experiments in Section 4.2. These capabilities highlight clear opportunities for creative and valuable scientific tools built on multimodal generation. However, as with any model capable of manipulating or synthesizing faces, there is a risk of misuse in harmful contexts, including unauthorized identity manipulation or deceptive content creation. We emphasize the importance of responsible deployment, transparency about synthetic outputs, and adherence to ethical guidelines when applying such models.

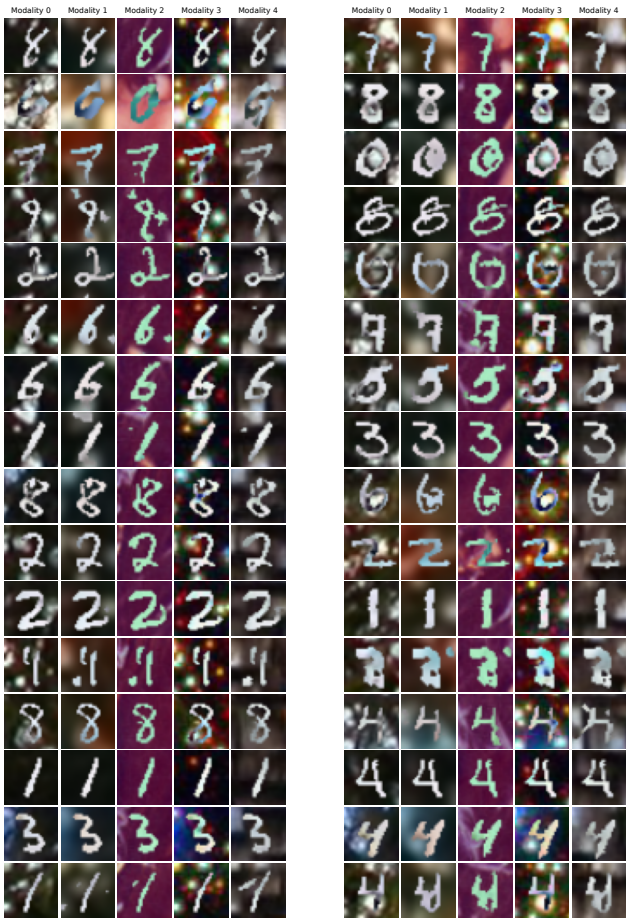

Figure 14: Unconditional generation batched samples on PolyMNIST. All modalities share the same structure while only changing the backgrounds.

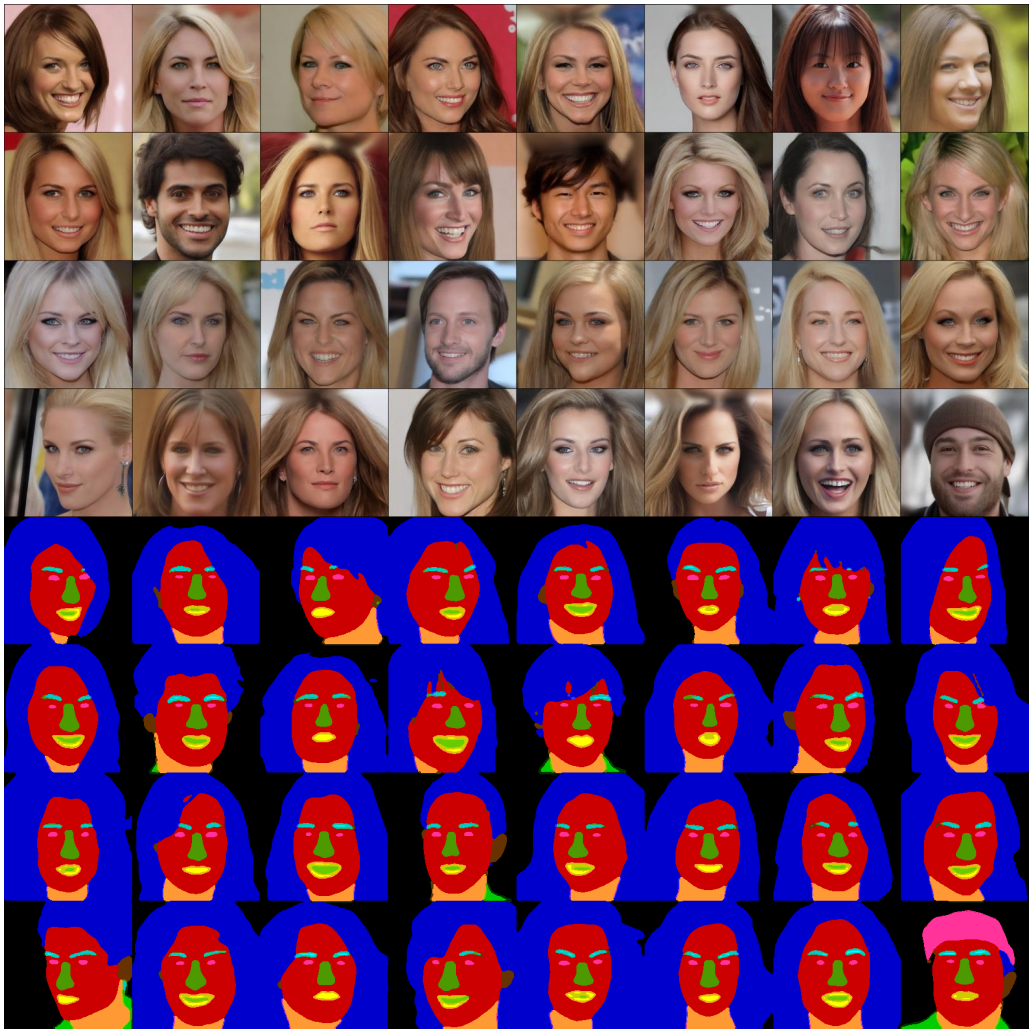

Figure 15: Unconditional generation batched samples. In all cases, the poses and face structure is shared between the segmentation masks and faces. We notice a slight bias towards blonde females in unconditional generation

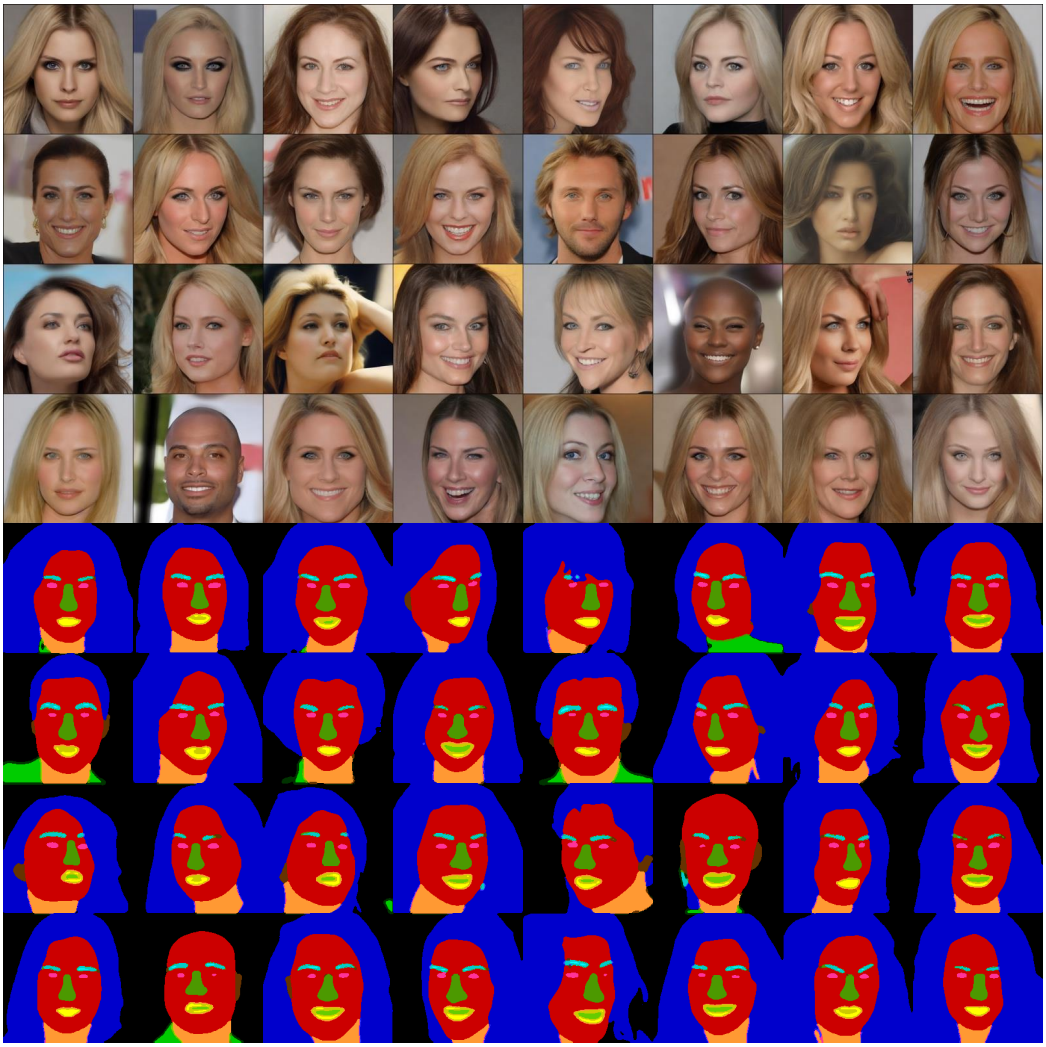

Figure 16: Unconditional generation batched samples. In all cases, the poses and face structure is shared between the segmentation masks and faces. We notice a slight bias towards blonde females in unconditional generation



Figure 17: Unlike previous VAE-based approaches, we are able to generate high-resolution 512×512 samples. These samples provide much more detail and quality which increase the scalability of joint unconditional generation.

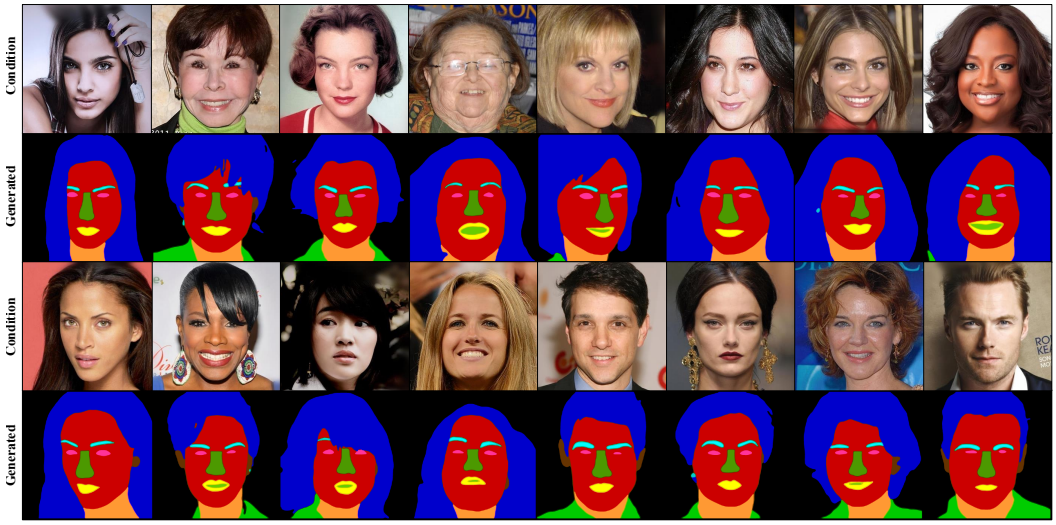

Figure 18: Face $\rightarrow$ mask conditional generation. The generated masks are aligned with the given conditioning faces. The poses are consistent, while some hair details may not be always captured

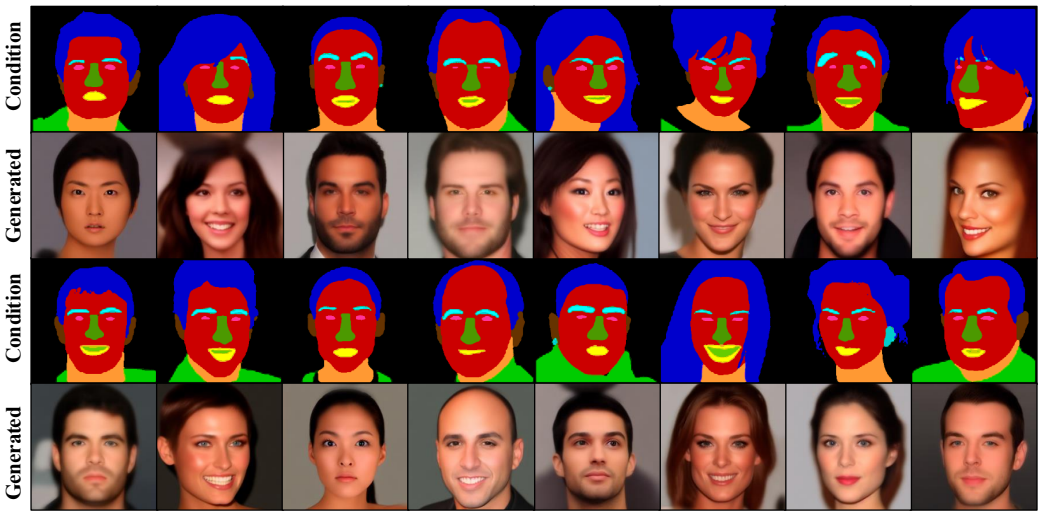

Figure 19: Mask $\rightarrow$ Face conditional generation. The generated faces are aligned with the given conditioning masks. The poses, hair length, and expressions are consistent.

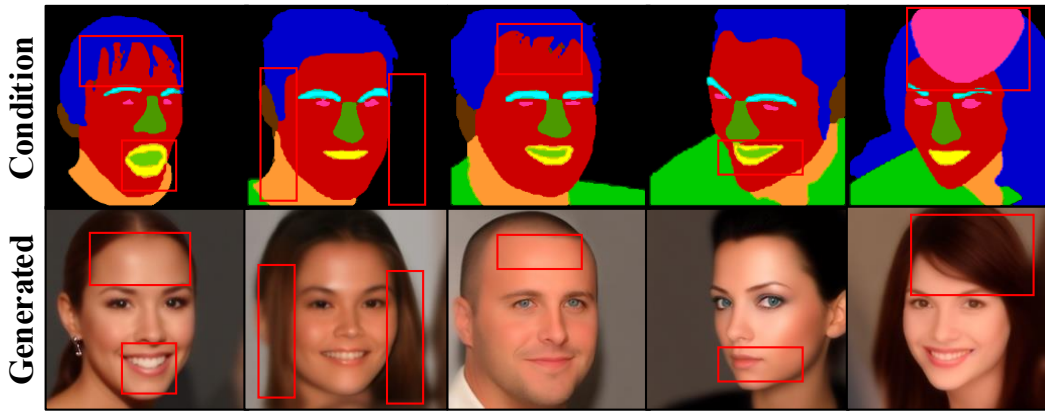

Figure 20: Failure cases in conditional generation on CelebAMask-HQ. Some generated faces misalign with the conditioning masks, particularly for accessories such as glasses, hats, and earrings (rightmost image). We also observe occasional inconsistencies in hair styles or lengths.

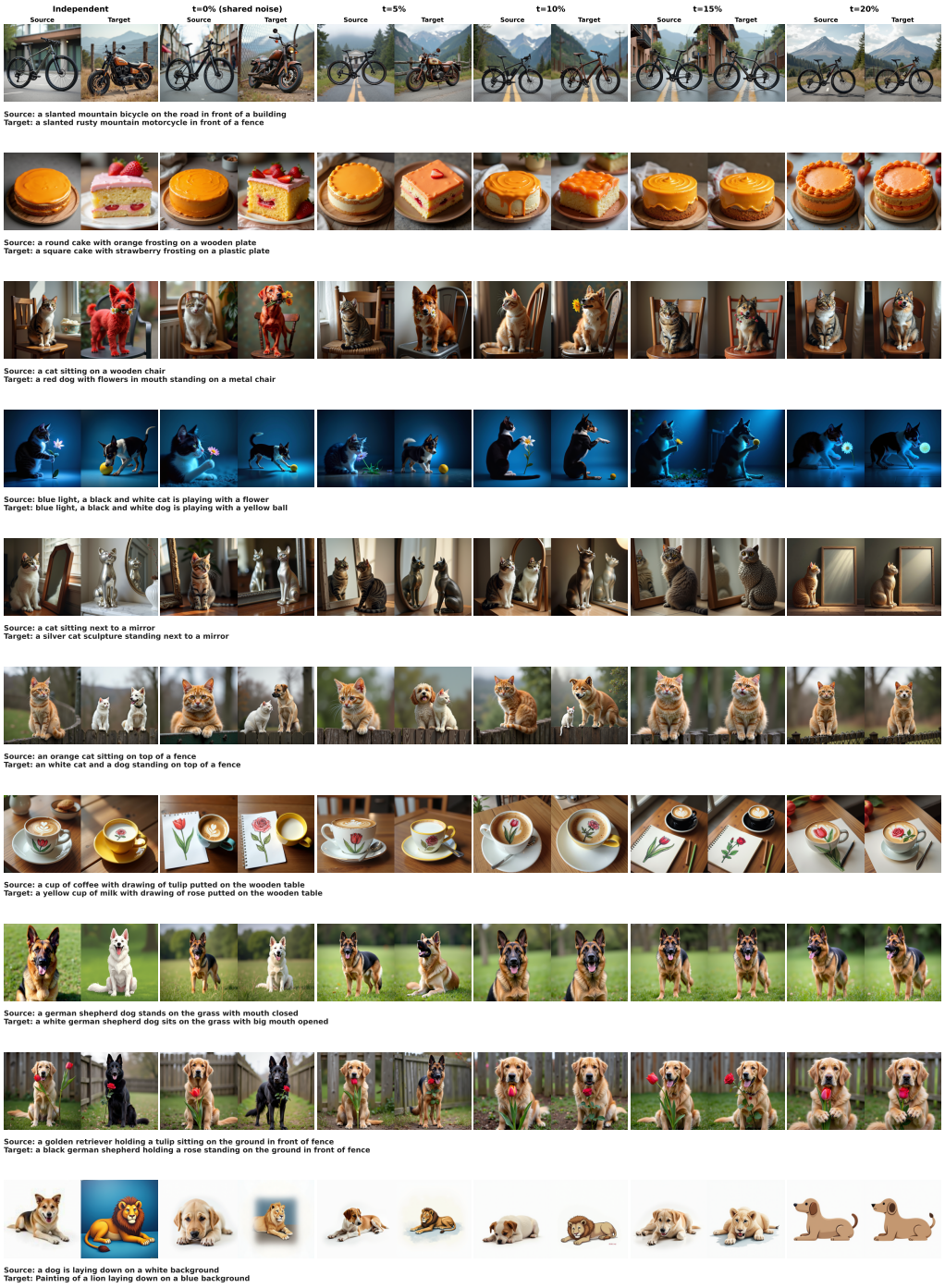

Figure 21: Interleave cogeneration on random samples from PIE-Bench++. Using SnB always leads to correlated pairs compared to independently using pretrained models. These results were generated through FLUX.

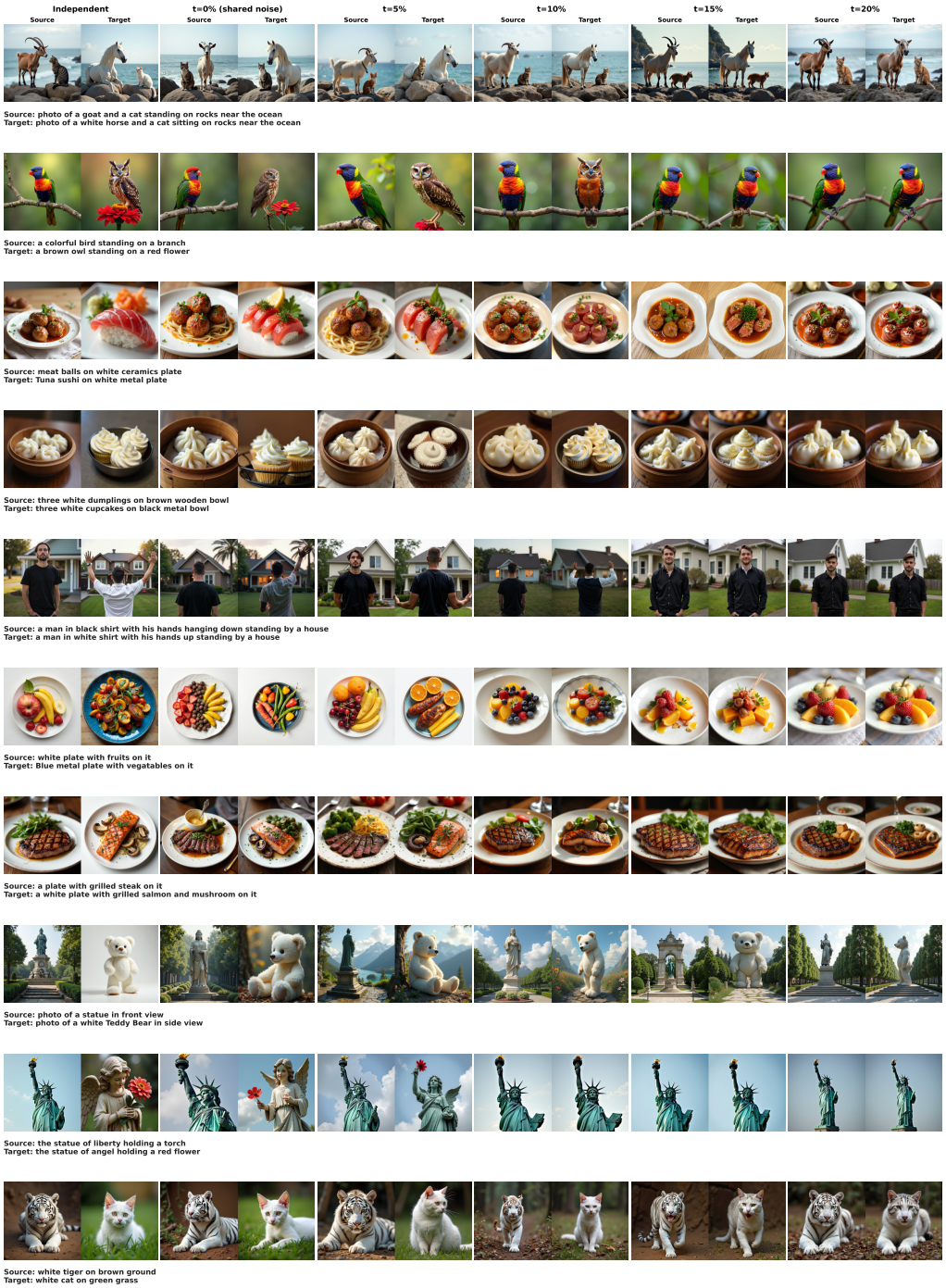

Figure 22: Interleave cogeneration on random samples from PIE-Bench++. Using SnB always leads to correlated pairs compared to independently using pretrained models. These results were generated through FLUX.

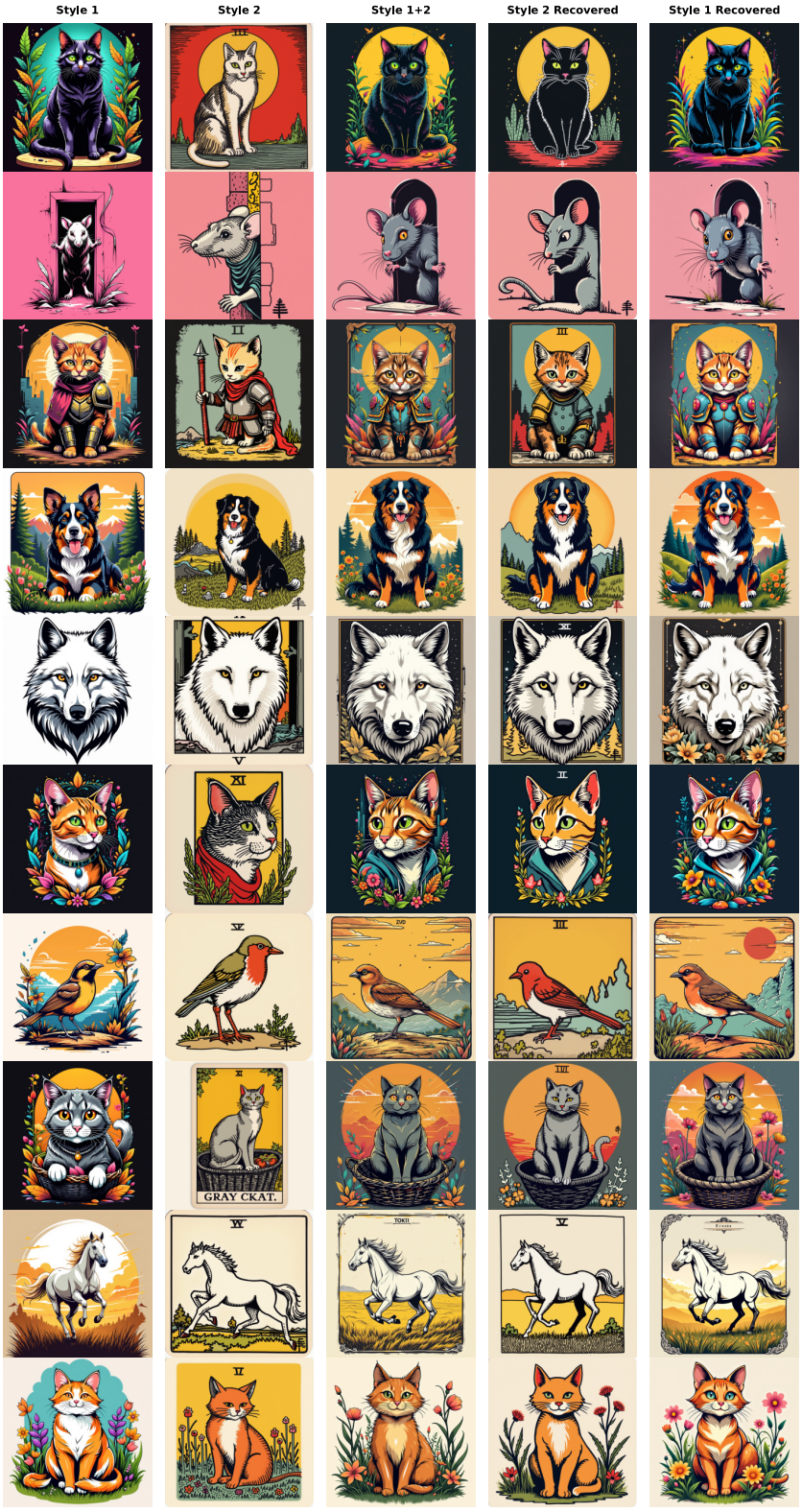

Figure 23: Style coupling and decoupling using parameter-based score interventions with an identity bridge. We sequentially apply two LoRA adaptations at different timesteps, resulting in a composition of both styles (middle). The left columns show individual styles, while the right columns demonstrate decoupling, recovering each style from the combined output.



Figure 24: Style coupling and decoupling using parameter-based score interventions with an identity bridge. We sequentially apply two LoRA adaptations at different timesteps, resulting in a composition of both styles (middle). The left columns show individual styles, while the right columns demonstrate decoupling, recovering each style from the combined output.

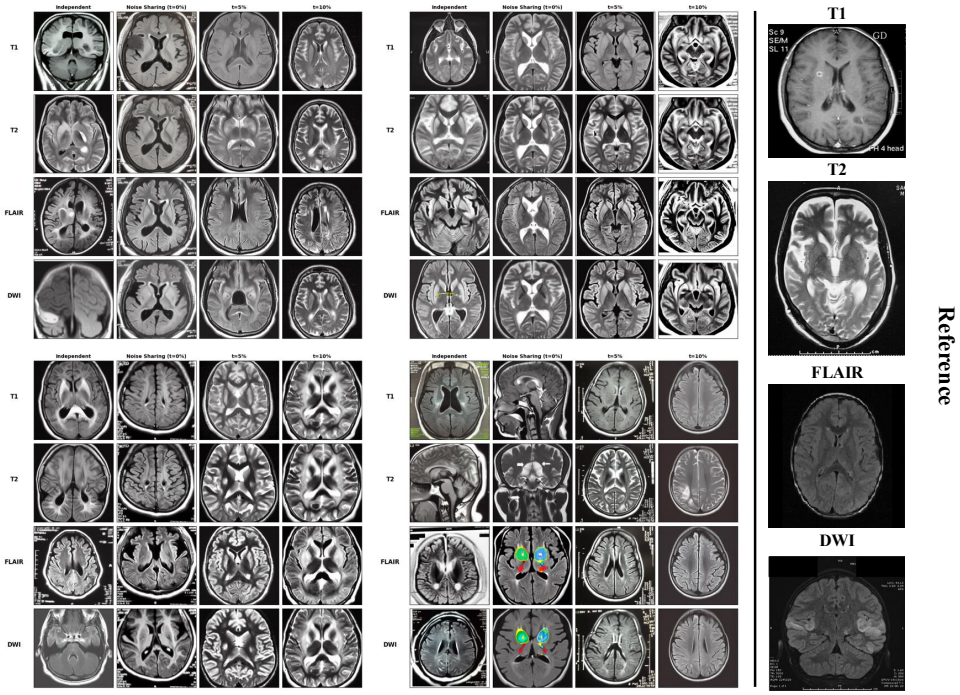

Figure 25: 4 Randomly sampled brain MRIs from SnB. We provide modality reference images from ROCO for those not familiar with MRI contrast modalities. Notice how SnB is both faithful to modality specific feature while maintaining coherence.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We claim advantage over past multimodal methods, and validate them in the experiments section. We achieve better or comparable performance on standard benchmarks, and we extend SnB to downstream applications, which is connected to the flexibility and scalability claims.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We added a limitation section as a part of the conclusion section. We mention multiple limitations of the deisgn of SnB.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: We provide no proofs of our method as our experiments and evaluations were empirical.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We provide experimental detail about our implementation and the backbones, benchmarks, hardware used in both the main paper and appendix B. Additionally, we plan to publish our code upon publication.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: we plan to publish our code upon publication. All of our experiments were on open-source datasets.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We include all training and testing detail in appendix B. We refer to the used benchmarks for more detial.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: It is infeasible for us to calculate statistical significance given the size of models and resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide hardware detail regarding devices used for training and inference in appendix B. Our method is small as our only learnable network is the bridge module.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Our work aligns with the code of ethics. We additionally include a societal impact section in appendix F.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We included a societal impact section in the appendix (appendix F), where we mention the potential benefit of our work. Our experiments also show how it has potential applications in the medical sciences. We mention possible risks also, and how responsible deployment is always needed.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We use pretrained open-source models which already have safeguards.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Every model and benchmark used in this work was properly cited and acknowledged.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: the paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: Our work does not use LLMs.

    Guidelines:

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