FLOWBIND: EFFICIENT ANY-TO-ANY GENERATION WITH BIDIRECTIONAL FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Any-to-any generation seeks to translate between arbitrary subsets of modalities, enabling flexible cross-modal synthesis. Despite recent success, existing flow-based approaches are challenged by its inefficiency, as they require large-scale datasets often with restrictive pairing constraints, incur high computational cost from modeling joint distribution, and multi-stage training pipeline. We propose **FlowBind**, an efficient framework for any-to-any generation. Our approach is distinguished by its simplicity: it learns a shared latent space capturing cross-modal information, with modality-specific invertible flows bridging this latent to each modality. Both components are optimized jointly under a single flow-matching objective, and at inference the invertible flows act as encoders and decoders for direct translation across modalities. By factorizing interactions through the shared latent, FlowBind naturally leverages arbitrary subsets of modalities for training, and achieves competitive generation quality while substantially reducing data requirements and computational cost. Experiments on text, image, and audio demonstrate that FlowBind attains comparable quality while requiring up to 6× fewer parameters and training 10× faster than prior methods.

1 INTRODUCTION

Recent progress in flow-based generative models has delivered state-of-the-art performance in multi-modal generation. By conditioning on a given input modality, these models excel at specialist tasks such as text-to-image (Esser et al., 2024; Labs et al., 2025) or text-to-audio synthesis (Liu et al., 2024; Huang et al., 2023), demonstrating their strength in learning continuous cross-modal transformations. However, these successes are largely confined to fixed input and output mapping, and extending flow models to support true any-to-any generation, where arbitrary subsets of modalities can be generated given any other subsets, remains an open challenge.

Bridging the gap from specialist to generalist flow models introduces fundamental hurdles, primarily due to requirements of multi-modal data and computational cost. Frameworks that rely on a central anchor modality, typically text (Tang et al., 2023), requires each modality to be paired with text during training so that all modalities can be aligned through the shared text representation. This design is restrictive, as it prevents the model from learning the rich, direct correlations that exist beyond language. Conversely, methods that model the full joint conditioning of all modalities (Li et al., 2025b) can achieve expressive generation performance but at a steep cost: they require some fully-paired data for stable training, which is scarce, and their computational complexity often scales quadratically with the number of modalities. These data and compute issue render them impractical for real-world scenarios with a large and diverse set of modalities.

Beyond the computational cost, a significant hurdle for generalist models is the complexity of their training pipelines. Rather than a single, unified process, these frameworks often rely on intricate, multi-stage procedures. These stages separately optimize the encoding components for modality alignment and the decoding components responsible for the model's generative capabilities. This staged approach is evident in prominent models; for instance, CoDi (Tang et al., 2023) employs a multi-stage process that separates modality alignment from joint generation. Similarly, OmniFlow (Li et al., 2025b) requires a distinct post-training phase after merging its core components. Such multi-stage pipelines can be brittle, difficult to optimize, and hinder the development of truly seamless, end-to-end generative models.

We introduce FlowBind, a simple flow-based model that addresses these limitations. FlowBind introduces a learnable shared latent capturing cross-modal commonality, and connects each modality to this latent through its own invertible flow. All components are trained jointly under a single flow-matching objective, while the learned flows enable direct any-to-any translation at inference. Because each flow requires only its modality paired with the latent, the method naturally supports training with partially paired data while reducing computational cost. This design yields a simple, efficient, and data-flexible solution for general-purpose any-to-any generation.

In summary, our main contributions are as follows: **(1)** We introduce a flow-based framework for any-to-any generation that factorizes multi-modal interactions through a learnable shared latent, enabling training from arbitrary paired data with low computation budget. **(2)** Our method jointly optimizes both the shared latent and all modality-specific flows under a single flow-matching loss, avoiding the multi-stage pipelines. **(3)** Experiments on text, image, and audio demonstrate that FlowBind achieves competitive quality with substantially reduced data and computation compared to representative baselines, while flexibly supporting any-to-any translation.

2 Preliminaries

Flow Matching Conditional Flow Matching (Lipman et al., 2023) is a simulation-free framework for learning a continuous transformation between a source distribution p_0 and a target distribution p_1. This transformation is defined by an Ordinary Differential Equation (ODE), $\frac{dz_t}{dt} = v_\theta(z_t, t)$, where a drift network v_θ parametrizes the time-dependent vector field. With linear interpolation path $z_t = (1-t)\,z_0 + t\,z_1$ with $(z_0, z_1) \sim (p_0, p_1)$, the target velocity is simply $z_1 - z_0$, and the objective becomes:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}\left[\|v_\theta(z_t, t) - (z_1 - z_0)\|^2\right]. \tag{1}$$

At the optimum, Eq. 1 yields the conditional expectation of the target velocity:

$$v_\theta^\star(x, t) \;=\; \mathbb{E}[\, z_1 - z_0 \mid z_t = x \,]. \tag{2}$$

Generation is then performed by integrating the learned drifts over time:

$$z_{t_1} = z_{t_0} + \int_{t_0}^{t_1} v_\theta(z_t, t)\, dt = \text{ODESolve}(z_{t_0}, v_\theta, t_0, t_1) \tag{3}$$

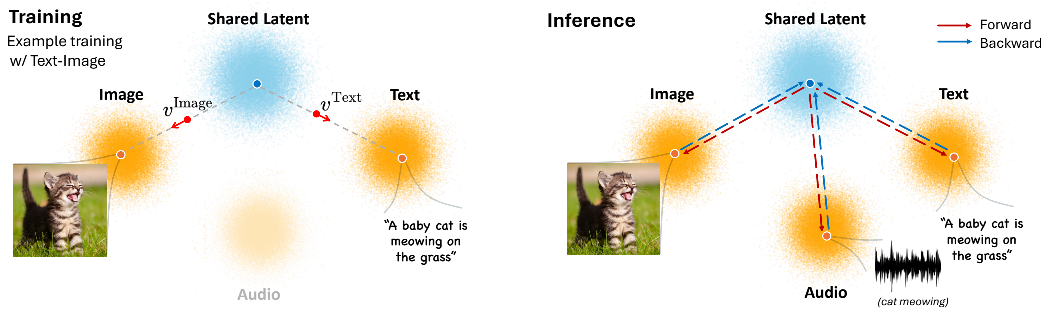
Note that, under standard Lipschitz conditions, the induced flow is invertible *i.e.*, the ODE can be integrated forward or backward in time to induce samples from p_0 or p_1.

Any-to-Any Generative Flows The goal of any-to-any generation is to learn a unified model that can translate between arbitrary subsets of modalities. Given N modalities $\mathbf{z} = (z^1, \ldots, z^N)$, this amounts to modeling their joint distribution $p(\mathbf{z})$ so that for any $S_{\text{in}}, S_{\text{out}} \subseteq \{1, \ldots, N\}$, the model can perform any-to-any generation by sampling from conditional probability $p(\mathbf{z}^{S_{\text{out}}}|\mathbf{z}^{S_{\text{in}}})$.

Existing flow-based approaches address this problem by constructing continuous trajectories that transform i.i.d. Gaussian noise $\mathbf{z}_0 \sim \pi_{\text{prior}}$ into data samples $\mathbf{z}_1 \sim \pi_{\text{data}}$. Representative examples include CoDi (Tang et al., 2023) and OmniFlow (Li et al., 2025b), which mainly differ in how they synchronize trajectories across modalities. CoDi learns modality-specific encoders that align all modalities to a shared text embedding, which then serves as the conditioning signal for per-modality denoising networks $\epsilon^i(z_t^i, t, \hat{c}^{\text{text}})$. In contrast, OmniFlow learns a time-decoupled joint velocity field $v(z_{t_1}^1, \ldots, z_{t_N}^N, t_1, \ldots, t_N)$, where the interpolation path for each modality is explicitly conditioned with the other modalities to ensure alignment.

Despite their empirical success, existing flow-based methods face several limitations. First, they cannot fully leverage arbitrary paired modalities for any-to-any generation: CoDi requires each modality to be paired with text to establish a canonical embedding, while OmniFlow relies heavily on fully paired data for stable training [1]. Second, both methods require multi-stage training: CoDi separately learns the shared representation and denoising networks, whereas OmniFlow pre-trains drift networks for each modality pair before joint training. Finally, they operate in high-dimensional representations, leading to substantial computational cost and slow convergence.

[1] Although partially paired data can be used for training in principle, performance and stability are reported to depend strongly on fully paired data; see Appendix B.2 of Li et al. (2025b).



(a) **Training** both shared latent and drifts.

(b) **Inference** with per-modality drifts.

Figure 1: An overview of FlowBind. (a) During training, we jointly learn the shared latent and per-modality drift networks in a single stage. (b) At inference, the learned drift networks perform flexible any-to-any generation by solving per-modality ODEs forward and backward in time.

3 FlowBind

To address the aforementioned challenges, we propose **FlowBind**, a unified flow-based framework for any-to-any generation. FlowBind is designed to overcome key drawbacks of prior methods: it supports single unified training procedure, operates with lower computational overhead, and fully exploits partially paired data for effective learning.

The key idea of FlowBind is to replace the fixed Gaussian prior with a *learnable, shared* distribution that encapsulates common information across modalities. This acts as a latent anchor, where each modality is connected to it via their own invertible, per-modality flows (Figure 1). With this factorization, FlowBind achieves alignment across modalities naturally via the shared distribution, unlike existing approaches that anchor all modalities to text (Tang et al., 2023) or couples them through a joint drift (Li et al., 2025b). Meanwhile, both the shared distribution and the per-modality flows are learned jointly with only standard flow matching loss using partially paired data.

Formally, consider a subset of multi-modal data $\mathbf{z}^S = \{z^i | i \in S\}$ with $S \subset \{1, \dots, N\}$, which is sampled from a joint distribution $\mathbf{z}^S \sim \pi^S_{\text{data}}$. Assume that there exists a shared latent $z^* \sim \pi^S_{\text{shared}}$ that encompasses the common information of all individual modality in $\mathbf{z}^S$. Then for each $i \in S$, FlowBind learns a straight interpolation path that bridges the data z^i to the shared latent z^* by:

$$z^i_t = tz^i + (1-t)z^* \tag{4}$$

$$\frac{\partial z^i_t}{\partial t} = v^i(z^i_t, t), \tag{5}$$

where v^i denotes the modality-specific velocity field. Note that multi-modal flows are factorized per modality given the shared latent (Eq. 4), and the shared latent implicitly aligns these flows across modality (Eq. 5). During training, the shared latent is instantiated as $z^* = H_\phi(\mathbf{z}^S)$ through an auxiliary encoder H_ϕ, whose marginal approximates π_{shared} and is optimized jointly with the per-modality drift networks v_{θ^i} (Figure 1(a)). At inference, FlowBind relies only on the learned drift networks: owing to the invertibility of the direct flows, both inferring the shared latent from input modalities and generating outputs from the latent are achieved by a single drift network per modality (Figure 1(b)). Details of the training and inference procedures are provided in Section 3.1.

Following prior works (Esser et al., 2024; Liu et al., 2024), FlowBind operates in a compressed latent space obtained by per-modality autoencoders. However, instead of high-dimensional latent, we adopt compact and semantic representations extracted by strong encoders in each modality, paired with decoders that reconstruct modality-specific details from the encoded feature. This design enables FlowBind to focus on shared structure in a low-dimensional space, making cross-modality alignment simpler and training both faster and efficient.

Taken together, FlowBind provides several advantages over existing approaches. By introducing a shared latent space, FlowBind factorizes the multi-modal flow into independent per-modality drifts, allowing them to operate in isolation with reduced computational cost. This factorization also naturally enables training with arbitrary paired modalities: since each drift network learns only to connect its modality to the shared latent, learning does not depend on specific modalities or fully-paired data. Finally, both the shared latent and modality-specific drifts are optimized jointly with a single flow matching objective, avoiding the multi-stage training pipelines of prior works and yielding a simple and efficient framework.

3.1 Training and Inference

Learning Objective During training, the auxiliary encoder H_ϕ and the set of modality-specific drift networks $\{v_{\theta^i}\}_{i=1}^N$ are optimized jointly under the flow matching framework in Eq. 4 and 5. Given a partially paired sample $\mathbf{z}^S$, the auxiliary encoder produces a shared latent $z^* = H_\phi(\mathbf{z}^S)$, and for each modality $i \in S$, the drift network v_{θ^i} is trained to approximate the velocity field along the path between z^i and z^*. This leads to the training objective:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{t, \mathbf{z}^S, z^*} \left[\sum_{i \in S} \left(\left\| v_{\theta^i}(z_t^i, t) - (z^i - z^*) \right\|^2 \right) \right], \tag{6}$$

where $\theta = \{\theta^1, ..., \theta^N\}$. In principle, this couples the two components: drift networks learn to predict the displacement toward each modality endpoint, while the auxiliary encoder is encouraged to provide a shared latent from which every modality can be recovered to aid drift networks.

However, this formulation admits degenerate solutions. For example, if the encoder collapses to a constant output such as $z^* = 0$, the drift networks can trivially fit $v^i(z_t^i, t) = z^i$ with $z_t^i = tz^i$ and achieve the zero loss for $t \in (0, 1]$, leaving the encoder with no meaningful supervision (Kim et al., 2024). The underlying reason is that flow matching enforces transportation between two fixed endpoints but does not itself constrain the distribution of encoder outputs. Prior works on direct flow (Liu et al., 2025; He et al., 2025) address this by adding explicit regularizers, such as contrastive losses on the encoder, but these introduce additional computation, hyperparameters, and scalability bottlenecks especially with increasing number of modalities.

In contrast, we show that both stabilization and meaningful learning of the encoder can be achieved within the flow-matching objective itself. Our approach is simple: for $t \in (0, 1]$, we stop gradients through the auxiliary encoder to stably train the drift networks, while at t=0, the encoder is directly updated together with the drifts. Despite its simplicity, this scheme effectively prevents collapse and provides the encoder with a meaningful learning signal, as we elaborate below.

Analysis on Encoder Objective To understand what the auxiliary encoder learns under our training strategy, we analyze the flow matching loss at t=0. Substituting the Bayes-optimal drift $v^\star(z_t, t)$ (Eq. 2) into Eq. 6 at $t = 0$ gives encoder's effective objective:

$$\mathcal{L}(\phi) = \mathbb{E}\big[\| \, v^\star(z^*, 0) - (z^i - z^*) \, \|_2^2\big] = \mathbb{E}\big[\| \, \mathbb{E}[z^i \mid z^*] - z^i \, \|_2^2\big] = \mathbb{E}\big[\mathrm{Var}(z^i \mid z^*)\big]. \tag{7}$$

This shows that, at t=0, the encoder is explicitly optimized to minimize the conditional variance of each modality given the shared latent. The term $\mathbb{E}\big[\mathrm{Var}(z^i \mid z^*)\big]$, often referred to as the *unexplained variance*, measures how much information about modality i remains outside of z^*. By the law of total variance (Grimmett & Stirzaker, 2001), reducing this quantity equivalently increases the explained variance of z^i by z^*. Since the optimization is carried out jointly across all modalities, the encoder is therefore driven to shape z^* so that it retains predictive information about each modality, ensuring that the shared latent becomes increasingly informative for cross-modal alignment.

More generally, when the drift networks are not optimal, Eq. 6 at t=0 decomposes into unexplained variance and an approximation error of the drifts:

Proposition 1 (Equivalence at $t = 0$) *For any parameters (θ, ϕ) and modality subset S, the flow matching loss (Eq. 6) at $t = 0$ decomposes as follows:*

$$\mathcal{L}(\theta, \phi) = \underbrace{\sum_{i \in S} \mathbb{E}\big[\mathrm{Var}(z^i \mid z^*)\big]}_{\textit{unexplained variance}} + \underbrace{\sum_{i \in S} \mathbb{E}\Big[\big\| \, v_{\theta^i}(z^*, 0) - \mathbb{E}[\, z^i - z^* \mid z^*] \, \big\|^2\Big]}_{\textit{approximation error of } v_\theta^i}.$$

A formal proof is provided in Appendix A.1.

This decomposition reveals that even when the drift networks are imperfect, the auxiliary encoder H_ϕ is consistently driven to minimize unexplained conditional variance while simultaneously optimizing the shared latent to enable better drift approximation of their targets. In this way, our training strategy encourages the auxiliary encoder and the drift networks remain tightly coupled: the drifts learn to predict each modality endpoint from the shared latent ($t \in [0, 1]$), while the encoder is driven to shape the latent into a representation from which all modalities can be reliably recovered ($t = 0$). During training, we balance the drifts and encoder training by sampling from the mixture $t \sim (1 - \alpha)\text{Unif}(0, 1) + \alpha\delta(t = 0)$. The training procedure is given at Algorithm 1.

Inference After training, FlowBind performs versatile any-to-any generation relying solely on the learned per-modality flows, without utilizing the auxiliary encoder. Given a source modality i, we first project it onto the shared latent by integrating its backward flow, and then map the shared latent to the target modality j via the corresponding forward flow:

$$\hat{z}^* = \text{ODESolve}(z^i, v_\theta^i, 1, 0), \qquad \hat{z}^j = \text{ODESolve}(\hat{z}^*, v_\theta^j, 0, 1) \tag{8}$$

When conditioning on multiple source modalities $\mathbf{z}^S$, FlowBind obtains per-modality latent estimates $\hat{z}^{(*,i)}$ by solving the corresponding backward flows independently. These estimates are then aggregated into the shared latent $\hat{z}^*$ by simple averaging. Finally, the target modality is generated by integrating its forward flow starting from $\hat{z}^*$. The inference procedure is given at Algorithm 2.

4 Related Work

Any-to-Any Generation A prominent paradigm for any-to-any generation tokenizes all modalities into a discrete space and trains a single sequence model to predict the unified stream autoregressively. In this setup, a powerful large language model performs cross-modal sequence generation, with tokenized data of all modalities. Some works (Team, 2024) focus on interleaved generation solely on text-image generation, while others (Wu et al., 2024; Zhan et al., 2024) extend to broader multi-modal scenarios including speech (Wang et al., 2024) and even robotics (Lu et al., 2024). Training these models typically involves multi-stage procedures and often instruction finetuning which requires the dataset with detailed textual descriptions. Additionally, these works can be computationally demanding during both training and inference.

Another line of work utilizes discrete diffusion models, often by adapting them to operate on discrete token spaces (Rojas et al., 2025; Shi et al., 2025). These methods, which typically focus on text-image generation tasks, leverage the high-quality synthesis capabilities of diffusion for multimodal scenario. For instance, UniDisc (Swerdlow et al., 2025) highlights the controllability of this approach by framing various conditional generation tasks, such as inpainting.

Direct Flow-based Models Recent flow-based models (Liu et al., 2025; He et al., 2025) have explored learning direct, data-to-data invertible mappings between two modalities, predominantly focusing on text-image pairs. This approach represents a fundamental departure from traditional generative flows that typically learn bridging from fixed prior distributions (e.g., standard Gaussian) to target data distributions through conditional generation mechanisms. To facilitate these direct transformations, existing methodologies designate latent distribution of one modality (i.e., source distribution) as a learnable embedding. This is achieved by introducing an encoder for the source modality and constructing additional loss terms that align the source and target modalities, such as contrastive learning objectives.

While our approach shares foundational ideas with prior work, its emphasis and formulation differ. Existing methods typically rely on multiple loss terms to stabilize training and to optimize endpoint embeddings; in contrast, we employ a single, unified flow objective to achieve the same optimization. Moreover, we pursue direct flows for multi-modality connectivity, whereas most prior efforts have concentrated on two-modality settings—especially text–image generation.

5 Experiments

We conduct an extensive evaluations on any-to-any generation tasks across text, image, and audio modalities. For baselines, we mainly consider previous approaches on flow-based any-to-any generative modeling, namely CoDi (Tang et al., 2023) and OmniFlow (Li et al., 2025b). Qualitative results on various input and output modality combinations are provided in the anonymized website: `https://sites.google.com/view/flowbind`.

Table 1: Comparison of computational cost. #(A-B) indicates the number of training samples for each dataset combination. Training time for CoDi is omitted due to absence of training code and details. For OmniFlow, we report the training time only for the final joint training stage.

Model	Train Param.	GPU-hr	Number of Traning Data				Joint Training
			#(T–I)	#(T–A)	#(I–A)	#(T–A–I)	
CoDi	4.3B	-	400M	3.5M	1.9K	-	NO
OmniFlow	3.2B	480hr*	28M	2.4M	200K	2.2M	NO
FlowBind	568M	48hr	310K	96K	180K	-	YES

Tasks and Evaluation Protocol We consider all six possible one-to-one generation tasks that consist of text, image and audio, and discuss its result in Sec. 5.2. Furthermore, we conduct qualitative analysis under more complex many-to-many generation tasks in Sec. 5.3 to validate cross-modal generation capability of FlowBind. We use an automated metrics for comprehensive evaluation on one-to-one generation, where well-defined quality and alignment metrics are available. Specifically, generation quality is assessed using established modality-specific measures: FID (Heusel et al., 2017) for images, FAD (Kilgour et al., 2019) for audio, and CIDEr (Vedantam et al., 2015) for text captions. Cross-modal alignment is evaluated through pairwise similarity metrics: CLIP scores for text-image pairs (Hessel et al., 2021), CLAP scores for text-audio pairs (Elizalde et al., 2023), and Audio-Image-Similarity (AIS) (Wu et al., 2022) for image-audio pairs. Evaluations are done at held-out test set for image-audio and audio-text tasks, while we employ widely adopted zero-shot benchmark in MS-COCO for text-to-image and image-to-text tasks. Detailed descriptions on datasets and evaluation protocols are provided in Appendix D.

Implementation Details We employ EmbeddingGemma (Team et al., 2025) for textual semantic latent, CLIP (Radford et al., 2021) for visual latent with Stable-UnCLIP (HuggingFace, 2025) as decoder, and CLAP (Elizalde et al., 2023) features for audio synthesis conditioning. Note that these modality-specific encoders and decoders are frozen during the training of FlowBind. We employ MLP-based architecture with residual connections for both auxiliary encoders and drift networks, with AdaLN-zero for time modulation (Peebles & Xie, 2023). More detailed information, including the architectural and training specifications, can be found in Appendix C.1.

5.1 Instantiation of FlowBind

To highlight the claimed efficiency of FlowBind, we instantiate FlowBind as a relatively lightweight model, and also train it on smaller dataset with simple single-stage training. We summarize the details of our instantiation of FlowBind in Table 1, making comparison to previous flow-based any-to-any generation models. Compared to baselines, FlowBind achieves any-to-any generation with considerably less computations and efforts. When comparing the computational cost, FlowBind operates on low-dimensional, compact representation space, yielding a lightweight model with less than 1B trainable parameters. This design choice makes FlowBind to be trained much faster, using about 10× less compute compared to OmniFlow, in terms of GPU-hours. We also use much smaller data compared to baselines (0.15 % of CoDi or 1.79 % of OmniFlow). In subsequent sections, we now demonstrate that our efficient any-to-any generation model can achieve strong cross-modal generation capabilities.

5.2 Results on One-to-One Generation

Effectiveness of FlowBind We demonstrate the effectiveness of FlowBind under all six pairwise one-to-one generation scenarios in Table 2 and Table 3. While the core goal of FlowBind lies on efficient modeling of any-to-any generation, we also observe the resulting model shows strong capability in cross-modal generation tasks. Compared to CoDi and OmniFlow, FlowBind achieves the best quality metrics in all six one-to-one generation tasks, while showing superior alignment score on four tasks among six. We also note that baselines such as OmniFlow are initialized from strong specialist model (*i.e.*, SD3-Medium) that excels at text-image alignment, which explains their particularly good performance on text-to-image alignment scores. As an overall, we conclude that FlowBind shows promising performance on the evaluated one-to-one generation tasks.

Table 2: Fidelity assessment on one-to-one evaluation benchmarks.

Category	**Model**	**T → I**	**I → T**	**T → A**	**A → T**	**I → A**	**A → I**
		FID ↓	**CIDEr ↑**	**FAD ↓**	**CIDEr ↑**	**FAD ↓**	**FID ↓**
Specialists	SD3-Medium	25.40	–	–	–	–	–
	FLUX.1	22.06	–	–	–	–	–
	LLaVA-NeXT	–	109.3	–	–	–	–
	TangoFlux	–	–	1.41	–	–	–
	AudioX	–	–	3.09	–	–	–
	Qwen2-Audio	–	–	–	4.64	–	–
	Seeing & Hearing	–	–	–	–	5.31	–
	Sound2Vision	–	–	–	–	–	42.55
Generalists	UnifiedIO2-L	21.54	134.7*	8.31	12.15	–	–
	CoDi	24.80	16.40	9.84	6.62	14.58	50.4
	OmniFlow	22.97	44.20	4.20	31.79	5.67	106.03
	FlowBind	**17.39**	**46.26**	**4.19**	**55.11**	**2.50**	**26.60**

Table 3: Alignment results on one-to-one evaluation benchmarks.

Category	**Model**	**T → I**	**I → T**	**T → A**	**A → T**	**I → A**	**A → I**
		CLIP ↑	**CLIP ↑**	**CLAP ↑**	**CLAP ↑**	**AIS ↑**	**AIS ↑**
Specialists	SD3-Medium	31.60	–	–	–	–	–
	FLUX.1	31.06	–	–	–	–	–
	LLaVA-NeXT	–	32.14	–	–	–	–
	TangoFlux	–	–	42.71	–	–	–
	AudioX	–	–	29.29	–	–	–
	Qwen2-Audio	–	–	–	17.09	–	–
	Seeing & Hearing	–	–	–	–	75.11	–
	Sound2Vision	–	–	–	–	–	62.39
Generalists	UnifiedIO2-L	30.71	30.73	13.48	18.68	–	–
	CoDi	30.26	26.24	10.79	17.94	61.55	74.26
	OmniFlow	**31.52**	27.71	24.23	**45.08**	71.71	59.22
	FlowBind	28.35	**29.74**	**29.08**	36.70	**82.89**	**78.17**

An interesting observation is that FlowBind exhibits substantial gains in the image-audio generation, where it significantly outperforms among generalists and even dedicated specialist, without making modality-specific adjustments. We conjecture the impressive performance of FlowBind at audio-image correspondence stems from the introduction of learnable shared latent space, which is designed to contain meaningful information about each modality (Section 3.1) and learned directly from audio-image pair. Instead of learning a shared latent space from arbitrarily paired data, CoDi employs an text-anchored design, using only text-paired data during its multimodal alignment stage. This design choice of CoDi makes alignment between non-text modality, such as audio-image alignment, to be indirectly captured with the aid of text. We also note that OmniFlow is also implicitly relying on text representation, given the fact that its weights are initialized from pretrained text-to-image and text-to-audio models at the beginning of second-stage any-to-any training. In contrast, FlowBind can learn a shared latent space directly from given train pair, offering more suitable space for cross-modal generation tasks.

Train Efficiency While showing promising performance compared to previous any-to-any generation models, we emphasize that FlowBind is trained with much less computations and efforts. As previously shown in Table 1, the demonstrated strong performance of FlowBind is achieved using 6 times less training parameters and 10 times less compute compared to OmniFlow, which employs the joint modeling approach. Our formulation of factorizing multimodal flow into per-modality flows result in this efficiency, as it avoids the exponential scaling of parameters and computational load inherent in the joint modeling approach.

Moreover, FlowBind employs a unified training objective, in contrast to prior works that require complex multi-stage training pipelines. Consequently, FlowBind can be trained with less effort

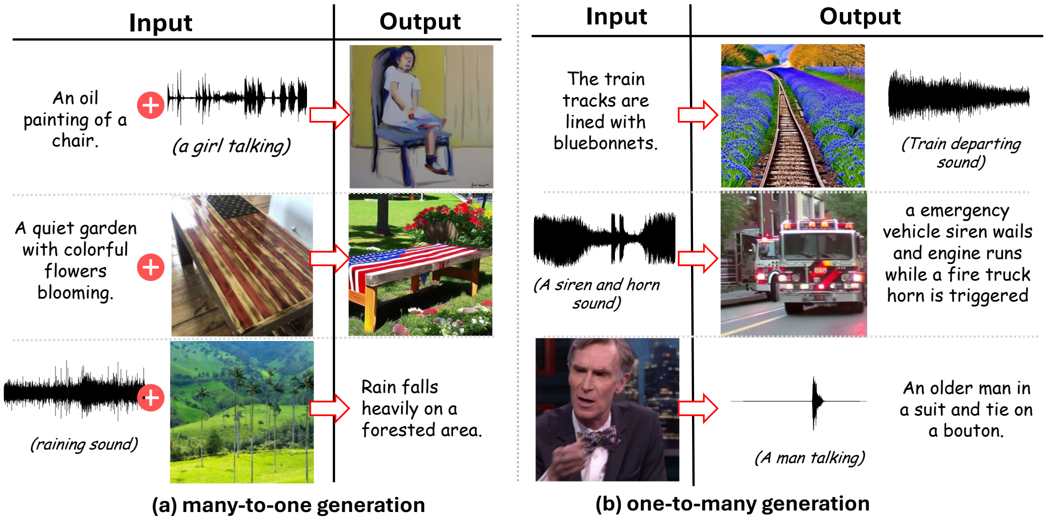

Input
Output
An oil painting of a chair.
(a girl talking)
A quiet garden with colorful flowers blooming.
(raining sound)
Rain falls heavily on a forested area.
(a) many-to-one generation
Input
Output
The train tracks are lined with bluebonnets.
(Train departing sound)
(A siren and horn sound)
a emergency vehicle siren wails and engine runs while a fire truck horn is triggered
(A man talking)
An older man in a suit and tie on a bouton.
(b) one-to-many generation

Figure 2: Qualitative results on various many-to-many generation tasks. More results and comparisons with baselines are presented in Appendix E.

without cumbersome hyperparameters and additional computations that emerge from more complex training procedure.

Data Efficiency In terms of data efficiency, FlowBind is able to achieve any-to-any generation with much smaller training dataset, using 0.15 % of CoDi and 1.79 % of OmniFlow. We conjecture the training can be done with much smaller dataset because we choose to model flow between low-dimensional representations. By doing so, the cross-modal generation capability is decomposed into inter-modal alignment and intra-modal generation in FlowBind. Our drift network is only required to capture inter-modal correspondence, as per-modality frozen encoder-decoders take charge in capturing intra-modal generative capability. This would enable FlowBind to quickly capture cross-modal alignment with fewer datasets.

5.3 Results on Many-to-Many Generation

Beyond the extensive evaluation on one-to-one generation, we conduct a qualitative analysis to assess FlowBind's capability as an *any-to-any* generation model, on more complex cross-modal generation tasks. As shown in Figure 2, FlowBind is capable of handling complex cross-modal generation tasks, faithfully reflecting the input conditions in its outputs. Interestingly, we see even some detailed components (Stars and Stripes printed on table) in input data appear again in output modalities, as shown in the second row of Figure 2(a). This highlights the expressiveness of learned shared space in FlowBind for cross-modal generation tasks, which enables aggregating necessary information from multiple input conditions by averaging on latent. More qualitative examples are presented in Appendix. E and the anonymized website: `https://sites.google.com/view/flowbind`.

6 Analysis

Fixed v.s. Learnable Shared Anchor Theoretical analysis in Section 3.1 implies that our training objective yields a meaning shared anchor space. To further support this claim, we conduct an empirical comparison between having text modality as a fixed anchor and having learnable, shared latent space as an anchor. Similar to the alignment procedure in CoDi, we consider a text-anchoring baseline that directly utilizes text modality as a fixed anchor. Since the image-audio

Table 4: Comparison of alignment scores between model that uses fixed text anchor and learnable shared anchor. I-A represents the image-audio dataset.

Model	I → T	A → T	I → A
Text-anchoring	27.94	36.72	55.48
FlowBind w/o I-A	**30.04**	**37.04**	**61.88**

pair cannot be used in this setting, we compare text-anchoring baseline with a variant of FlowBind that excludes image-audio pair during training. The resulting data-controlled comparison, as reported in Table 4, shows that cross-modal alignment can be improved by introducing learned shared latent space. Specifically, FlowBind variant trained without image-audio pair still outperforms text-anchoring variant in all three measured alignment scores. This suggests that employing learnable shared latent space can be beneficial for cross-modal alignment in general, validating our proposed objective in Equation. 7

Results in Table 4 demonstrate that FlowBind shows a better alignment than text-anchor variant for image-to-audio task, which does not include the text modality. For this task, FlowBind benefits from its relaxed data restriction, effectively modeling image-audio correspondence by directly learning from paired data.

Analysis on Shared Latent As mentioned in Section 3.1, our learning objective is designed to produce a shared latent representation that unifies information from all input modalities. We analyze the characteristics of the learned space, hypothesizing that it should exhibit strong cross-modal alignment. To quantitatively evaluate the alignment of shared latent representations across modalities, we measure the CKNNA metric introduced in Huh et al. (2024), comparing cross-modal alignment in shared latent space and in per-modality encoder space. We follow the suggested procedure for computing CKNNA measure, using at most 1024 samples with neighborhood size k set to 10. The analysis is done for text-audio and audio-image alignment, the settings where held-out test set is available.

Table 5: Shared latent space yields higher alignment measured in CKNNA.

Model	T-A	A-I
Latent	0.1965	0.1343
Shared Latent	**0.2872**	**0.3026**

As shown in Table 5, our learned representation exhibits higher alignment scores with the learnable shared latents, compared to the latents that are obtained from per-modality encoders. This quantitatively measured improvement in alignment validates our claim that shared latent space is not merely for co-embedding features, but is for building a truly shared semantic space. Our framework successfully learns a coherent, well-aligned latent space that bridges the semantic gap between different modalities, thereby handling complex any-to-any generation task effectively.

In addition to the quantitative analysis, we conduct a qualitative analysis via exploration of shared latent space by interpolation between two latents. As shown in Figure 3, we empirically observe that the shared latent space is indeed a well-aligned, semantically meaningful space, enabling the semantic of decoded image change gradually between two input images.

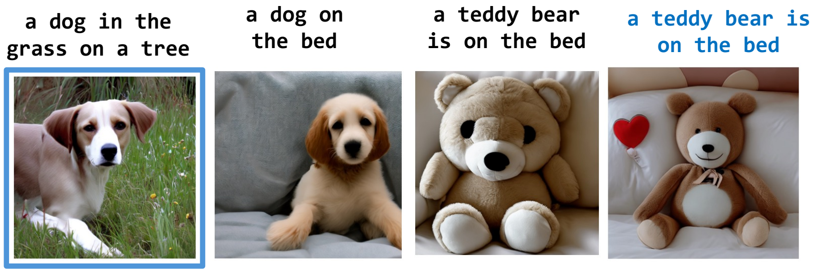

a dog in the grass on a tree
a dog on the bed
a teddy bear is on the bed
a teddy bear is on the bed

Figure 3: FlowBind's shared latent space learn semantically meaningful space, allowing smooth transition when interpolating between two latents. Data with blue boundary indicates input.

7 Conclusion

In this work, we introduce a novel framework for any-to-any multi-modal generation that directly addresses the critical limitations of data scarcity and computational complexity inherent in prior methods. By learning from arbitrarily paired data, our model alleviates the need for impractical fully-paired or anchor-based datasets. The core of our approach is a shared latent space trained end-to-end with a single, unified flow matching objective. This design not only simplifies the training pipeline but also yields a computationally efficient and highly scalable system. Our experiments demonstrate that this approach achieves competitive performance, particularly in non-text-anchored tasks, and learns a well-structured, semantically aligned latent space. We claim this data-flexible and efficient framework represents a significant step towards building generalist generative models.

Ethics Statement

We have carefully reviewed the Code of Ethics and confirm that we adhere to the principles. To the best of our knowledge, this work raises no ethical concerns.

Reproducibility Statement

We have made our best efforts to ensure the reproducibility of our experiments. We will release the code in public when published, to enable others to replicate our results. For dataset, exact lists/splits and preprocessing scripts for the datasets will be included in future code release. Regarding experiment details, we believe the appendix provides comprehensive details of our method, including algorithm pseudo-code (Algorithm B) and implementation details (Appendix C), and descriptions of all datasets involved (Appendix C.3).

References

Burak Can Biner, Farrin Marouf Sofian, Umur Berkay Karakas, Duygu Ceylan, Erkut Erdem, and Aykut Erdem. Sonicdiffusion: Audio-driven image generation and editing with pretrained diffusion models. *CoRR*, abs/2405.00878, 2024.

Black Forest Labs. Announcing Black Forest Labs. Blog post on Black Forest Labs website, August 2024. URL `https://blackforestlabs.ai/announcing-black-forest-labs/`. Accessed: 2025-05-14.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

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

LAION eV. laion/relaion-coco, 2025.

Geoffrey Grimmett and David Stirzaker. *Probability and Random Processes*. 2001.

Ju He, Qihang Yu, Qihao Liu, and Liang-Chieh Chen. Flowtok: Flowing seamlessly across text and image tokens. *CoRR*, abs/2503.10772, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *NeurIPS*, 2017.

Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *CoRR*, abs/2305.18474, 2023.

HuggingFace. Stable unclip documentation, 2025. URL `https://huggingface.co/docs/diffusers/api/pipelines/stable_unclip`. Accessed: 2025-09-25.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *ICML*, 2024.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *CoRR*, abs/2412.21037, 2024.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In Gernot Kubin and Zdravko Kacic (eds.), *20th Annual Conference of the International Speech Communication Association, Interspeech*, 2019.

Semin Kim, Jaehoon Yoo, Jinwoo Kim, Yeonwoo Cha, Saehoon Kim, and Seunghoon Hong. Simulation-free training of neural odes on paired data. In *NeurIPS*, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 kontext: Flow matching for in-context image generation and editing in latent space. *CoRR*, abs/2506.15742, 2025.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Omniflow: Any-to-any generation with multi-modal rectified flows. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 13178–13188. Computer Vision Foundation / IEEE, 2025b.

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

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

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE ACM Trans. Audio Speech Lang. Process.*, 2024.

Qihao Liu, Xi Yin, Alan L. Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A noise-free framework for cross-modality evolution. In *CVPR*, 2025.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *CVPR*, 2024.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

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

Kevin Rojas, Yuchen Zhu, Sichen Zhu, Felix X.-F. Ye, and Molei Tao. Diffuse everything: Multimodal diffusion models on arbitrary state spaces. *CoRR*, abs/2506.07903, 2025.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, and Shuicheng Yan. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *CoRR*, abs/2505.23606, 2025.

Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2974–2983, 2018.

Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, and Tae-Hyun Oh. Sound2vision: Generating diverse visuals from audio through cross-modal latent alignment. *arXiv preprint arXiv:2412.06209*, 2024.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. abs/2503.20853, 2025.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *NeurIPS*, 2023.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *CoRR*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report. *CoRR*, abs/2503.19786, 2025.

Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox: Diffusion transformer for anything-to-audio generation. abs/2503.10522, 2025.

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

Zekun Wang, King Zhu, Chunpu Xu, Wangchunshu Zhou, Jiaheng Liu, Yibo Zhang, Jiashuo Wang, Ning Shi, Siyu Li, Yizhi Li, Haoran Que, Zhaoxiang Zhang, Yuanxing Zhang, Ge Zhang, Ke Xu, Jie Fu, and Wenhao Huang. MIO: A foundation model on multimodal tokens. *CoRR*, abs/2409.17692, 2024.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal LLM. In *ICML*, 2024.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4541–4550, 2019.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. Anygpt: Unified multimodal LLM with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

A Proofs and Justifications

A.1 Expected Conditional Variance

Setup. Let $N \in \mathbb{N}$ be the number of modalities and define the shared latent $X := z^* = H_\phi(z^1, \ldots, z^N) \in \mathbb{R}^{dX}$. Fix $i \in \{1, \ldots, N\}$ and set

$$Y := z^i - z^* \in \mathbb{R}^d, \qquad f(X) := v_\theta^i(X, 0) \in \mathbb{R}^d, \qquad m(X) := \mathbb{E}[Y \mid X] = \mathbb{E}[z^i - z^* \mid z^*].$$

Assume square–integrability: $\mathbb{E}\|Y\|_2^2 < \infty$ and $\mathbb{E}\|f(X)\|_2^2 < \infty$. (For vectors, $\mathrm{Var}(Z \mid X) := \mathrm{tr}\,\mathrm{Cov}(Z \mid X)$.)

Objective at t=0.

$$L_i(\theta, \phi) \;=\; \mathbb{E}\big[\|f(X) - Y\|_2^2\big]\,.$$

Decomposition with orthogonality. Add and subtract $m(X)$ and expand:

$$\|f(X) - Y\|_2^2 = \|f(X) - m(X)\|_2^2 + \|m(X) - Y\|_2^2 + 2\langle f(X) - m(X),\, m(X) - Y\rangle.$$

Taking expectations and conditioning on X,

$$\mathbb{E}[\langle f(X) - m(X),\, m(X) - Y\rangle] = \mathbb{E}\big[\big\langle f(X) - m(X),\, \mathbb{E}[m(X) - Y \mid X]\big\rangle\big] = 0,$$

since $\mathbb{E}[m(X) - Y \mid X] = m(X) - \mathbb{E}[Y \mid X] = 0$. Equivalently,

$$m(X) - Y \;\perp\; L^2(\sigma(X)) \quad \text{and} \quad f(X) - m(X) \in L^2(\sigma(X)).$$

Thus,

$$\begin{aligned} L_i(\theta, \phi) &= \mathbb{E}\big[\|Y - m(X)\|_2^2\big] + \mathbb{E}\big[\|f(X) - m(X)\|_2^2\big] \\ &= \underbrace{\mathbb{E}\big[\mathrm{Var}(z^i \mid z^*)\big]}_{\text{unexplained variance}} \;+\; \underbrace{\mathbb{E}\big[\|\, v_\theta^i(z^*, 0) - \mathbb{E}[z^i - z^* \mid z^*]\,\|_2^2\big]}_{\text{distance to Bayes}}\,. \end{aligned}$$

Consequently,

$$\min_\theta L_i(\theta, \phi) = \mathbb{E}\big[\mathrm{Var}(z^i \mid z^*)\big]\,, \qquad \text{attained by} \quad {v_\theta^i}^\star(z^*, 0) = \mathbb{E}[z^i \mid z^*] - z^*.$$

Summed objective. For $S \subseteq \{1, \ldots, N\}$, define

$$L_{t=0}(\theta, \phi) := \mathbb{E}\left[\sum_{i \in S} \|\, v_\theta^i(z^*, 0) - (z^i - z^*)\,\|_2^2\right].$$

Summing the above identity over $i \in S$ and using linearity of expectation,

$$L_{t=0}(\theta, \phi) = \sum_{i \in S} \mathbb{E}\big[\mathrm{Var}(z^i \mid z^*)\big] + \sum_{i \in S} \mathbb{E}\big[\|\, v_\theta^i(z^*, 0) - \mathbb{E}[z^i - z^* \mid z^*]\,\|_2^2\big]\,,$$

hence

$$\min_\theta L_{t=0}(\theta, \phi) = \sum_{i \in S} \mathbb{E}\big[\mathrm{Var}(z^i \mid z^*)\big]\,, \quad \text{with } {v_\theta^i}^\star(z^*, 0) = \mathbb{E}[z^i \mid z^*] - z^* \;\; \forall i \in S.$$

Implication. By the Law of Total Variance, $\mathrm{Var}(z^i) = \mathbb{E}[\mathrm{Var}(z^i \mid z^*)] + \mathrm{Var}(\mathbb{E}[z^i \mid z^*])$, so minimizing the unexplained part $\mathbb{E}[\mathrm{Var}(z^i \mid z^*)]$ equivalently maximizes the explained variance $\mathrm{Var}(\mathbb{E}[z^i \mid z^*])$. Because the decomposition holds for any (θ, ϕ), gradients w.r.t. ϕ (the encoder) continually act to reduce the summed unexplained variance across modalities.

A.2 Discussion on Law of Total Variance

Recall. From the t=0 decomposition in the previous section, for each $i \in S$ we have

$$L_i(\theta, \phi) = \underbrace{\mathbb{E}\big[\mathrm{Var}(z^i \mid z^*)\big]}_{\text{unexplained}} + \underbrace{\mathbb{E}\big[\|\, v_\theta^i(z^*, 0) - \mathbb{E}[z^i - z^* \mid z^*]\,\|_2^2\big]}_{\text{distance to Bayes}}\,.$$

The first term depends only on the encoder via z^*.

Law of Total Variance our t=0 formulation

$$L_i(\theta,\phi) = \underbrace{\mathbb{E}\big[\mathrm{Var}(z^i \mid z^*)\big]}_{\text{unexplained}} + \underbrace{\mathbb{E}\Big[\big\|\, v^i_\theta(z^*,0) - \mathbb{E}[z^i - z^* \mid z^*] \,\big\|_2^2\Big]}_{\text{distance to Bayes}},$$

there are concrete benefits to reducing it:

By the law of total variance,

$$\mathrm{Var}(z^i) \;=\; \mathbb{E}\big[\mathrm{Var}(z^i \mid z^*)\big] \;+\; \mathrm{Var}\big(\mathbb{E}[z^i \mid z^*]\big). \tag{9}$$

Since $\mathrm{Var}(z^i)$ is fixed, any reduction of the unexplained term $\mathbb{E}[\mathrm{Var}(z^i \mid z^*)]$ necessarily increases the explained term $\mathrm{Var}(\mathbb{E}[z^i \mid z^*])$. Equivalently, a larger fraction of the variability of z^i is captured through the *same* shared latent z^*. In the multimodal setting, applying equation 9 to each i concentrates cross-modal structure in z^* and thereby promotes alignment across modalities.

Consequences for Latent Design and Alignment Aggregating over $i \in S$, minimizing $\sum_{i\in S} \mathbb{E}[\mathrm{Var}(z^i \mid z^*)]$ compels the shared latent z^* to encode information that is jointly predictive for all modalities, which in turn increases each $\mathrm{Var}(\mathbb{E}[z^i \mid z^*])$ through the *same* bottleneck. At t=0, writing $f^i_\theta(z^*) := v^i_\theta(z^*,0)$ and $m_i(z^*) := \mathbb{E}[z^i - z^* \mid z^*]$, this strategy simultaneously (i) drives $f^i_\theta(z^*)$ toward the Bayes target $m_i(z^*)$ and (ii) reallocates variability from unexplained to explained, yielding an aligned latent space that strengthens downstream predictors for all $i \in S$.

B Training and Inference

This section presents the detailed training and inference algorithms to provide a clear understanding of each procedural formally.

Algorithm 1: Training

```
Input  : Minibatch $\{\mathbf{z}^{S_b}\}_{b=1}^{B}$;
Aux encoder $H_\phi$;
Flows $\{v^i_\theta\}_{i=1}^{N}$ (params $\theta$);
Time sampler $t \sim p(t)$.
Output: Loss $\mathcal{L}$
for each step do
    Sample $\{\mathbf{z}^{S_b}\}_{b=1}^{B}$;
    for $b = 1$ to $B$ do
        $z^*_b \leftarrow H_\phi(\mathbf{z}^{S_b})$
    Draw $t_b \sim p(t)$ for $b = 1, \dots, B$;
    $\mathcal{L} \leftarrow 0,\ M \leftarrow 0$ for $b = 1$ to $B$ do
        for each $i \in S_b$ do
            $z_t \leftarrow (1 - t_b) z^*_b + t_b z^i_b$;
            $\hat{u} \leftarrow v^i_\theta(z_t, t_b)$;
            $u^\star \leftarrow z^i_b - z^*_b$;
            $\mathcal{L} \leftarrow \mathcal{L} + \|\hat{u} - u^\star\|_2^2$;
            $M \leftarrow M + 1$;
    if $M > 0$ then
        $\mathcal{L} \leftarrow \mathcal{L}/M$;
    return $\mathcal{L}$
```

Algorithm 2: Inference

```
Input  : Sources $S$ with $\{z^i\}_{i\in S}$; target $j$;
Learned flows $\{v^i_\theta\}_{i=1}^{N}$; ODESolve.
Output: $\hat{z}^j$.
// Encode sources to shared
   latent ($t : 1 \to 0$)
for each $i \in S$ do
    $\hat{z}^{*,i} \leftarrow$ ODESolve$(z^i,\ v^i_\theta,\ 1,\ 0)$;
$\hat{z}^* \leftarrow \frac{1}{|S|} \sum_{i\in S} \hat{z}^{*,i}$;
// Decode to target ($t : 0 \to 1$)
$\hat{z}^j \leftarrow$ ODESolve$(\hat{z}^*,\ v^j_\theta,\ 0,\ 1)$;
return $\hat{z}^j$
```

C Implementation Details

C.1 Encoders and Decoders for Each Modality

For image, we use CLIP (Radford et al., 2021) for visual latent with Stable-UnCLIP (HuggingFace, 2025) as decoder. For audio, we use CLAP (Elizalde et al., 2023) features for conditioning on

AudioLDM (Liu et al., 2023). For text, we find that existing text autoencoders such as Optimus (Li et al., 2020) has limited reconstruction abilities. Therefore, we use EmbeddingGemma (Team et al., 2025) for text encoder, and train its decoder with simple reconstruction objective. We use pretrained Gemma3-1B (Team et al., 2025) for initialization and finetune it on two epochs of all texts used in Table 6. Note that these modality-specific encoders and decoders are frozen during the training of FlowBind, thereby not counted as a trained parameters when reporting trainable parameters.

C.2 Architecture

We employ Multi-Layer Perceptron (MLP) for both the flow models $\{v_{\theta^i}\}_{i=1}^N$ and the joint estimator H_ϕ. AdaLN (Peebles & Xie, 2023) is applied to all drift networks for better time modulation. For the auxiliary encoder, each modality input is processed by lightweight modality-specific modules (modality-specific parameters), which are subsequently averaged across all output. To enhance training robustness, we incorporate a fixed variance term as a hyperparameter that regularizes the learned representations.

C.3 Training Dataset

We employ all three types of paired data across text, image and audio. We do not use triple data in our experiments. We summarize the details about training dataset in Table 6.

Table 6: Dataset summary.

Type	**Dataset name**	**Size**	**Description**
Text–Image	LAION-COCO	242k	available subset of eV (2025), filtered by aesthetic scores > 5.0. Captions are synthetically generated.
	Flickr-30k	30k	Sentence-based image descriptions.
Text–Audio	AudioCaps v2	91k	Natural language description audio captioning dataset that is parsed from Youtube.
Audio–Image	VGGSound	184k	Large-scale dataset from YouTube

C.4 Training Recipe

We trained the model for 200k iterations using the Adam optimizer and a global batch size of 1024. The total training process requires approximately 48 GPU-hours on NVIDIA H100. To train each drift network, we normalized the latent representations of each modality to match their respective scales. During training, we follow Kim et al. (2024) to randomly apply the velocity prediction objective at the endpoint of the flow (*i.e.*, $t = 1$) which empirically improves training stability.

D Evaluation Setup

Our experiment was done by below benchmarks.

	Eval Dataset	**Speicalists**
Text-to-Image	MSCOCO-30k	Stable-Diffusion3 (Esser et al., 2024) FLUX.1 (Black Forest Labs, 2024)
Image-to-Text	COCO Kaparthy	Llava-Next (Li et al., 2025a)
Text-to-Audio Audio-to-Text	AudioCaps Test set	TangoFlux (Hung et al., 2024) AudioX (Tian et al., 2025) Qwen2-Audio (Mei et al., 2024)
Image-to-Audio Audio-to-Image	VGGSound Test set	Seeing&Heering (Xing et al., 2024) Sound2Vision (Sung-Bin et al., 2024)

Audio-Image-Similarity (AIS) We followed SonicDiffusion (Biner et al., 2024) to measure relative AIS on audio-image evaluations. In contrast to other alignment metrics, AIS is a reference-based metric to compensate different scales of measured cosine similarity. For audio-to-image evaluation, AIS is defined as a ratio of audios in testset that achieve worse cosine similarity than conditioning audio, measured with genrated image. For example, AIS is zero if the generated image is not aligned at all and thus shows least cosine similarity among test set audios. The similarity is measured using wav2clip (Wu et al., 2022) audio embedding and image CLIP (Radford et al., 2021) embedding from ViT-B/32 model. We generalize AIS metric for image-to-audio generation in a symmetric way, counting the ratio of images in test set that gives lower cosine similarity compared to conditioning image.

E Qualitative Results

In this section, we present more qualitative results on various any-to-any generation, includig one-to-one (Figure 4 and 5), one-to-many (Figure 6 and 7), and many-to-one (Figure 8 and 10) generation. Qualitative results on various input and output modality combinations are provided in the anonymized website: `https://sites.google.com/view/flowbind`.

It shows that FlowBind faithfully translate input modalities into another modalities while preserving content. Compared to baseline, FlowBind exhibits stronger qualitative results especially on challenging many-to-one generation tasks. Specifically, we observe that the baselines struggle to preserve the heterogeneous contents of different modalities, often failing to produce content of one of two modalities. Compared to this, FlowBind faithfully generates outputs that preserve the content of all input modalities, showcasing the advantage of FlowBind in any-to-any geneation tasks.

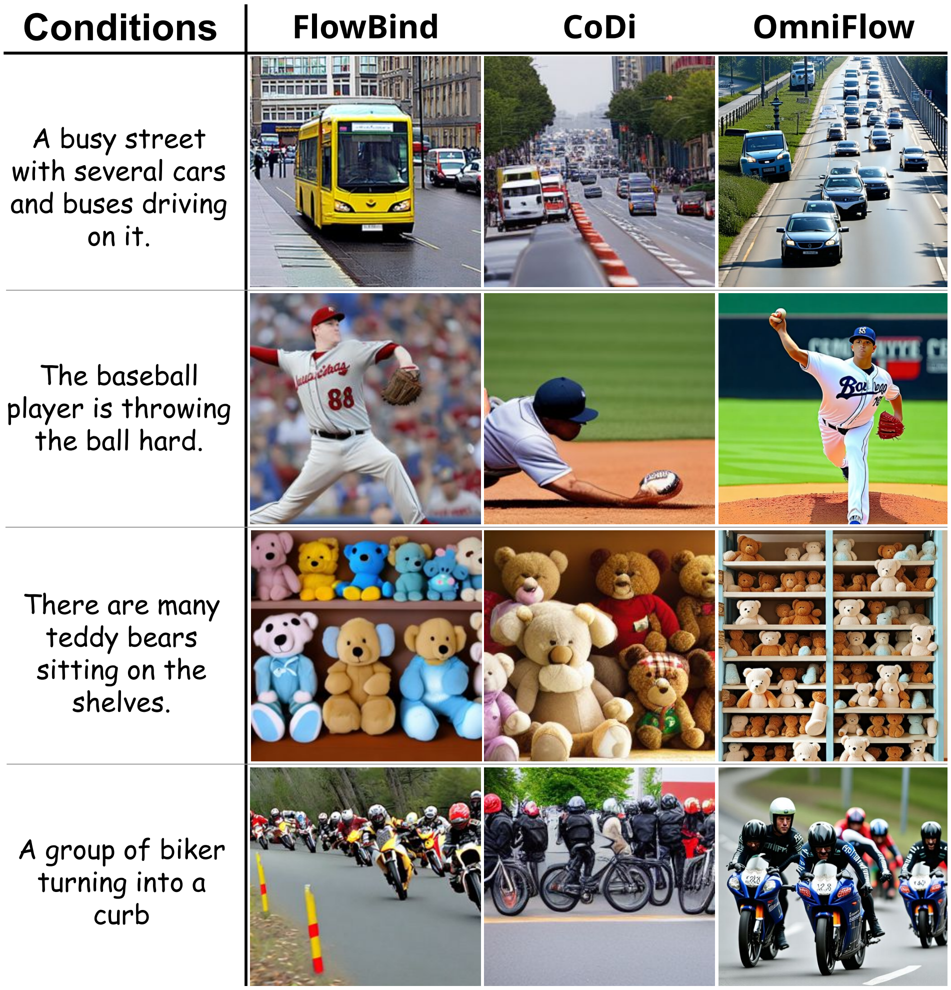

Conditions
FlowBind
CoDi
OmniFlow
A busy street with several cars and buses driving on it.
The baseball player is throwing the ball hard.
There are many teddy bears sitting on the shelves.
A group of biker turning into a curb

Figure 4: Results on text-to-image generation.

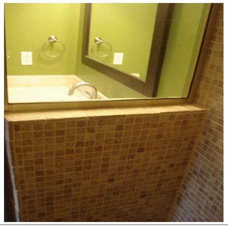
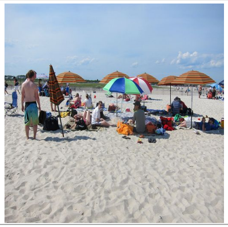
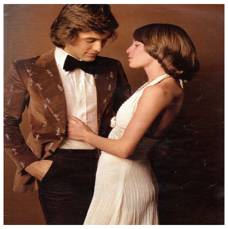
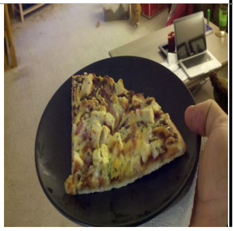
Conditions	FlowBind	CoDi	OmniFlow
	A bathroom with tile and sink, and a mirror above the wall.	new bathroom chairs get messy so bathroom rooms have bathroom colors	a bathroom with a wall-mounted garden tray
	People at the beach in the sand are gathered on a sunny day.	three men are beach at the beach home	a group of people and children sitting at a beach with several trucks parked nearby
	A man and woman in formal attire standing next to each other.	couple and dress wear an opera and an attractive man	a man and woman standing together while he is kissing her
	A plate with a pizza and food on top.	pizza and tomato, turkey and pizza	a plate of fires and a piece of chicken on plate

Figure 5: Results on image-to-text generation.

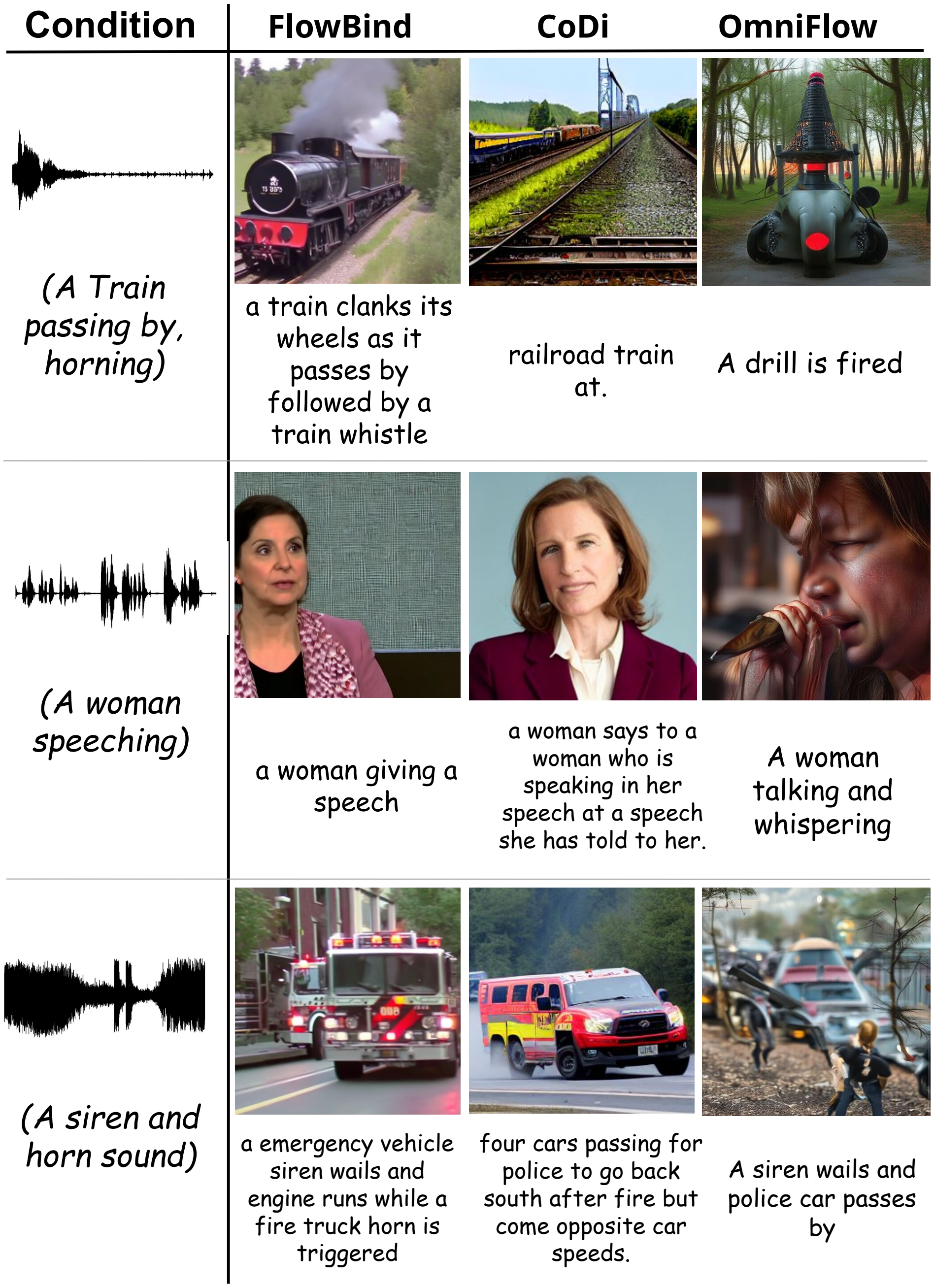

Condition
FlowBind
CoDi
OmniFlow
(A Train passing by, horning)
a train clanks its wheels as it passes by followed by a train whistle
railroad train at.
A drill is fired
(A woman speeching)
a woman giving a speech
a woman says to a woman who is speaking in her speech at a speech she has told to her.
A woman talking and whispering
(A siren and horn sound)
a emergency vehicle siren wails and engine runs while a fire truck horn is triggered
four cars passing for police to go back south after fire but come opposite car speeds.
A siren wails and police car passes by

Figure 6: Results on audio-to-{text, image} generation.

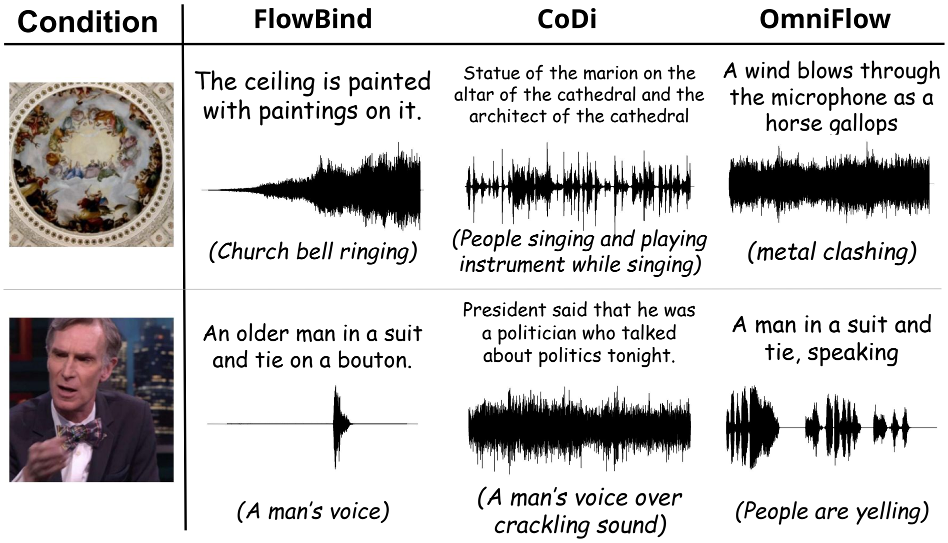

Condition
FlowBind
CoDi
OmniFlow
The ceiling is painted with paintings on it.
Statue of the marion on the altar of the cathedral and the architect of the cathedral
A wind blows through the microphone as a horse gallops
(Church bell ringing)
(People singing and playing instrument while singing)
(metal clashing)
An older man in a suit and tie on a bouton.
President said that he was a politician who talked about politics tonight.
A man in a suit and tie, speaking
(A man's voice)
(A man's voice over crackling sound)
(People are yelling)

Figure 7: Results on image-to-{text+audio} generation.

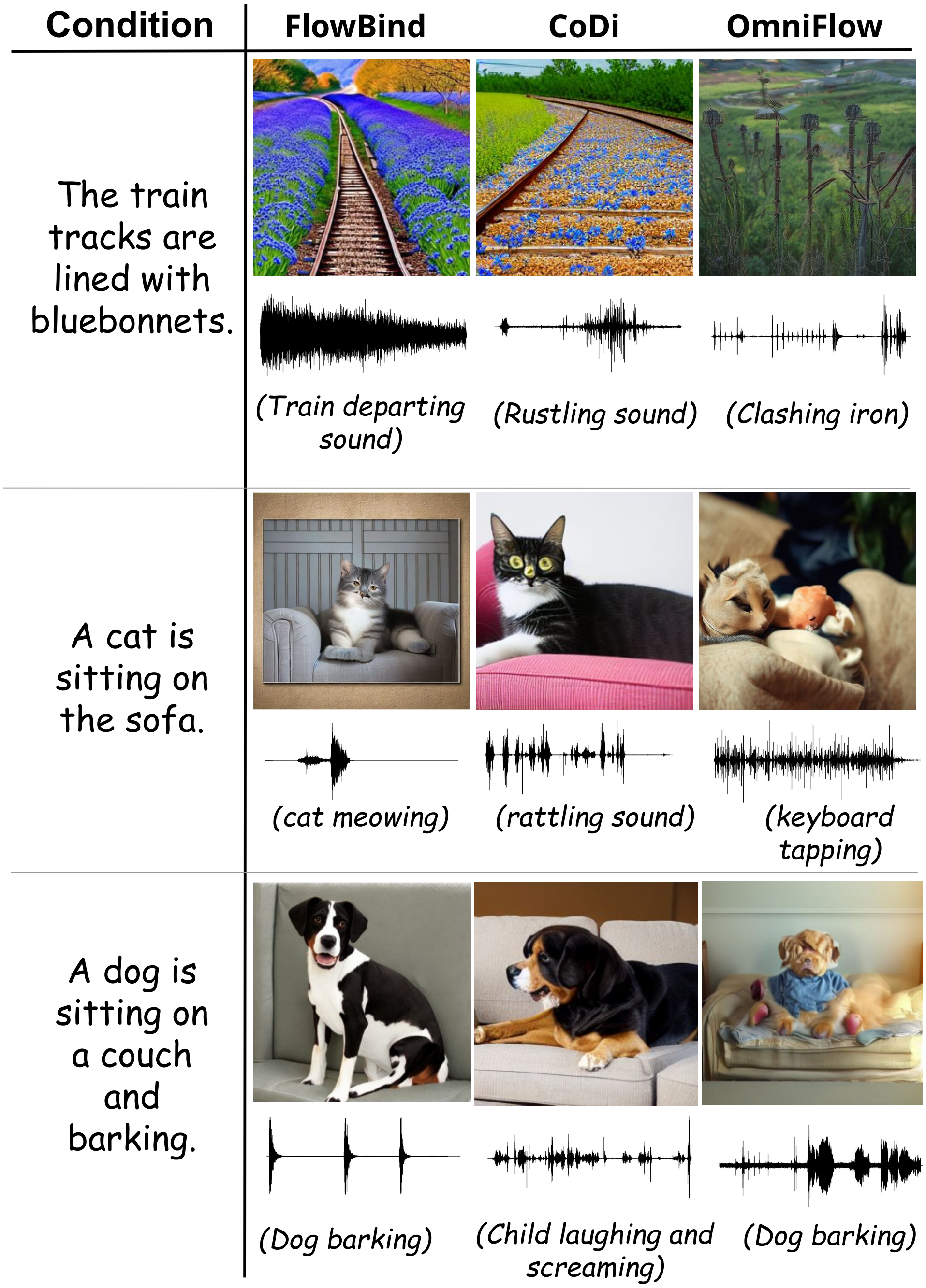


Figure 8: Results on text-to-{image+audio} generation.

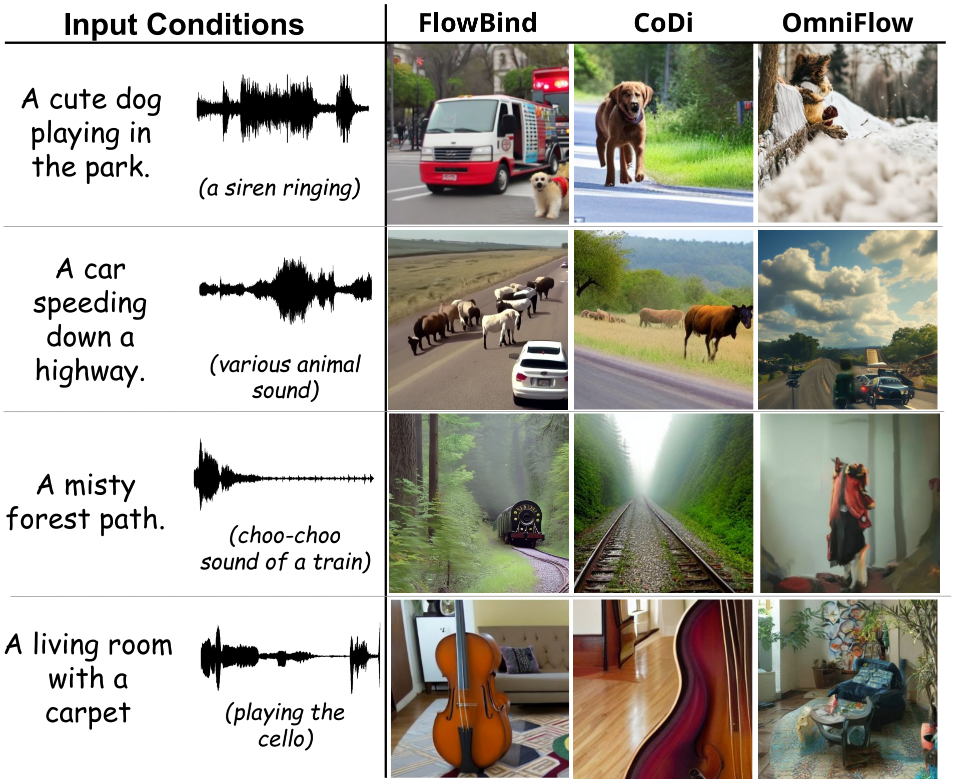

Input Conditions
FlowBind
CoDi
OmniFlow
A cute dog playing in the park.
(a siren ringing)
A car speeding down a highway.
(various animal sound)
A misty forest path.
(choo-choo sound of a train)
A living room with a carpet
(playing the cello)

Figure 9: Results on {text+audio}-to-image generation.

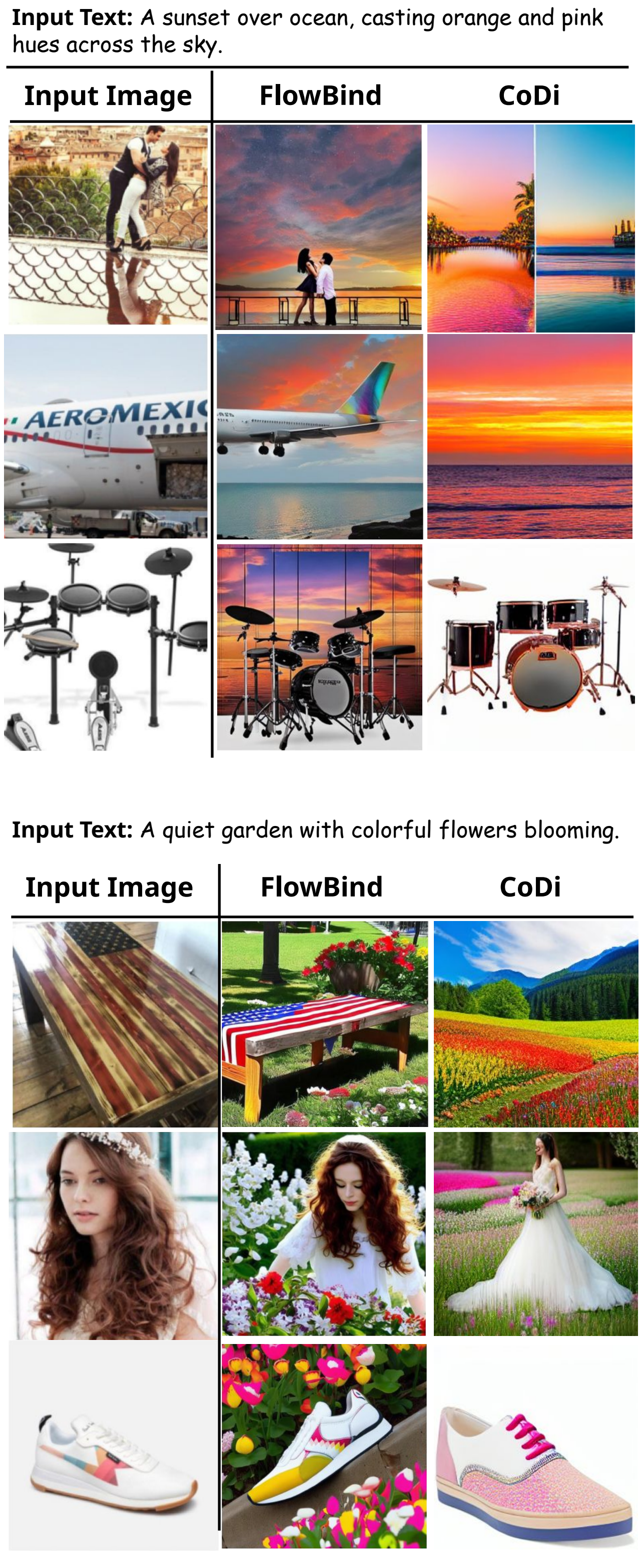

Input Text: A sunset over ocean, casting orange and pink hues across the sky.
Input Image
FlowBind
CoDi
AEROMEXIC
Input Text: A quiet garden with colorful flowers blooming.
Input Image
FlowBind
CoDi

Figure 10: Results on {text+image}-to-image generation.

F Quantitative Results on Many-to-Many Generation Tasks

In this section, we quantitatively evaluate FlowBind on many-to-one and one-to-many generation tasks to assess its performance in realistic any-to-any generation scenarios. To this end, we construct a synthetic triplet dataset by extending the AudioCaps text–audio pairs. Following a protocol similar to OmniFlow, we generate the missing image modality using FLUX.1-schnell (Labs et al., 2025), conditioned on the text annotations. This yields a triplet (text, audio, image) dataset that enables quantitative evaluation of many-to-many generation.

Tables 7 and 8 report the results for many-to-one and one-to-many settings, respectively, comparing FlowBind with other flow-based models. Note that for all alignment metrics (CLIP, CLAP, AIS), higher values indicate better alignment. FlowBind achieves competitive or superior alignment across these tasks and, in particular, exhibits a reduced tendency to ignore either modality in the many-to-one generation setting.

Table 7: Many-to-one generation alignment performances.

	(T+A) → I		(T+I) → A		(I+A) → T	
Method	CLIP (T→I)	AIS (A→I)	CLAP (T→A)	AIS (I→A)	CLIP (I→T)	CLAP (A→T)
CoDi	25.17	57.52	4.85	61.28	24.04	20.66
OmniFlow	24.06	54.90	7.68	59.32	26.38	**36.07**
FlowBind	**25.57**	**57.93**	**28.13**	**76.02**	**27.83**	35.21

Table 8: One-to-many generation alignment performances.

	T → (I+A)		I → (T+A)		A → (T+I)	
Method	CLIP (T→I)	CLAP (T→A)	CLIP (I→T)	AIS (I→A)	CLAP (A→T)	AIS (A→I)
CoDi	**26.61**	10.99	25.73	58.65	18.03	57.14
OmniFlow	24.71	12.92	26.36	63.99	36.07	54.22
FlowBind	25.02	**29.12**	**27.98**	**74.34**	**36.79**	**59.99**

G Data Flexibility of FlowBind

FlowBind demonstrates data flexibility, effectively working with arbitrary partially paired data. The ablation experiment in Table 4 supports this claim, showing that the zero-shot performance of FlowBind on the image-to-audio task are reasonable even without trained with image-audio pairs. In this section, we extend the evaluation by testing the ratio of paired data. Specifically, we comprehensively assess our method under varying fractions of partially paired data by expanding the experiments in Table 4. We vary the ratio of Image–Audio paired data (i.e., VGGSound) subsampled to different fractions (e.g., 0%, 1%, 3%, 10%, 30%, 100%), and alignment scores are measured. The results, shown in Figure 11, demonstrate FlowBind's robustness to varying ratios of partially paired data, showing reasonable perfomrance on image-to-audio generation even with 1% or 3% of the subset.

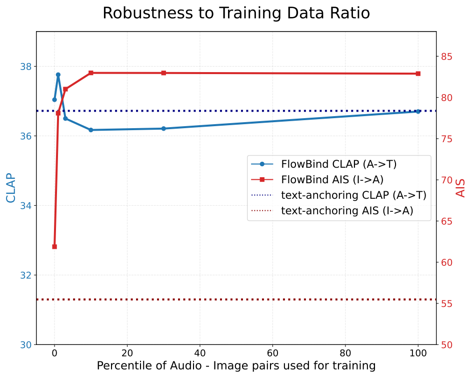


Figure 11: Performance of FlowBind varying fractions of Image–Audio data

H Robustness of Plain Averaging

To further assess the robustness of FlowBind under competing source modalities in many-to-one generation, we constructed a conflict set by randomly pairing audio clips with text prompts that deliberately describe different semantics. We then performed (T + A) → I generation with plain averaging in the shared latent space, and present the results in Figure 12.

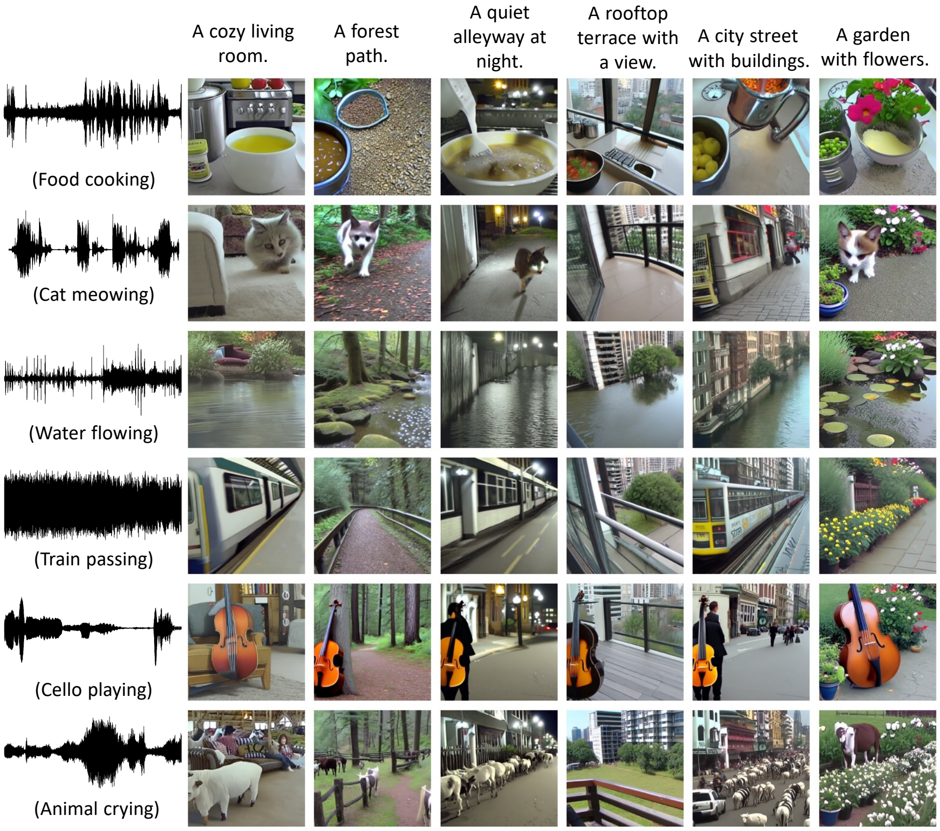


Figure 12: Results on conflicting conditions of {text+audio}-to-image generation.

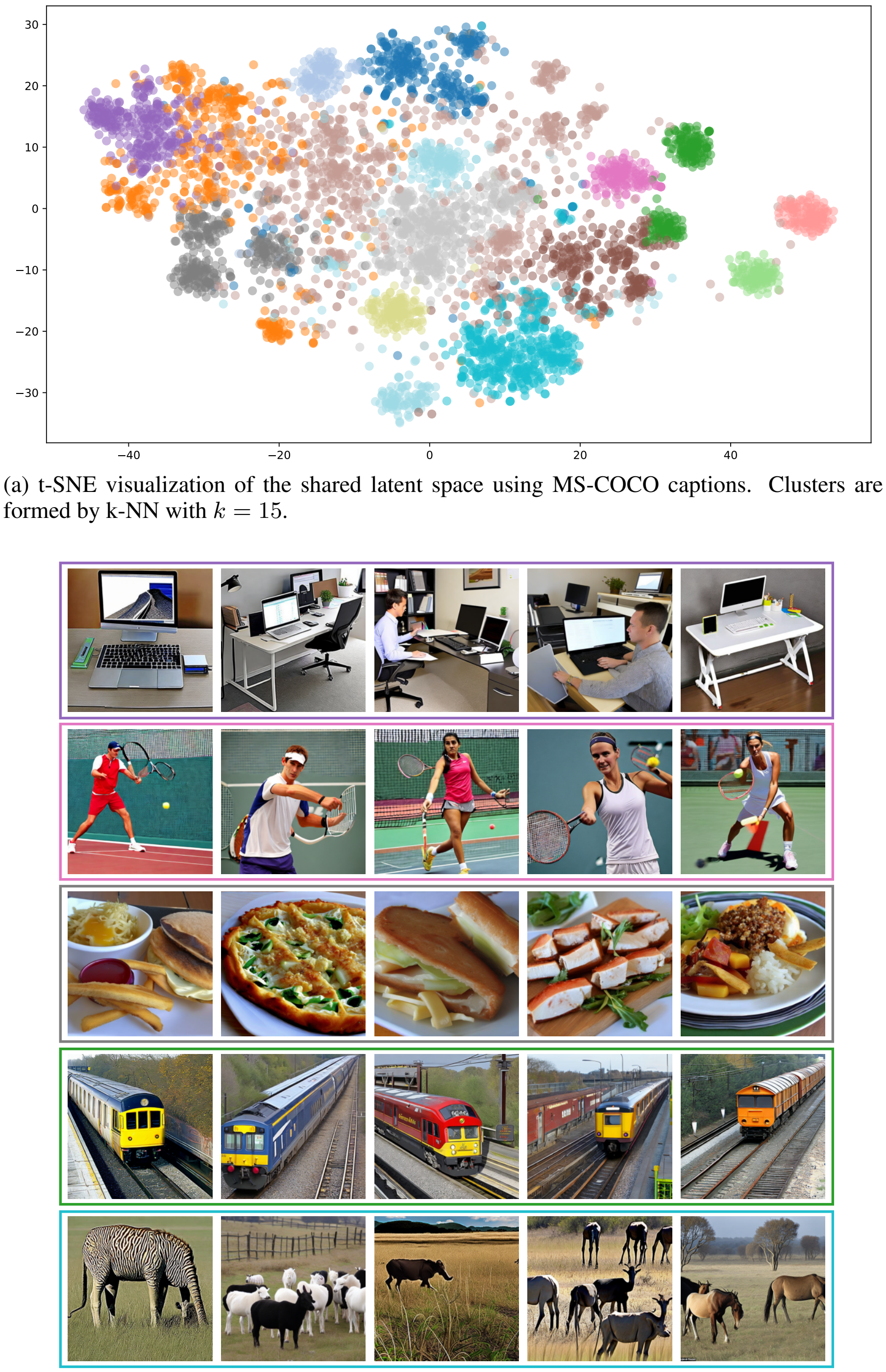
In this challenging setup, FlowBind faithfully reflects the two conflicting conditions in most cases, rather than collapsing to an incoherent blend or ignoring one modality.

We believe this robustness is the benefits of the shared latent space learned by FlowBind: as mentioned in Table 5, the shared latent achieves strong cross-modal alignment. As the shared latent space is well-structured and semantically aligned, even simple averaging leads to stable and meaningful behavior under conflicting conditions.

I FlowBind Latent Visualization

To further analyze the interpretability of the shared latent space and visualize the relationship between latents and generated content, we provide an additional t-SNE plot of FlowBind's shared latent space along with representative generated images.

In detail, we sampled 5,000 random text prompts from the MS-COCO evaluation set, encoded each prompt into FlowBind's shared latent space, and then performed clustering in this space using k-NN with k=15 which is shown in Figure 13a. For some clusters, we decoded the top 5 center features into images, as shown in Figure 13b. The images within the same cluster appear semantically very close, indicating that the shared latent space aligns meaningful semantic structure and that nearby latents correspond to coherent variations in the generated content.



(a) t-SNE visualization of the shared latent space using MS-COCO captions. Clusters are formed by k-NN with $k = 15$.

(b) Example image clusters decoded from latent points within a randomly selected cluster. Each color boundary represents a distinct cluster, as shown in 13a.

Figure 13: Visualization of shared latent space of FlowBind and corresponding generated images.

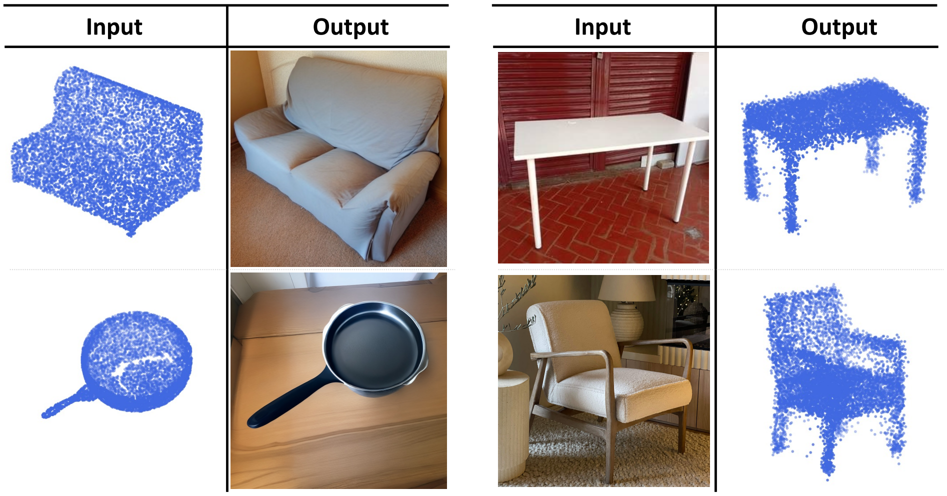
These examples in Figure 13b show that samples drawn from the same cluster in the shared latent space are semantically coherent (e.g., office scenes, tennis players, food dishes, trains, animals), while different clusters capture clearly distinct concepts. This supports that our shared latent space forms representations according to high-level semantics, so that nearby latent points correspond to consistent and meaningful variations in the generated images.

J FlowBind with Additional Modality

To demonstrate the scalability of FlowBind, we extend our framework to an additional modality, namely 3D point clouds. We use the Pix3D dataset (Sun et al., 2018), which contains 10k pairs of (Image, Point cloud), and adopt a pre-trained modality-specific autoencoder from (Yang et al., 2019). All other settings are kept the same as in our main experiments (Section 5); adding a new modality only introduces its modality-specific drift network, leading to approximately linear growth in the total number of parameters.

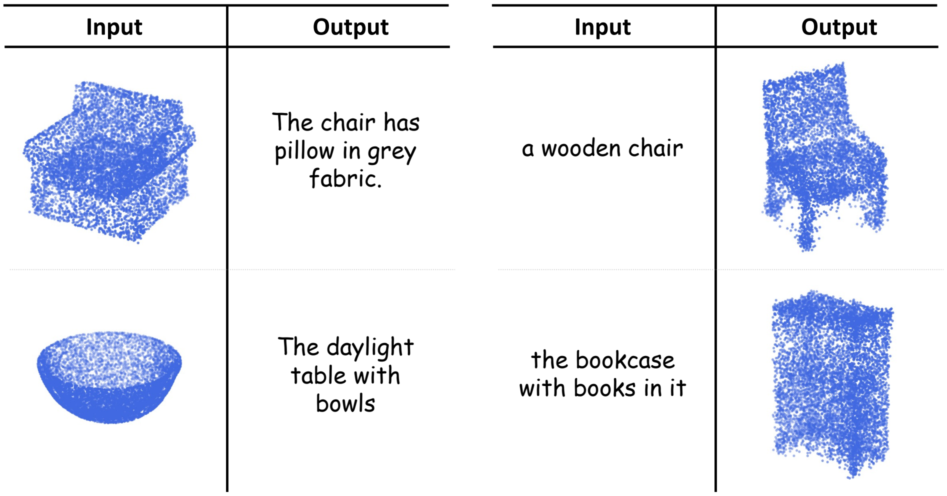
Figure 14 presents the qualitative results for cross-modal generation of image-point clouds, demonstrating strong performance while preserving the geometry of the underlying object and overall consistency of the shape.

More importantly, as shown in Figure 15, FlowBind also achieves reasonable performance on **unseen** cross-modal combinations (e.g., text → point clouds and point clouds → text), indicating that our framework can effectively exploit arbitrarily partially paired data, owing to its central learnable anchor design.



(a) Results on point clouds-to-image generation (b) Results on image-to-point clouds generation

Figure 14: Cross modal generation results on image–point clouds



The chair has pillow in grey fabric.

a wooden chair

The daylight table with bowls

the bookcase with books in it

(a) Results on point clouds-to-text generation (b) Results on text-to-point clouds generation

Figure 15: Cross-modal generation results on text–point clouds. FlowBind handles cross-modal generations *unseen during training* by effectively leveraging arbitrarily partially paired data.