FLOWSEED: DECOUPLED SOURCE LEARNING FOR CROSS-MODAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow-based generative models typically start from condition-independent Gaussian sources, leaving source initialization largely unexplored. We show that source initialization substantially affects generation quality, while jointly learning condition-dependent sources and generators can undesirably entangle source learning with generative modeling. We therefore propose **FlowSeed**, a decoupled framework that learns condition-dependent sources independently of the image generator. FlowSeed first encodes paired cross-modal correspondence in an aligned representation space, from which independently trained flows recover compatible source structures. For text-to-image generation, a lightweight text flow provides condition-dependent initialization for the aligned semantic component, while a token-wise FlowEdit-style strategy bridges it to the image generation flow. By decoupling source learning from generation, FlowSeed preserves the efficient and diverse generation paradigm of standard Flow Matching while enabling more effective condition-aware initialization. Experiments across multiple representation spaces consistently demonstrate improved generation performance over standard Flow Matching and learned-source baselines. Moreover, FlowSeed supports plug-and-play adaptation of pretrained generators to new conditioning modalities without retraining the generator.

1 SOURCE INITIALIZATION MATTERS

Flow-based and diffusion-based generative models (Zheng et al., 2026; Singh et al., 2026; Wang et al., 2026) typically initialize the generation process from a simple source distribution, e.g., an isotropic Gaussian distribution, and progressively transport source samples toward the target data distribution. While the source distribution is commonly treated as an uninformative starting point, we ask a more fine-grained question: *does the choice of an individual source sample matter for the final generation?*

To investigate this, we conduct a controlled experiment by fixing both the generation model and the text condition while varying only the initial source sample. Specifically, for each text prompt c, we independently sample K Gaussian noises and generate the corresponding images:

$$z_i \sim \mathcal{N}(0, I),\ x_i = G(z_i, c),\ i = 1, \ldots, K. \tag{1}$$

We then rank the generated samples according to their evaluation scores and compare the best and worst performing source initializations.

As shown in Fig. 1a, different source samples lead to noticeably different generation outcomes even when the generator and text condition are kept identical. Moreover, as the candidate pool grows, the gap between the best and worst initializations consistently increases. For example, the DPG-Bench gap increases from 0.5 at $K = 5$ to 2.7 at $K = 20$. These results suggest that the standard Gaussian source contains both favorable and unfavorable regions for a given condition, yet random sampling provides no mechanism for preferentially selecting the former.

This empirical observation is also consistent with the recent analysis of condition-dependent source learning in CSFM (Kim et al., 2026). For linear Flow Matching with $\Delta = X_1 - X_0$ and $X_t =$

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

$(1-t)X_0 + tX_1$, the FM objective can be decomposed as

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\Big[\|v_\theta(X_t,t) - u_t(X_t)\|_2^2\Big] + \mathbb{E}[\mathrm{Var}(\Delta \mid X_t)]\,, \tag{2}$$

where the second term denotes the *intrinsic variance* induced by the source–target coupling. A condition-agnostic Gaussian source is independently paired with condition-specific targets, which can induce intersecting transport paths and conflicting velocity supervision. In contrast, allowing the source distribution to depend on the condition, $p(X_0 \mid c)$, provides a mechanism to adapt the source–target coupling and reduce such ambiguity.

Together, the empirical variation among individual source samples and the coupling-level analysis above motivate a natural question: rather than randomly searching for favorable Gaussian samples, can we directly infer a better source from the conditioning information itself? Existing approaches such as CSFM address this by jointly optimizing a condition-dependent source and the generator. In the following, we show that such joint optimization can introduce undesirable interactions between source learning and generative modeling, motivating our decoupled formulation in FlowSeed.

2 Limitation of Joint Source-Generator Optimization

Given the observation that source initialization has a significant impact on generation quality, a natural question is how to obtain better source samples. Existing approaches mainly follow two directions: *source refinement* and *condition-dependent source construction*.

Source refinement. The first line of work improves an initially sampled Gaussian source while keeping the pretrained generator fixed. InitNO (Guo et al., 2024) optimizes the initial noise using attention-derived signals to improve text–image alignment, while ReNO (Eyring et al., 2024) and DNO (Tang et al., 2024) directly optimize the source according to downstream reward functions. Instead of performing per-sample optimization, Golden Noise (Zhou et al., 2025) and NoiseRefine (Ahn et al., 2026) learn auxiliary networks that transform random Gaussian samples into more favorable source samples. Related approaches also explore lightweight or prompt-aware noise refinement (Mannering et al., 2025; Tong et al., 2025).

In general, these methods can be written as

$$z^* = \mathcal{R}_\phi(z,c),\ z \sim \mathcal{N}(0,I), \tag{3}$$

where $\mathcal{R}_\phi$ denotes either an explicit optimization procedure or a learned refinement mapping. Although effective, such approaches still start from an independently sampled Gaussian source and mainly perform optimization around this initial sample. Moreover, the optimized source is typically coupled with the specific pretrained generator, since the quality of the refined source is evaluated through the downstream generation process:

$$z^* = \arg\max_z R(G_\theta(z,c),c), \tag{4}$$

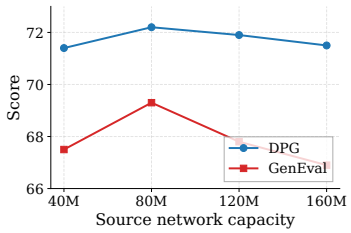
where $R(\cdot)$ denotes a generation quality metric. Therefore, source refinement is often tied to a particular generator or reward signal and may require additional optimization or source-quality supervision.

Condition-dependent source construction. Another line of work directly incorporates conditioning information into the source distribution or source representation. CPD (Issachar et al., 2025) constructs a conditional prior centered around a condition-dependent point in the target space, while CAR-Flow (Chen et al., 2026) learns condition-aware reparameterizations to improve the alignment between source and target distributions. CrossFlow (Liu et al., 2025) and FlowTok (He et al., 2025) further replace the conventional Gaussian source with learned representations from the conditioning modality, directly modeling cross-modal transport from text to image representations. Noise PPO (Miao et al., 2025) learns a prompt-conditioned initial-noise distribution while keeping the pretrained image generator fixed.

Most closely related to our setting, CSFM (Kim et al., 2026) introduces a learnable condition-dependent source distribution and jointly optimizes the source model with the generative flow.

	DPG-Bench			GenEval		
K	Best	Worst	Gap	Best	Worst	Gap
5	70.3	69.8	0.5	64.8	64.6	0.2
10	70.7	69.2	1.5	65.0	64.3	0.7
15	71.0	68.8	2.2	65.1	64.1	1.0
20	71.2	68.5	**2.7**	65.1	64.0	**1.1**

(a) Effect of source initialization. For each prompt, we sample K Gaussian sources and compare the best and worst generations.



(b) Effect of source-network capacity under joint source-generator optimization. Increasing source capacity beyond a moderate scale degrades generation performance.

Figure 1: **Motivation for decoupled source learning. Left:** Different Gaussian source samples can lead to substantially different generation quality under the same generator and condition, and the best–worst gap increases with the number of candidate sources. **Right:** Directly increasing the capacity of a jointly optimized source network does not monotonically improve generation quality, suggesting an undesirable interaction between source learning and generative modeling.

Specifically, a source network S_ϕ generates condition-dependent initial states:

$$z_0 = S_\phi(c), \tag{5}$$

and the source network and flow model are jointly optimized:

$$\min_{\phi,\theta} \mathbb{E}_{t,x_1,c}\left[\|v_\theta(x_t,t,c) - (x_1 - z_0)\|^2\right], \tag{6}$$

where the interpolation trajectory is defined as:

$$x_t = (1-t)z_0 + tx_1. \tag{7}$$

By learning a source distribution adapted to the conditioning information, CSFM reduces ambiguity in source–target transport and facilitates Flow Matching optimization.

However, we find that directly coupling source learning with generator optimization introduces a trade-off between source quality and generative capability. As shown in Fig. 1b, increasing the capacity of the source network does not consistently improve generation performance. Instead, after reaching a certain scale, a stronger source network leads to degraded generation quality.

We hypothesize that this phenomenon is caused by the interaction between source optimization and velocity field learning. In Flow Matching, the model learns the transport velocity:

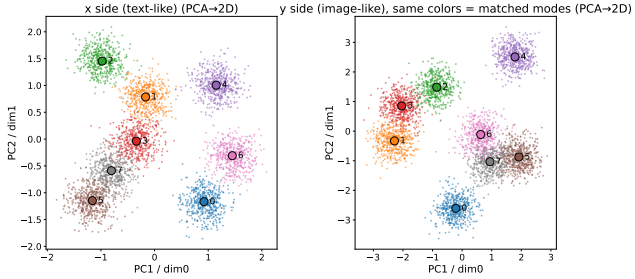
$$v^*(x_t,t) = x_1 - z_0. \tag{8}$$

When the source network becomes increasingly expressive, the learned source z_0 can already encode more information about the target sample, reducing the difficulty of the velocity prediction task. Consequently, the flow model may receive weaker supervision for learning a meaningful transport process and gradually lose its generation capability. In other words, *a better source does not necessarily lead to a better generator when the two components are optimized jointly*.

These observations reveal a limitation of joint source-generator optimization: source learning and generation modeling are entangled under the same training objective. To obtain high-quality condition-dependent sources while preserving the original generation ability, we propose to decouple source learning from generator training. Instead of modifying the source distribution through joint optimization, we directly learn a condition-to-source mapping and use the generated source as the initialization for an independently trained generator.

3 Source Alignment from Paired Data via Decoupled Flow Matching

The previous analysis suggests that jointly optimizing a condition-dependent source and the generator can introduce undesirable interactions between source learning and velocity-field modeling. We



(a) Visualization of the paired distributions in the 512-dimensional setting, projected to 2D using PCA. The two modalities exhibit different marginal geometries, while the same colors indicate corresponding semantic modes across modalities.

Dim.	Good-noise ↑	Prior	Lift ↑
5	59.2%	5.0%	11.8×
10	42.2%	5.0%	8.4×
15	42.6%	5.0%	8.5×
512	54.8%	5.0%	11.0×

(b) Good-noise rate across representation dimensions. The Gaussian prior baseline is fixed to 5% by construction.

Figure 2: **Toy verification of decoupled source alignment. Left:** A representative 512-dimensional paired distribution, where corresponding semantic modes are preserved across two different marginal geometries. **Right:** Sources recovered through the paired modality are 8.4–11.8× more likely to fall into the corresponding source region than random Gaussian samples, and the effect remains strong even at 512 dimensions.

therefore ask a more fundamental question: *can a condition-dependent source be recovered without jointly training the two modalities?* Our key hypothesis is that paired observations already contain sufficient correspondence information. Once this correspondence is encoded into aligned representations, two independently trained flows can recover compatible source structures even without parameter sharing or cross-modal supervision during Flow Matching.

Decoupled source alignment. Consider two domains $\mathcal{X}$ and $\mathcal{Y}$ with paired observations $(x, y) \sim p(x, y)$. For the toy verification below, we first learn aligned representations using pairwise alignment together with reconstruction regularization,

$$h_x = E_x(x),\ h_y = E_y(y),\ \mathcal{L}_{\text{repr}} = \|h_x - h_y\|_2^2 + \lambda \left(\|x - D_x(h_x)\|_2^2 + \|y - D_y(h_y)\|_2^2 \right), \quad (9)$$

where the alignment term transfers pairwise correspondence into the shared representation space, while the reconstruction term prevents representation collapse and preserves modality-specific information. In our toy experiments, we additionally use standard contrastive and variance regularization for stable representation learning.

After representation alignment, we train two Flow Matching models *independently* on the marginal distributions of h_x and h_y. Denoting their data-to-source ODE maps as T_x and T_y, respectively,

$$T_x : h_x \to \mathcal{Z},\ T_y : h_y \to \mathcal{Z},\ \mathcal{Z} = \mathcal{N}(0, I). \quad (10)$$

Importantly, the two flows do not share parameters and never observe paired samples during Flow Matching training. If the paired correspondence encoded in the representation space induces compatible transport structures, then for a matched pair we expect

$$T_x(E_x(x)) \approx T_y(E_y(y)), \quad (11)$$

despite the two flows being optimized independently. This provides the central mechanism behind FlowSeed: semantic correspondence can first be learned at the representation level and subsequently inherited by the source space, without coupling source learning to image-generator optimization.

Toy Verification. We verify whether paired cross-modal correspondence can be preserved in the source space when the two Flow Matching models are trained independently. We construct paired Gaussian mixtures with eight semantic modes and vary the representation dimensionality from 5 to 512.

To quantify source-space correspondence, we introduce the *good-noise rate*. We first select a source region $\mathcal{A} \subset \mathcal{Z}$ with 5% Gaussian prior mass and transport its samples through one modality and the paired cross-modal mapping. The resulting representations are then mapped back to the source

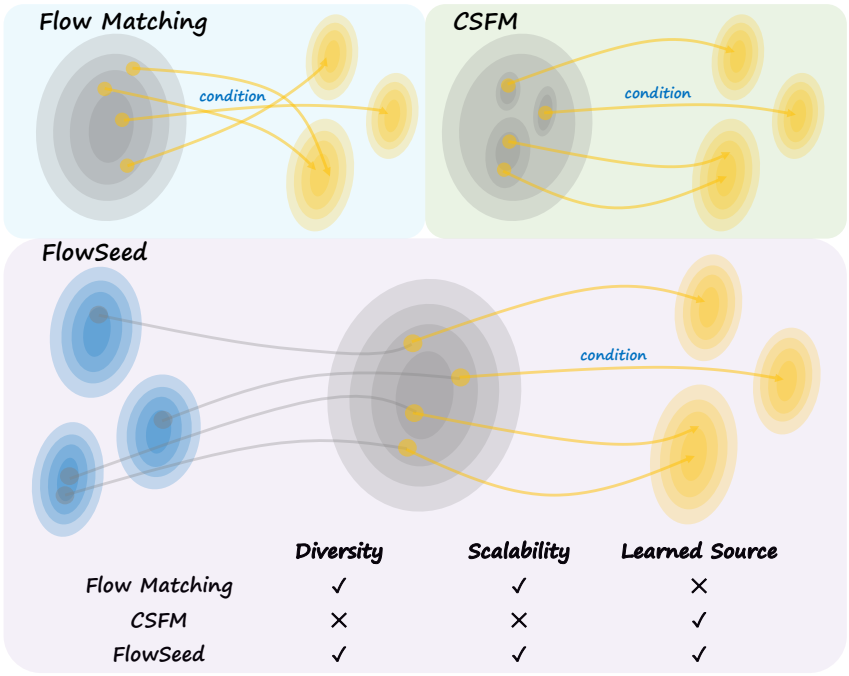


Figure 3: **Conceptual comparison of source initialization strategies.** Standard Flow Matching starts from condition-independent random Gaussian sources, preserving generation diversity and scalability but providing no mechanism for learning favorable source initializations. End-to-end methods such as CSFM introduce condition-dependent learned sources by jointly optimizing the source model and the generator, which couples source learning with generative modeling. In contrast, **FlowSeed** decouples the two processes: a condition-dependent source is learned independently and subsequently used to initialize an independently trained generative flow. This design combines learned source initialization with the diversity and scalability of conventional Flow Matching.

space by the independently trained flow of the other modality, forming a returned source set $\mathcal{Z}_{\text{ret}}$. We define its neighborhood

$$\mathcal{G} = \left\{ z : \min_{\tilde{z} \in \mathcal{Z}_{\text{ret}}} \|z - \tilde{z}\|_2 < \epsilon \right\}, \tag{12}$$

where ϵ is calibrated such that $P_{z \sim \mathcal{N}(0,I)}(z \in \mathcal{G}) = 5\%$. The good-noise rate is then

$$R_{\text{good}} = P_{z \sim \mathcal{A}}(z \in \mathcal{G}), \tag{13}$$

with a 5% random-Gaussian baseline by construction. A higher rate therefore indicates stronger overlap between source regions recovered independently from the paired modalities.

As shown in Fig. 2b, sources recovered through the paired modality fall into the corresponding source region far more frequently than random Gaussian samples. The good-noise rate reaches 59.2% in 5 dimensions and remains 54.8% even in 512 dimensions, corresponding to $11.8\times$ and $11.0\times$ improvements over the Gaussian baseline, respectively. This result supports our central hypothesis that *paired cross-modal information can be encoded in the representation space and subsequently preserved by independently trained flows in a shared Gaussian source space*. This makes it possible to learn condition-dependent sources without jointly optimizing the source model and the image generator.

4 FlowSeed: Decoupled Source Learning for Cross-Modal Generation

The previous experiments show that paired cross-modal representations provide a useful alignment signal for learning correlated source distributions. Motivated by this observation, we propose

FlowSeed, a decoupled framework that learns condition-dependent source initialization without modifying the original image generation objective. As shown in Fig. 3, the key idea is to separate cross-modal generation into two components: (i) learning a source distribution aligned with the input condition, and (ii) preserving the original conditional image generation process. FlowSeed therefore learns the source model and image generator independently, and only couples them during sampling.

Condition-Dependent Source Learning. Given a paired text–image sample (c, x), we extract aligned multimodal representations using modality encoders aligned in a shared representation space,

$$e_c = E_{\text{text}}(c),\ e_x = E_{\text{image}}(x). \tag{14}$$

We then independently train a text-side Flow Matching model on the distribution of e_c. Specifically, with $z_0 \sim \mathcal{N}(0, I)$, $z_1 = e_c$, and $z_t = (1-t)z_0 + tz_1$, the model learns

$$v_{\theta_c}(z_t, t) \approx z_1 - z_0. \tag{15}$$

At inference time, instead of sampling a new text representation, we start from the given condition embedding e_c at $t = 1$ and integrate the learned flow backward to $t = 0$:

$$z_c = \Phi_{\theta_c}^{1\to 0}(e_c),\ \frac{dz_t}{dt} = v_{\theta_c}(z_t, t), \tag{16}$$

where z_c serves as the condition-dependent source associated with c. Empirically, this text-side source model can be substantially smaller than the image generator, suggesting that source prediction mainly requires modeling cross-modal semantic correspondence rather than full visual synthesis.

Representation-Space Image Generation. To inject the learned source while preserving the original image generation capability, we adopt a factorized representation space inspired by REG (Wu et al., 2026). Each image representation is decomposed as

$$e_x = [g_x, l_x], \tag{17}$$

where g_x denotes the aligned global representation and l_x contains local visual features, e.g., VAE, DINO, or dense SigLIP features. A conditional image Flow Matching model $v_{\theta_x}(z_t, t, e_c)$ is trained with target $[g_x, l_x]$ using the standard Flow Matching objective.

During inference, FlowSeed only replaces the source of the aligned global component:

$$z_0 = [z_g, z_l],\ z_g = z_c,\ z_l \sim \mathcal{N}(0, I). \tag{18}$$

Thus, the global semantic token inherits a condition-dependent source predicted by the text flow, while the local tokens remain randomly initialized. This factorization injects condition information into the aligned semantic subspace without sacrificing the stochastic diversity of local image details.

Token-wise FlowEdit Sampling. The above formulation can be viewed as first mapping the text representation back to its aligned source and then running the image generation flow. We further simplify this two-stage procedure by interpreting the transformation from the text-aligned representation to the image-aligned representation as a token-wise editing process.

Specifically, for the global semantic token, we perturb the text representation along a FlowEdit-style trajectory,

$$\tilde{g}_t = (1-t)\epsilon + te_c,\ \epsilon \sim \mathcal{N}(0, I), \tag{19}$$

and remove the semantic transport already modeled by the text flow from the image velocity. Local tokens are left unchanged and follow the original image generation dynamics. The resulting token-wise velocity is

$$\frac{dz_t}{dt} = \left[v_{\theta_x}^{g}(z_t, t, e_c) - v_{\theta_c}(\tilde{g}_t, t),\ v_{\theta_x}^{l}(z_t, t, e_c)\right]. \tag{20}$$

In this way, FlowSeed avoids explicitly recovering a source noise and restarting the full image sampling process. Instead, it directly edits the aligned semantic component toward the target image representation while retaining the original dynamics for local visual features. This token-wise decoupling provides condition-dependent source control with minimal interference to the pretrained image generator.

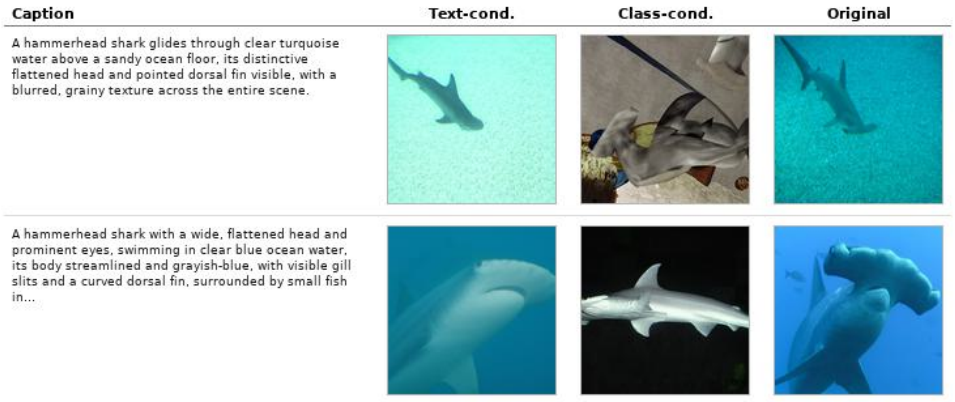


Figure 4: **Plug-and-play text conditioning with FlowSeed.** We apply FlowSeed to a pretrained class-conditional image generator RAE without updating its parameters. Given a natural-language caption, FlowSeed learns a text-conditioned source that enables the frozen generator to produce images consistent with the textual description. We show the resulting text-conditioned generations alongside the original class-conditioned outputs and reference images.

5 Plug-and-Play with Pretrained Generators

A key benefit of the decoupled formulation is that new conditioning modalities can be introduced without retraining the generator. We verify this property on a pretrained class-conditional RAE generator with DINO representations. Using paired ImageNet captions, we learn a lightweight text encoder aligned with the frozen image representation and independently train a text flow to predict its condition-dependent source. At inference, this source directly initializes the frozen generator. As shown in Fig. 4, FlowSeed enables the original class-conditional model to follow natural-language descriptions without updating any generator parameters.

More generally, once a generator has learned a strong source-to-image mapping, FlowSeed can extend it to new conditions by modifying only the source initialization. This decoupled design provides a simple and scalable alternative to jointly retraining the condition encoder and image generator.

6 Experiments

6.1 Experimental Setup

Data and training. For text-to-image generation, we construct a training set of approximately 1.49M image–text pairs by combining MS-COCO 2017, ShareGPT4V, Image-Generation-CoT, and OpenGPT-4o-Image (Lin et al., 2014; Chen et al., 2024; Liao et al., 2025; Chen et al., 2025). This collection provides high-quality image–text alignment and broad content diversity while requiring substantially less data and computation than large-scale text-to-image pretraining. Unless otherwise specified, all models are trained for 80 epochs with a global batch size of 256.

We evaluate FlowSeed across multiple image representation spaces, including VAE (Rombach et al., 2022), DINOv2 (Oquab et al., 2023), SigLIP (Tschannen et al., 2025), and CLIP (Radford et al., 2021), and use DINOv2, SigLIP, or CLIP as the global alignment space. Since DINOv2 does not provide a naturally aligned text representation, we additionally train a text encoder to align text features with its global image representation. During inference, standard Flow Matching and CSFM use a 50-step ODE solver. For FlowSeed, local image tokens are sampled with 50 ODE steps, while the aligned global token switches to the FlowEdit-style update after the first 15 steps.

Baselines and evaluation. We compare against standard Flow Matching and recent methods that modify or replace the conventional Gaussian source, including CrossFlow (Liu et al., 2025) and CSFM (Kim et al., 2026). We evaluate compositional prompt following using GenEval (Ghosh

Table 1: Comparison across different representation spaces under the same training protocol. CrossFlow denotes our reimplementation adapted to each representation setting. For the CLIP representation, no additional global alignment representation is used, since CLIP itself provides a globally aligned image–text representation; generation follows a DALL·E 2-style conditional diffusion decoder. Gray-shaded FlowSeed entries denote its standard FM sampling special case, where only the independently trained image flow is used. We report this fallback when it outperforms the full FlowSeed sampler.

Representation		Method	GenEval ↑	DPG-Bench ↑	FID ↓	CLIP Score ↑
Image	Global					
DINOv2	SigLIP	Standard FM	67.4	72.9	**2.78**	0.342
		CrossFlow	68.6	71.8	2.89	0.344
		CSFM	69.3	72.2	2.85	0.348
		FlowSeed	**72.8**	**75.1**	**2.78**	**0.361**
DINOv2	CLIP	Standard FM	64.0	71.4	**2.91**	0.335
		CrossFlow	64.6	70.8	2.99	0.336
		CSFM	65.2	71.1	2.98	0.339
		FlowSeed	**70.7**	**73.4**	**2.91**	**0.354**
DINOv2	DINOv2	Standard FM	67.3	72.1	2.88	0.341
		CrossFlow	65.1	71.3	2.97	0.336
		CSFM	65.7	71.8	2.94	0.338
		FlowSeed	**71.2**	**74.2**	**2.86**	**0.356**
VAE	SigLIP	Standard FM	66.2	62.9	3.12	0.328
		CrossFlow	64.8	63.1	3.19	0.329
		CSFM	65.4	63.5	3.15	0.331
		FlowSeed	**68.1**	**64.8**	**2.97**	**0.345**
SigLIP	SigLIP	Standard FM	67.8	73.3	**2.75**	0.345
		CrossFlow	68.0	73.4	2.78	0.347
		CSFM	68.5	73.8	2.81	0.349
		FlowSeed	**73.1**	**75.5**	**2.75**	**0.365**
CLIP	—	Standard FM	58.4	61.2	3.45	0.312
		CrossFlow	59.5	64.6	3.36	0.321
		CSFM	60.1	65.1	3.32	0.324
		FlowSeed	**64.7**	**70.8**	**3.10**	**0.338**

et al., 2023) and DPG-Bench (Hu et al., 2024), and report FID (Heusel et al., 2017) and CLIP Score on the COCO test set to measure image quality and text–image alignment.

6.2 System-level Comparison

Table 1 compares FlowSeed with standard Flow Matching, CrossFlow, and CSFM across multiple representation settings. Overall, FlowSeed consistently achieves the strongest text-to-image performance, yielding the best GenEval and DPG-Bench scores in all evaluated settings while also improving CLIP Score. On image quality, FlowSeed either obtains the best FID or matches standard Flow Matching when the learned source degenerates to the original image-flow initialization.

The gains are consistent across substantially different representation spaces, including VAE, DINOv2, SigLIP, and CLIP, suggesting that the proposed decoupled source learning is not tied to a specific latent representation. Compared with CrossFlow, which directly learns cross-modal transport, and CSFM, which jointly optimizes the source and generator, FlowSeed preserves the original image-generation capability while introducing a more effective condition-dependent source. This leads to improved semantic alignment without sacrificing generation quality, demonstrating the robustness and generality of the proposed framework.

Table 2: **Component-wise analysis of FlowSeed. Left:** Effect of source-module capacity. "Source Encoder/FM" denotes the source encoder used by CSFM and the text Flow Matching model used by FlowSeed, respectively. **Right:** Effect of the starting step of FlowEdit-style sampling. Higher is better for both GenEval and DPG-Bench.

Source Encoder/FM	Method	GenEval ↑	DPG-Bench ↑
40M	CSFM	67.5	71.4
	FlowSeed	70.9	72.6
80M	CSFM	**69.3**	**72.2**
	FlowSeed	72.8	75.1
120M	CSFM	67.8	71.9
	FlowSeed	73.2	75.7
160M	CSFM	66.9	71.5
	FlowSeed	**73.5**	**76.0**

FlowEdit Start Step	GenEval ↑	DPG-Bench ↑
0	71.9	74.4
5	72.3	74.7
10	72.5	74.9
15	**72.8**	**75.1**
20	72.4	74.6
25	71.6	73.8

6.3 Component-wise Analysis

We further analyze two key design choices of FlowSeed: the capacity of the source modeling module and the starting step of FlowEdit-style sampling. Unless otherwise specified, the ablations are conducted under the DINOv2/SigLIP representation setting.

Source-module capacity. The left part of Table 2 reveals a clear difference between joint and decoupled source learning. CSFM achieves its best performance with an 80M source encoder, while further increasing its capacity leads to degraded GenEval and DPG-Bench scores. In contrast, FlowSeed benefits consistently from a stronger text flow, improving from $70.9/72.6$ at 40M to $73.5/76.0$ at 160M. This result further supports our motivation for decoupling: increasing source model capacity does not interfere with the independently trained image generator, allowing the source model to scale without degrading its generative capability.

FlowEdit starting step. The right part of Table 2 studies when the aligned global token switches to the FlowEdit-style update. Performance improves as the starting step increases from 0 to 15, reaching the best GenEval and DPG-Bench scores of 72.8 and 75.1, respectively, and then decreases when the transition is further delayed. This suggests that a moderate transition point provides a better balance between following the image-flow trajectory and exploiting the condition-dependent source. We therefore use step 15 as the default setting in all main experiments.

7 Conclusion and Future Work

In this work, we introduce **FlowSeed**, a decoupled framework for condition-dependent source learning in cross-modal generation. Our key observation is that source initialization has a substantial impact on generation quality, while jointly optimizing the source distribution and the generator can undesirably entangle source learning with generative modeling. FlowSeed instead learns cross-modal correspondence in an aligned representation space and independently recovers condition-dependent sources through Flow Matching. Combined with token-wise FlowEdit sampling, this design improves semantic alignment while preserving the original generation capability. Experiments across multiple representation spaces demonstrate consistent gains over standard Flow Matching and existing learned-source approaches, while the plug-and-play results further show that source learning can serve as a lightweight interface for adapting pretrained generators to new conditioning modalities.

Looking forward, we believe decoupled source learning provides a general perspective for cross-modal generation beyond text-to-image synthesis. Future work includes extending FlowSeed to richer conditioning modalities such as audio and video, studying its scalability on larger pretrained generative models, and exploring more efficient source transport and sampling strategies for few-step generation.

AI USE STATEMENT

In this work, we used generative AI tools only for literature research and writing assistance, including searching for and summarizing relevant literature, polishing the wording and readability of the manuscript, and proofreading for language and presentation errors. We did not use generative AI tools for other research tasks requiring disclosure, including developing theoretical or conceptual frameworks, formulating mathematical claims or proofs, proposing hypotheses, designing the methodology or experiments, implementing the proposed method, processing datasets, or analyzing and interpreting experimental results. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content, claims, and results presented in this work.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics and all authors have read and adhered to the Code of Ethics. In this study, no human subjects is involved. The use of all datasets, including ImageNet, MS-COCO 2017, ShareGPT4V, Image-Generation-CoT and OpenGPT-4o-Image (Deng et al., 2009; Lin et al., 2014; Chen et al., 2024; Liao et al., 2025; Chen et al., 2025), follows the relevant usage guidelines and public licenses, ensuring no violation of privacy. We have been careful to avoid any biased or discriminatory results during our research process. No personally identifiable information is used, and no privacy or security concerns will be raised due to our experiments. We are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. All code and datasets have been made publicly available in an anonymous repository to facilitate replication and verification. The experimental setup, including training steps, model configurations, and hardware details, is described in detail in the paper. We have also provided a full description of FlowSeed to assist others in reproducing our experiments.

Additionally, the datasets used in our experiments are publicly available, ensuring consistent and reproducible evaluation results.

We believe these measures will enable other researchers to reproduce our work and further advance the field.

REFERENCES

Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Jaewon Min, Minjae Kim, Wooseok Jang, Hyoungwon Cho, Sayak Paul, SeonHwa Kim, Eunju Cha, et al. A noise is worth diffusion guidance. In *International Conference on Learning Representations*, volume 2026, pp. 94450–94498, 2026.

Chen Chen, Pengsheng Guo, Liangchen Song, Jiasen Lu, Rui Qian, Tsu-Jui Fu, Xinze Wang, Wei Liu, Yinfei Yang, and Alex Schwing. Car-flow: Condition-aware reparameterization aligns source and target for better flow matching. *Advances in Neural Information Processing Systems*, 38: 94919–94945, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024.

Zhihong Chen, Xuehai Bai, Yang Shi, Chaoyou Fu, Huanyu Zhang, Haotian Wang, Xiaoyan Sun, Zhang Zhang, Liang Wang, Yuanxing Zhang, et al. Opengpt-4o-image: A comprehensive dataset for advanced image generation and editing. *arXiv preprint arXiv:2509.24900*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems*, 37:125487–125519, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9380–9389. IEEE, 2024.

Ju He, Qihang Yu, Qihao Liu, and Liang-Chieh Chen. Flowtok: Flowing seamlessly across text and image tokens. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–12. IEEE, 2025.

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. *Advances in neural information processing systems*, 30, 2017.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Noam Issachar, Mohammad Salama, Raanan Fattal, and Sagie Benaim. Designing a conditional prior distribution for flow-based generative models. *arXiv preprint arXiv:2502.09611*, 2025.

Junwan Kim, Jiho Park, Seonghu Jeon, and Seungryong Kim. Better source, better flow: Learning condition-dependent source distribution for flow matching. *arXiv preprint arXiv:2602.05951*, 2026.

Jiaqi Liao, Zhengyuan Yang, Linjie Li, Dianqi Li, Kevin Lin, Yu Cheng, and Lijuan Wang. Imagegen-cot: Enhancing text-to-image in-context learning with chain-of-thought reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17214–17223, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A noise-free framework for cross-modality evolution. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2755–2765. IEEE, 2025.

Harvey Mannering, Zhiwu Huang, and Adam Prugel-Bennett. Noise-level diffusion guidance: Well begun is half done. *arXiv preprint arXiv:2509.13936*, 2025.

Yanting Miao, William Loh, Pacal Poupart, and Suraj Kothawade. A minimalist method for fine-tuning text-to-image diffusion models. *arXiv preprint arXiv:2506.12036*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Inference-time alignment of diffusion models with direct noise optimization. *arXiv preprint arXiv:2405.18881*, 2024.

Yunze Tong, Didi Zhu, Zijing Hu, Jinluan Yang, and Ziyu Zhao. Noise projection: Closing the prompt-agnostic gap behind text-to-image misalignment in diffusion models. *arXiv preprint arXiv:2510.14526*, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 40633–40642, 2026.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Ming-Ming Cheng, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *Advances in Neural Information Processing Systems*, 38:7714–7743, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, volume 2026, pp. 35791–35820, 2026.

Zikai Zhou, Shitong Shao, Lichen Bai, Shufei Zhang, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17688–17697. IEEE, 2025.

A Additional Analysis of Decoupled Source Alignment

This section provides additional details and experiments for the source alignment analysis in Sec. 3. Our goal is to examine whether paired cross-modal correspondence can be preserved in a shared Gaussian source space when the two Flow Matching models are trained *independently*. Importantly, paired supervision is used only for representation alignment; once the representations are obtained, the two flow models observe only their respective marginal distributions and share neither parameters nor paired samples.

A.1 Synthetic Paired Distributions

We construct several controlled joint distributions $p(x, y)$ with increasing levels of ambiguity and geometric mismatch. The two domains are referred to as the x side (text-like) and the y side (image-like).

Coordinate-wise monotonic mapping. We first consider a two-dimensional deterministic mapping. The first coordinate of x follows a two-component Gaussian mixture and the second coordinate is uniformly distributed:

$$\begin{aligned} x_1 &\sim 0.5\,\mathcal{N}(-1.3, 0.45^2) + 0.5\,\mathcal{N}(1.3, 0.45^2), \\ x_2 &\sim \mathcal{U}(-1.5, 1.5). \end{aligned} \tag{21}$$

The paired observation is obtained through a coordinate-wise monotonic transformation

$$g(x) = \begin{bmatrix} 0.6x_1^3 + 0.7 \\ 2\tanh(x_2) - 0.4 \end{bmatrix}, \qquad y = g(x). \tag{22}$$

This setting provides a simple case in which the correspondence is deterministic and globally rank preserving.

We additionally construct a stochastic many-to-many variant by adding isotropic Gaussian perturbation:

$$y = g(x) + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad \sigma = 0.3. \tag{23}$$

Unlike Eq. 22, a single x now admits multiple valid partners in the y domain.

Eight-mode many-to-many mapping. To better mimic semantic text–image correspondence, we construct two mixtures containing $K = 8$ Gaussian modes. The mode centers on the x side lie on a circle of radius $r_x = 2.0$, while the y-side centers lie on a different circle of radius $r_y = 3.2$:

$$\begin{aligned} c_k^x &= r_x \begin{bmatrix} \cos\theta_k \\ \sin\theta_k \end{bmatrix}, \quad c_k^y = r_y \begin{bmatrix} \cos(\theta_k + \delta) \\ \sin(\theta_k + \delta) \end{bmatrix}, \\ \theta_k &= \frac{2\pi k}{K}, \qquad\qquad k = 0, \dots, K-1. \end{aligned} \tag{24}$$

Samples are independently drawn within the corresponding modes:

$$x \mid k \sim \mathcal{N}(c_k^x, 0.18^2 I), \qquad y \mid k \sim \mathcal{N}(c_{\pi(k)}^y, 0.30^2 I). \tag{25}$$

Thus, correspondence exists at the semantic-mode level rather than at the instance level.

We evaluate three geometric variants:

- **Standard:** $\delta = 0$ and $\pi(k) = k$;
- **Rotated:** the y distribution is globally rotated by $\delta = 0.3927$ radians ($22.5°$);
- **Permuted:** the marginal distributions are unchanged, but the semantic mode correspondence π is randomly permuted.

The latter two settings are important controls: they preserve similar marginal structures while deliberately breaking the trivial geometric correspondence between the two spaces.

High-dimensional semantic mixtures. We further extend the many-to-many experiment to $d \in \{5, 10, 15, 512\}$. Both domains contain $K = 8$ Gaussian modes in $\mathbb{R}^d$, but their mode centers are sampled independently:

$$\begin{aligned} c_k^x &= 2.5\, \frac{\tilde{c}_k^x}{\|\tilde{c}_k^x\|_2}, \quad c_k^y = 4.0\, \frac{\tilde{c}_k^y}{\|\tilde{c}_k^y\|_2}, \\ \tilde{c}_k^x, \tilde{c}_k^y &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d). \end{aligned} \tag{26}$$

Conditioned on semantic mode k,

$$x \mid k \sim \mathcal{N}(c_k^x, 0.22^2 I), \qquad y \mid k \sim \mathcal{N}(c_k^y, 0.35^2 I). \tag{27}$$

The two domains therefore have independent ambient geometries and share only the semantic mode correspondence. Within each mode, x and y are sampled independently, resulting in a genuinely many-to-many joint distribution.

A.2 Pair-supervised Representation Alignment

Before training the flow models, we learn modality-specific MLP encoders and decoders (E_x, E_y, D_x, D_y) from paired observations. Let

$$u_x = E_x(x), \qquad u_y = E_y(y). \tag{28}$$

The alignment objective used in the toy experiments contains four terms:

$$\mathcal{L}_{\text{repr}} = \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{nce}} + \mathcal{L}_{\text{var}} + \mathcal{L}_{\text{rec}}. \tag{29}$$

The direct pairwise alignment term is

$$\mathcal{L}_{\text{mse}} = \mathbb{E}_{(x,y)} \left[\|u_x - u_y\|_2^2 \right]. \tag{30}$$

We additionally use a symmetric contrastive objective

$$\mathcal{L}_{\text{nce}} = \frac{1}{2} \left[\text{CE}\left(\frac{\bar{U}_x \bar{U}_y^\top}{\tau}, I \right) + \text{CE}\left(\frac{\bar{U}_y \bar{U}_x^\top}{\tau}, I \right) \right], \tag{31}$$

where $\bar{U}_x$ and $\bar{U}_y$ denote ℓ_2-normalized batch features and $\tau = 0.1$.

To prevent feature collapse, we regularize each feature dimension to maintain non-vanishing variance:

$$\mathcal{L}_{\text{var}} = \frac{1}{2} \sum_{m \in \{x,y\}} \mathbb{E}_j \left[\max(0, 1 - \text{Std}(u_{m,j}))^2 \right]. \tag{32}$$

Finally, the reconstruction term is

$$\mathcal{L}_{\text{rec}} = \frac{1}{2} \left(\|D_x(u_x) - x\|_2^2 + \|D_y(u_y) - y\|_2^2 \right). \tag{33}$$

All four terms use unit weights in our toy experiments.

After training, the two representations are standardized independently:

$$h_x = \frac{u_x - \mu_x}{\sigma_x}, \qquad h_y = \frac{u_y - \mu_y}{\sigma_y}. \tag{34}$$

The subsequent Flow Matching models are trained only on the marginal distributions $p(h_x)$ and $p(h_y)$.

A.3 Independent Flow Matching

We train two unconditional Flow Matching models v_x and v_y independently:

$$T_x : h_x \leftrightarrow z_x, \qquad T_y : h_y \leftrightarrow z_y, \qquad z_x, z_y \sim \mathcal{N}(0, I). \tag{35}$$

For either modality, given a target representation h and Gaussian source z, we use the linear probability path

$$h_t = (1 - t)z + th, \qquad t \sim \mathcal{U}(0, 1), \tag{36}$$

and optimize

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{h,z,t} \left[\|v_\theta(h_t, t) - (h - z)\|_2^2 \right]. \tag{37}$$

No paired information is used in Eq. 37. In particular, v_x and v_y have separate parameters and are optimized on independently sampled batches from their respective marginal distributions.

At evaluation time, the data-to-source and source-to-data mappings are obtained by integrating the learned ODE:

$$\frac{dh_t}{dt} = v_\theta(h_t, t). \tag{38}$$

We use fourth-order Runge–Kutta integration with 100 steps for all reported toy experiments.

A.4 Evaluation Protocols and Metrics

We evaluate source correspondence from several complementary perspectives.

Protocol A: paired source alignment. For paired observations (x_i, y_i), we first compute

$$z_i^x = T_x(h_i^x), \qquad z_i^y = T_y(h_i^y) \tag{39}$$

through reverse ODE integration. We report the average paired source distance

$$D_{\text{pair}} = \frac{1}{N} \sum_i \|z_i^x - z_i^y\|_2, \tag{40}$$

together with a shuffled baseline

$$D_{\text{shuffle}} = \frac{1}{N} \sum_i \|z_i^x - z_{\pi(i)}^y\|_2, \tag{41}$$

where π is a random permutation. We additionally measure Pearson and Spearman correlations between the two source representations.

For retrieval, each z_i^x is used to rank the y-side sources by Euclidean distance. Recall@1% measures whether its paired z_i^y falls within the closest 1% of a retrieval pool of 2,048 samples.

Protocol A': cross-modal source transfer. To test whether the recovered source has generative meaning, we directly feed the source inferred from x into the independently trained y flow:

$$\hat{h}_y = T_y^{-1}(z_x). \tag{42}$$

For semantic-mixture experiments, we report *Mode Accuracy*, i.e., the probability that $\hat{h}_y$ belongs to the semantic mode paired with the original x. A randomly sampled Gaussian source serves as the baseline.

Protocol B: source-region round trip. We select a Gaussian source region

$$\mathcal{A} = \{z : \|z - c\|_2 < r\}, \tag{43}$$

where $c \sim \mathcal{N}(0, I)$ and r is calibrated such that

$$P_{z\sim\mathcal{N}(0,I)}(z \in \mathcal{A}) = 0.05. \tag{44}$$

We sample z_{in} from this region, generate an x-side sample, obtain a paired y-side observation from the known synthetic coupling, and map it back through the independently trained y flow:

$$z_{\text{in}} \xrightarrow{T_x^{-1}} h_x \to h_y \xrightarrow{T_y} z_{\text{back}}. \tag{45}$$

The *Region Return Rate* measures $P(z_{\text{back}} \in \mathcal{A})$, whose random baseline is 5%.

Good-noise rate. Exact pointwise source recovery is unnecessarily restrictive for stochastic many-to-many mappings. We therefore also measure whether the recovered cross-modal sources define a consistent *source region*.

Let $\mathcal{Z}_{\text{back}} = \{z_{\text{back}}^{(j)}\}$ denote the set of sources obtained after the round trip in Eq. 45. We define its ϵ-neighborhood as

$$\mathcal{G}_\epsilon = \left\{ z : \min_{\tilde{z}\in\mathcal{Z}_{\text{back}}} \|z - \tilde{z}\|_2 < \epsilon \right\}. \tag{46}$$

The radius ϵ is calibrated such that

$$P_{z\sim\mathcal{N}(0,I)}(z \in \mathcal{G}_\epsilon) = 0.05. \tag{47}$$

The *good-noise rate* is then defined as

$$R_{\text{good}} = P_{z\sim\mathcal{A}}(z \in \mathcal{G}_\epsilon), \tag{48}$$

and we report the lift

$$\text{Lift} = \frac{R_{\text{good}}}{0.05}. \tag{49}$$

Therefore, $R_{\text{good}} > 5\%$ indicates that the source regions induced independently by the paired modalities overlap substantially more than expected under random Gaussian sampling.

Protocol C: semantic source partition agreement. For mixture distributions, we additionally sample the same Gaussian source z and generate one sample through each independent flow:

$$\hat{h}_x = T_x^{-1}(z), \qquad \hat{h}_y = T_y^{-1}(z). \tag{50}$$

We assign each output to its nearest semantic mode and measure the fraction for which the two labels agree. With $K = 8$ modes, random agreement is $1/K = 12.5\%$.

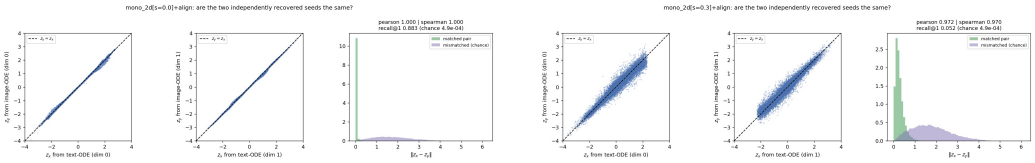
A.5 Low-dimensional Results

Continuous mappings. Table 3 reports results for the deterministic and noisy coordinate-wise mappings. For the deterministic mapping, independently recovered sources are nearly identical: the paired source distance is only 0.020, compared with 1.756 after shuffling, and the source Spearman correlation reaches 1.000. The selected source region is recovered with a 96.4% return rate compared with the 5% Gaussian baseline.

Introducing stochasticity weakens exact pointwise correspondence, as expected, but the alignment remains pronounced. With $\sigma = 0.3$, the paired source distance is 0.270 versus 1.766 for shuffled pairs, while Spearman correlation remains 0.970 and the region return rate is 61.5%.

Table 3: Source alignment on continuous 2D mappings. "Shuf." denotes randomly shuffled cross-modal pairs and "Chance" denotes random Gaussian initialization.

Setting	Source Dist. ↓	Spearman ↑	R@1% ↑	Region Return ↑
Monotonic	0.020 / 1.756	1.000	1.000	96.4 / 5.0%
Noisy monotonic	0.270 / 1.766	0.970	0.443	61.5 / 5.0%



(a) Deterministic monotonic correspondence. (b) Noisy many-to-many correspondence.

Figure 5: **Source alignment under continuous paired mappings.** Each panel compares independently recovered x-side and y-side sources. The deterministic case exhibits nearly pointwise source correspondence, while strong alignment remains under stochastic pairing.

Geometric and permutation controls. The eight-mode experiments allow us to distinguish alignment induced by paired information from alignment caused merely by similar marginal geometry. Table 4 compares models trained directly on the original spaces with models trained after pair-supervised representation alignment.

In the standard ring, the raw geometries are already approximately aligned, and the two independent flows naturally recover similar semantic partitions. However, rotating or permuting the y space strongly degrades cross-modal source transfer when no representation alignment is used. In particular, mode accuracy drops to 50.4% under rotation and 24.9% under permutation. After pair-supervised alignment, the same settings recover 99.6% mode accuracy, while source-partition agreement returns to approximately 99%. This confirms that source correspondence can be inherited from paired representations even when the original marginal geometries are incompatible.

A.6 High-dimensional Results

We next evaluate whether the phenomenon persists as the representation dimension increases. Table 5 summarizes the results for $d \in \{5, 10, 15, 512\}$.

The aligned representations remain highly consistent at all dimensions: their cosine similarity is above 0.997. More importantly, transferring a source inferred from the x modality through the independently trained y flow preserves the correct semantic mode with 98.8–99.9% accuracy, whereas random Gaussian initialization remains close to the 12.5% chance level.

The good-noise metric yields the same conclusion at the source-region level. Across all dimensions, the recovered source region receives 42.2–59.2% of selected sources despite being calibrated to contain only 5% of the Gaussian prior, corresponding to a 8.4–$11.8\times$ lift. Notably, at $d = 512$ the good-noise rate remains 54.8% ($11\times$ over prior), showing that the observed source correspondence is not restricted to low-dimensional toy spaces.

A.7 Implementation Details

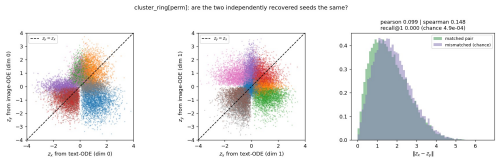
All experiments use MLP velocity fields with a 64-dimensional Fourier time embedding and SiLU activations. For the 2D experiments, the x-side (text-like) flow uses three hidden layers of width 96, while the y-side (image-like) flow uses four hidden layers of width 256. Representation encoders and decoders use three-layer MLPs with width 128.

For the high-dimensional experiments, the feature dimension is set equal to the ambient dimension d. The x-side flow width is

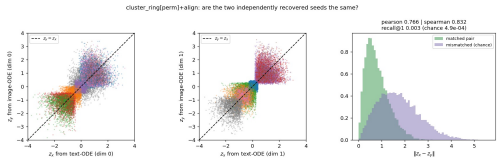
$$\min(512, \max(96, 16d)), \tag{51}$$

Table 4: Many-to-many eight-mode controls. Source Dist. reports paired / shuffled source distance. Mode Acc. reports transferred-source / random-source accuracy. Partition Agr. reports agreement / chance level.

Setting	Pair Align.	Source Dist. ↓	Mode Acc. ↑	Partition Agr. ↑
Standard ring	No	0.859 / 1.771	99.7 / 12.7%	99.3 / 12.5%
Standard ring	Yes	0.824 / 1.761	99.5 / 12.1%	99.0 / 12.5%
Rotated ring	No	0.971 / 1.769	50.4 / 12.2%	49.7 / 12.5%
Rotated ring	Yes	0.828 / 1.763	99.6 / 12.6%	99.0 / 12.5%
Permuted ring	No	1.649 / 1.777	24.9 / 12.7%	25.2 / 12.5%
Permuted ring	Yes	**0.805** / 1.743	**99.6** / 12.8%	98.4 / 12.5%



(a) Permuted correspondence without pair alignment.



(b) The same permutation after pair-supervised alignment.

Figure 6: **Pair supervision resolves incompatible cross-modal geometry.** Permuting the semantic correspondence substantially destroys source alignment when each flow is trained directly on its marginal distribution. Learning paired representations before the independent flows restores a strongly corresponding source structure.

and the y-side width is

$$\min(1024, \max(256, 24d)). \tag{52}$$

The batch size is 4,096 for $d < 256$ and 2,048 for $d = 512$.

The representation aligners are optimized for 5,000 steps in the 2D experiments and 6,000 steps in the high-dimensional experiments. The two Flow Matching models are subsequently trained independently for 10,000 and 12,000 steps, respectively. We use AdamW with learning rate 2×10^{-3} and cosine learning-rate decay. The flow models use an EMA with decay 0.999. Gradients are clipped to norm 1.0.

Unless otherwise stated, evaluation uses 20,000 paired samples and 100-step RK4 ODE integration. Source-region experiments use a Gaussian region with prior mass 5%. All reported experiments use random seed 0.

B Additional Experimental Details

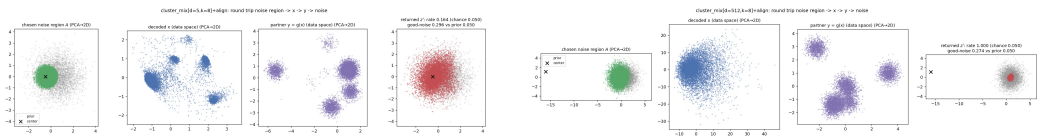
This section provides additional implementation and training details omitted from the main paper. Unless otherwise specified, all compared methods use the same data, image representation, generative backbone, optimization protocol, and sampling budget within each representation setting. The implementation is based on PyTorch and distributed data-parallel training.

Data preprocessing. As described in the main paper, we train on approximately 1.49M paired image–text samples collected from MS-COCO 2017, ShareGPT4V, Image-Generation-CoT, and OpenGPT-4o-Image. All images are center-cropped and resized to 256×256. We do not use random horizontal flipping or other image augmentations. The input normalization follows the corresponding frozen representation encoder. The paired caption associated with each image is used both for conditional image generation and for learning the cross-modal alignment. All representation encoders and decoders are kept frozen throughout Flow Matching training.

Common generative backbone. For fair comparison, Standard FM, CrossFlow, CSFM, and FlowSeed use the same image-flow architecture under each representation setting. We adopt a

Table 5: Source alignment on high-dimensional eight-mode Gaussian mixtures. "Pair Align." reports feature-space pair distance / cosine similarity. Mode Acc. compares transferred sources with random Gaussian sources. The good-noise baseline is calibrated to 5% prior mass. Partition Agr. is compared with the 12.5% chance level for eight modes.

Dim.	Pair Align. ↓ / ↑	Mode Acc. ↑	Good-noise Rate ↑	Partition Agr. ↑
5	0.067 / 0.999	99.6 / 12.7%	59.2 / 5.0% (11.8×)	96.6 / 12.5%
10	0.118 / 0.997	99.4 / 12.5%	42.2 / 5.0% (8.4×)	97.4 / 12.5%
15	0.147 / 0.998	98.8 / 12.3%	42.6 / 5.0% (8.5×)	97.8 / 12.5%
512	0.185 / 1.000	99.9 / 12.8%	54.8 / 5.0% (11.0×)	83.1 / 12.5%



(a) $d = 5$. (b) $d = 512$.

Figure 7: **Cross-modal source-region correspondence.** Visualization of the source-region round-trip at low and high dimensionality. For $d > 2$, the data-space panels are projected to two dimensions using PCA for visualization; the reported good-noise rates are computed in the original representation space.

NextDiT-style Transformer equipped with a lightweight DDT prediction head. For representation-space generation, the main image flow contains a 14-layer Transformer encoder with hidden dimension 864, followed by a 2-layer DDT head with hidden dimension 2048. The encoder and decoder use 12 and 16 attention heads, respectively, with 4 and 8 key/value heads for grouped-query attention. We additionally use two context-refiner blocks. QK normalization, RMSNorm, multi-axis rotary positional embeddings, and SiLU-gated feed-forward layers are used throughout the network, while no additional absolute positional embedding is introduced.

The input projection is adapted only to the dimensionality and spatial layout of the corresponding image representation. In particular, the DINOv2 RAE representation has shape $768 \times 16 \times 16$, for which we use patch size 1, while the VAE representation has shape $4 \times 32 \times 32$ and uses patch size 2. Other representation settings follow the same backbone design with the input/output dimensions adjusted accordingly. Thus, the model architecture and training recipe remain otherwise unchanged across representation spaces.

Frozen representation models. For representation-based generation, images are first encoded by the corresponding pretrained image encoder and the generative flow operates entirely in its representation space. The resulting representation is decoded by the associated frozen decoder after sampling. For the DINOv2 setting, we use the pretrained RAE encoder–decoder pair; for the VAE setting, we use the pretrained VAE with its standard latent scaling factor. SigLIP- and CLIP-based experiments similarly use their corresponding frozen representation models. When a representation does not provide a naturally aligned text embedding, we learn the lightweight text-to-representation alignment module described below.

Cross-modal representation alignment. Before training the text-side flow, we learn a mapping from frozen text features to the target image representation. We implement this mapping using a Perceiver-style network with learnable query tokens. The number of output queries is chosen to match the spatial token count of the target representation, such that its output can be directly reshaped into the corresponding image feature layout. The module uses a hidden dimension of 768, multi-head attention, RMSNorm, and SiLU-gated feed-forward layers with an expansion ratio of 4. The input text encoder remains frozen.

For the main FlowSeed experiments, the alignment module is deterministic and is trained by combining feature regression and directional alignment:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{MSE}} + 0.5\,\mathcal{L}_{\cos}, \qquad \mathcal{L}_{\cos} = 1 - \cos(h_c, h_x), \tag{53}$$

where h_c is the predicted text representation and h_x is the frozen image representation. We do not introduce a variational KL objective for this module in FlowSeed. Once trained, the alignment module is frozen when training the text-side Flow Matching model.

Text-side Flow Matching. The text flow is trained independently of the image generator. Given an aligned text representation h_c, we sample Gaussian noise $z \sim \mathcal{N}(0, I)$ and train an unconditional Flow Matching model along the linear path

$$h_t = (1-t)z + t h_c, \qquad u_t = h_c - z. \tag{54}$$

The model minimizes the standard velocity regression objective

$$\mathcal{L}_{\text{textFM}} = \mathbb{E}_{z,h_c,t}\left[\|v_{\theta_c}(h_t, t) - u_t\|_2^2\right]. \tag{55}$$

Importantly, the text flow receives no image input and no additional cross-modal supervision: the cross-modal correspondence has already been encoded into h_c by the frozen alignment module. Consequently, after representation alignment is learned, text-flow training only requires the captions and their aligned text representations.

We use the same NextDiT–DDT model family for the text flow but substantially reduce its capacity relative to the image generator. In the source-capacity ablation, we vary only the width and depth of this text flow to obtain approximately 40M, 80M, 120M, and 160M parameter variants while keeping the image generator, data, objective, and sampling procedure fixed.

Image Flow Matching. The image generator is trained independently with the standard conditional Flow Matching objective. Given an image representation x_1 and $x_0 \sim \mathcal{N}(0, I)$, we construct

$$x_t = (1-t)x_0 + t x_1, \qquad u_t = x_1 - x_0, \tag{56}$$

and optimize

$$\mathcal{L}_{\text{imageFM}} = \mathbb{E}\left[\|v_{\theta_x}(x_t, t, c) - u_t\|_2^2\right]. \tag{57}$$

The full token-level text representation is provided as context to the image flow. The image encoder, decoder, and text-conditioning encoder are frozen; only the generative flow is optimized. In particular, FlowSeed introduces no additional objective into image-flow training, which is important for preserving the original generative capability of standard FM.

Timestep sampling. We use a linear probability path and velocity prediction for all Flow Matching models. Following the implementation used for representation-space generation, we first draw $u \sim \mathcal{U}(0, 1)$ and apply a dimension-dependent timestep shift,

$$t = \frac{su}{1 + (s-1)u}, \qquad s = \sqrt{\frac{D}{4096}}, \tag{58}$$

where D denotes the total dimensionality of the modeled representation. For representations with $D = 4096$, this reduces to uniform timestep sampling. The same timestep schedule is used by all compared methods within a given representation setting.

Optimization details. Unless otherwise specified, all models in the main experiments are trained for 80 epochs with a global batch size of 256. We use AdamW with learning rate 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, and zero weight decay. The gradient norm is clipped to 1.0. We maintain an exponential moving average of trainable parameters with decay 0.9995 and use the EMA weights for evaluation. Training is performed in bfloat16 precision with distributed data parallelism. The global random seed is set to 0. No additional optimization of the source is performed at inference time.

For clarity, the common training configuration is summarized below.

Table 6: Common optimization settings used in the main experiments.

Setting	Value
Image resolution	256×256
Flow path	Linear
Prediction target	Velocity
Training epochs	80
Global batch size	256
Optimizer	AdamW
Learning rate	2×10^{-4}
AdamW β	$(0.9, 0.95)$
Weight decay	0
Gradient clipping	1.0
EMA decay	0.9995
Training precision	bfloat16
Sampling solver	Euler ODE
Sampling steps	50

FlowSeed sampling. For Standard FM and CSFM, we use the same 50-step Euler ODE sampler. FlowSeed likewise uses a total sampling budget of 50 image-flow steps. Local visual tokens follow the original image-flow trajectory throughout the entire sampling process. For the aligned global semantic component, we first follow the standard image flow and then switch to the token-wise FlowEdit-style update described in the main paper. Unless otherwise specified, the switch is performed after the first 15 sampling steps.

At an editing step, we construct a perturbed source state from the aligned text representation and Gaussian noise, evaluate the text-flow and image-flow velocities, and update the global component using their velocity difference. The local component is simultaneously updated using the original image-flow velocity. Therefore, the source-side flow only modifies the semantically aligned component and does not replace the local image-generation dynamics. The FlowEdit-start ablation varies the transition point among $\{0, 5, 10, 15, 20, 25\}$ while keeping all other settings unchanged.

Conditional generation uses the same classifier-free guidance configuration for all methods within each comparison. In our evaluation implementation, we use a guidance scale of 2.0. No additional guidance is applied specifically to FlowSeed.

Baseline implementation and controlled comparisons. Standard FM uses the same image-flow backbone and starts all representation components from independent Gaussian noise. Our Cross-Flow baseline is reimplemented using the same representation encoder, training data, image backbone, optimization schedule, and sampling budget as FlowSeed, with only the cross-modal transport formulation changed. For CSFM, we retain the same image generator and jointly optimize its condition-dependent source module following the CSFM formulation. In the source-capacity study, only the capacity of the source-side module is varied; the image generator is kept unchanged. These controls ensure that the differences in the main comparison arise from the source-learning strategy rather than from model capacity or training budget.

Evaluation protocol. We use identical prompt sets and sampling settings for all methods in each comparison. GenEval evaluation uses the complete 553-prompt benchmark set. DPG-Bench is evaluated using its benchmark prompts and evaluation protocol. For COCO evaluation, we compute FID using Inception features and CLIP Score between generated images and their corresponding captions. All reported samples are decoded using the same frozen decoder associated with each representation space. Unless explicitly stated otherwise, evaluation uses the EMA checkpoints.

Details of the source-initialization study. For the source-sensitivity experiment in the main paper, we fix both the text condition and the trained image generator and independently sample $K \in \{5, 10, 15, 20\}$ Gaussian initializations. All samples use exactly the same ODE solver, guidance configuration, and sampling budget. We rank generations only according to their evaluation score

and report the corresponding best–worst gap. Therefore, this experiment isolates the effect of the initial Gaussian sample without changing either the condition or the generator.

Toy verification. The toy study is designed to test source correspondence independently of the complexity of real image generation. We construct paired Gaussian mixtures with $K = 8$ semantic modes. The two modalities have independently generated mode geometries, and samples are paired according to their semantic mode while the exact samples within each mode are drawn independently. Thus, the pairing is many-to-many rather than an artificial point-wise deterministic mapping.

For each modality, we first learn an MLP encoder and decoder that map the two domains into a common feature space. The alignment networks use three-layer MLPs with hidden width 128. They are optimized with

$$\mathcal{L}_{\mathrm{toy-align}} = \mathcal{L}_{\mathrm{pair}} + \mathcal{L}_{\mathrm{NCE}} + \mathcal{L}_{\mathrm{var}} + \mathcal{L}_{\mathrm{rec}}, \tag{59}$$

where $\mathcal{L}_{\mathrm{pair}}$ is pairwise MSE, $\mathcal{L}_{\mathrm{NCE}}$ is a symmetric contrastive loss with temperature 0.1, $\mathcal{L}_{\mathrm{var}}$ prevents representation collapse, and $\mathcal{L}_{\mathrm{rec}}$ preserves information required to reconstruct the two modalities. All four terms have equal weight. We train the alignment networks with AdamW for 6,000 iterations, using batch size 4096, learning rate 2×10^{-3}, weight decay 10^{-4}, and cosine learning-rate decay.

After alignment, two velocity MLPs are trained *independently* on the two feature marginals. They use the same linear Flow Matching objective with Gaussian sources and receive no paired samples during flow training. We train each flow for 12,000 iterations with batch size 4096 and learning rate 2×10^{-3}, maintain an EMA with decay 0.999, and use gradient clipping of 1.0. ODE inversion and generation use a 100-step RK4 solver.

We evaluate dimensions 5, 10, 15, and 512. For the good-noise experiment, the selected source region $\mathcal{A}$ is calibrated to contain 5% probability mass under $\mathcal{N}(0, I)$. Samples from this region are transported through one modality, mapped to their paired modality, and inverted through the independently trained flow to obtain the returned source set $\mathcal{Z}_{\mathrm{ret}}$. We then define the good-noise region as the ϵ-neighborhood of $\mathcal{Z}_{\mathrm{ret}}$, where ϵ is numerically calibrated such that an independent Gaussian prior also falls into this neighborhood with probability 5%. Consequently, the random baseline is fixed to 5% by construction and the reported good-noise rate directly measures enrichment of corresponding source regions.

Plug-and-play experiment. For the plug-and-play experiment, we start from a pretrained representation-based RAE generator on ImageNet and freeze all of its parameters. The generator was originally trained with class conditioning and therefore does not directly support free-form text conditioning. Using the paired text annotations described in the main paper, we train a lightweight text encoder to regress to the frozen DINO image representation and then train the source-side Flow Matching model independently. During adaptation, neither the pretrained image flow nor its decoder is updated. At inference time, only the text-to-source pathway is added to the original model, and the resulting condition-dependent source is used by the frozen generator. This setting therefore isolates the ability of FlowSeed to introduce a new conditioning modality without retraining the original generative model.

C Additional Experimental Results

C.1 Additional Qualitative Results for Plug-and-Play Adaptation

We provide additional qualitative results for the plug-and-play experiment described in the main paper. The pretrained RAE generator is originally trained with class conditioning only, while FlowSeed keeps the entire generator frozen and introduces natural-language conditioning solely through the learned text-to-source pathway.

Figures 9 and 10 present 16 additional examples. Each row contains the input caption, the generation obtained with the FlowSeed text-conditioned source (**Text-cond.**), the output of the original class-conditioned generator (**Class-cond.**), and the corresponding reference image (**Original**). The captions contain substantially richer information than the original class labels, including object appearance, viewpoint, pose, illumination, background, and surrounding scene context.

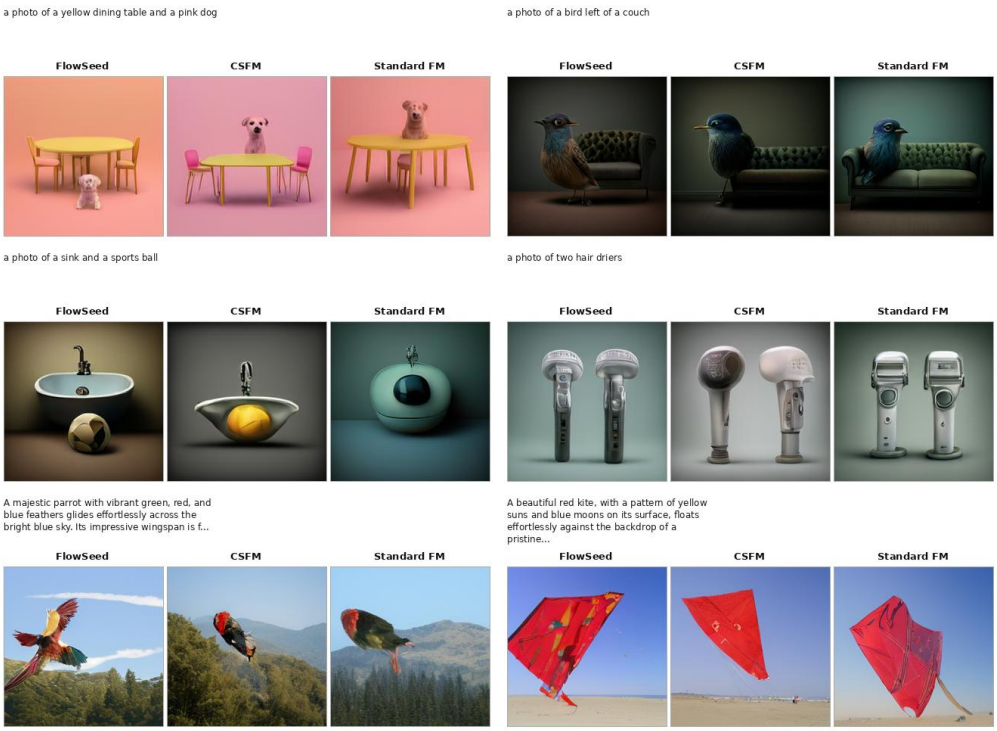


Figure 8: **Additional qualitative comparison among Standard FM, CSFM, and FlowSeed.** For each prompt, we compare generations produced by the three methods under the same setting. FlowSeed generally achieves better semantic alignment while preserving strong visual quality.

The qualitative results show that, without updating the pretrained generator, FlowSeed allows the model to respond to these fine-grained textual conditions while retaining the visual prior learned by the original generator. Compared with class conditioning, the text-conditioned results more directly reflect caption-specific attributes and scene configurations. These results further demonstrate that condition-dependent source learning can serve as a lightweight interface for introducing a new conditioning modality into an existing generator, without requiring joint retraining of its generative parameters.

C.2 Qualitative Comparison

We provide additional qualitative comparisons among **Standard FM**, **CSFM**, and **FlowSeed** to complement the quantitative results in the main paper. The purpose of this comparison is to more directly visualize the effect of different source-initialization strategies on text-to-image generation quality and semantic alignment.

Figure 8 shows representative examples under the same prompts and training setting. For each prompt, we compare generations produced by Standard FM, CSFM, and FlowSeed. Standard FM uses a condition-independent Gaussian source, CSFM jointly learns a condition-dependent source together with the generator, and FlowSeed learns the source in a decoupled manner and combines it with the independently trained image flow only at sampling time.

The qualitative results are consistent with the main quantitative findings. Compared with Standard FM, both CSFM and FlowSeed generally improve semantic alignment with the input prompt, indicating the benefit of introducing condition-aware source initialization. Compared with CSFM, FlowSeed tends to produce images that are better aligned with the prompt while preserving stronger visual quality and more coherent image structure. These examples further support our central claim that decoupling source learning from generator training provides a more effective way to exploit condition-dependent source initialization without weakening the underlying image-generation capability.

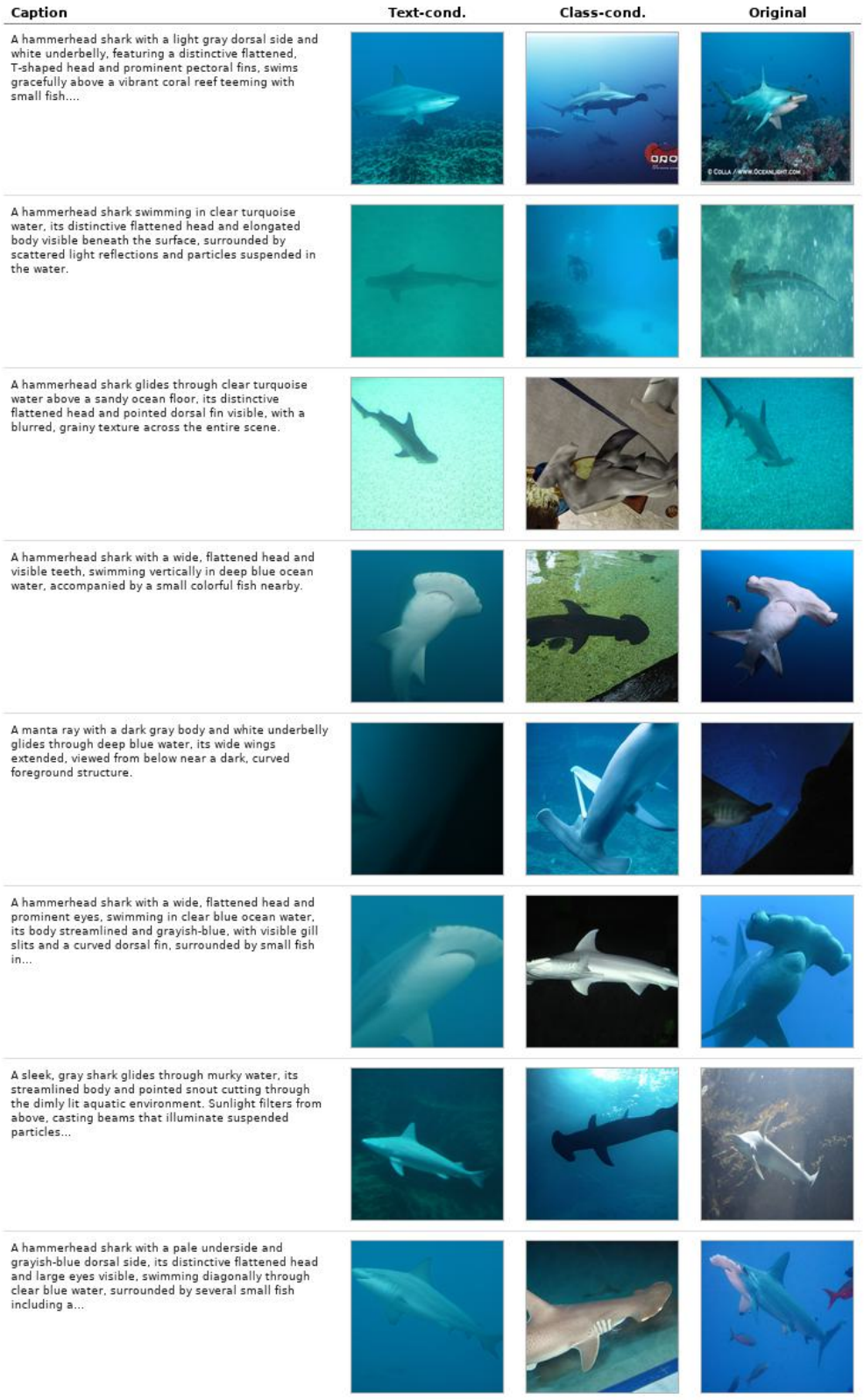


Figure 9: **Additional qualitative results of plug-and-play text conditioning (Part I).** Each row shows, from left to right, the input caption, FlowSeed text-conditioned generation (**Text-cond.**), output of the original pretrained RAE generator using class conditioning (**Class-cond.**), and the paired reference image (**Original**). FlowSeed introduces fine-grained text conditioning through source learning while keeping the pretrained image generator frozen.

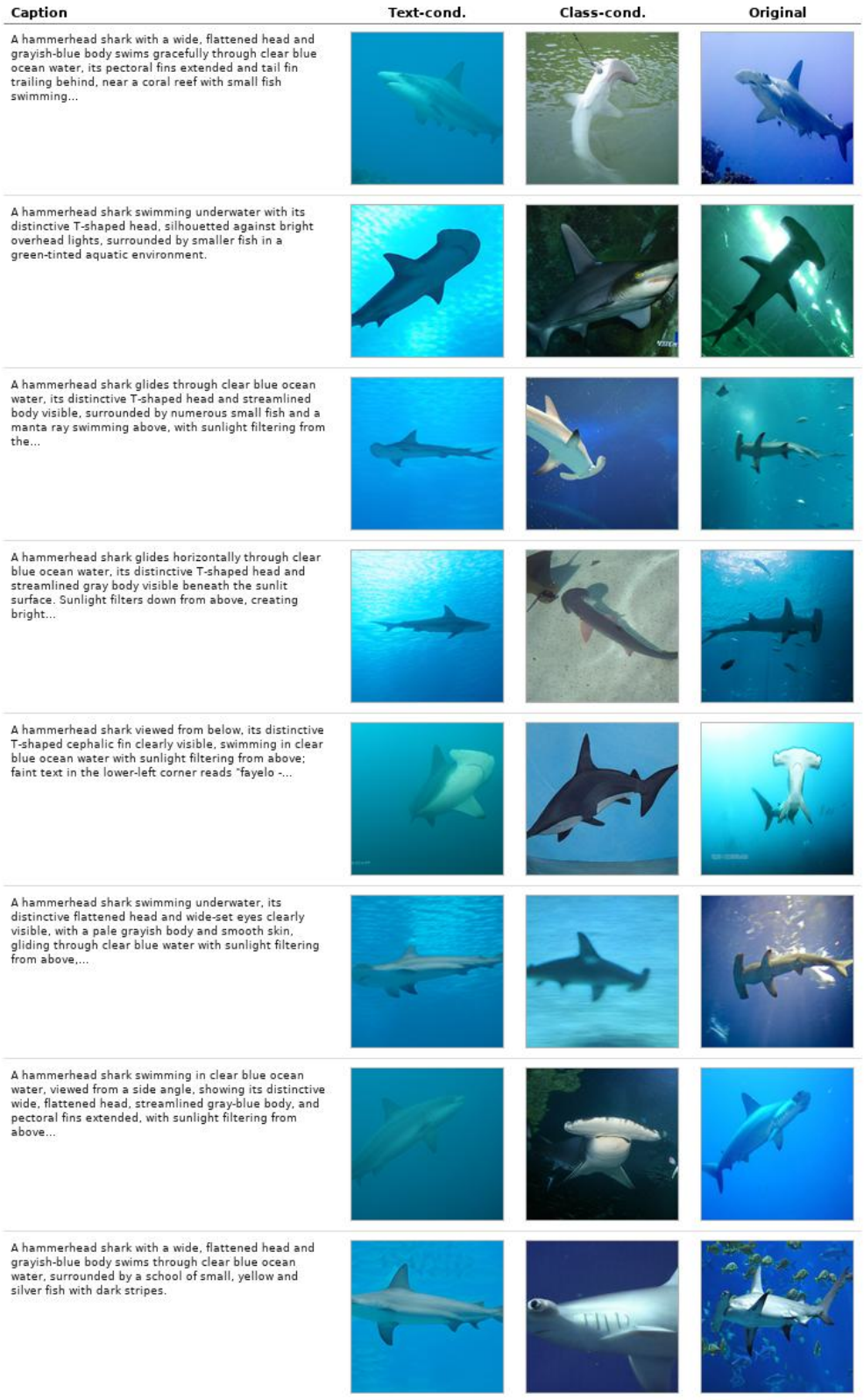


Figure 10: **Additional qualitative results of plug-and-play text conditioning (Part II).** More examples under the same setting as Fig. 9. FlowSeed enables the frozen class-conditional generator to follow detailed natural-language descriptions by modifying only the source-side conditioning pathway.