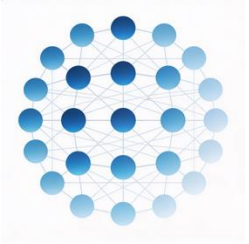
Dynamic and Structural Sampling for Interpretable Control in Multimodal Generation

Ye Zhu
Monge Assistant Professor
Department of Computer Science
Laboratoire d'Informatique (LIX)
École Polytechnique
ye.zhu@polytechnique.edu

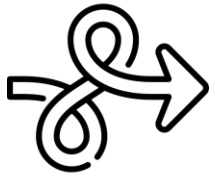
https://l-yezhu.github.io

About My Research in Generative Models

Generative Models (VAEs, GANs, DMs, etc.)



- ❑ [ZWSY **TPAMI**'24]
- ❑ [Z*Z*ZSY **ICASSP**'24]
- ❑ [ZWORTY **ICLR**'23]
- ❑ [ZOWACYT **ECCV**'22]
- ❑ [ZWYY **TPAMI**'21]
- ❑ [ZWHYY **ICASSP**'21]
- ❑ [ZWYY **ECCV**'20]



Generation Dynamics (probabilistic properties, asymptotic behaviors, etc.)

Multimodality (Vision and other modalities)

- ❑ [ZWDRY **NeurIPS**'23]
- ❑ [W*Y*QZ+W+ **ICLR**'24]
- ❑ [YZDR **ICML**'24]
- ❑ [WHZRW **ICCV**'25 Highlight]
- ❑ [ZMMHWZ **NeurIPS**'25]
- ❑ [ZWXDYR **NeurIPS**'25 Workshop]

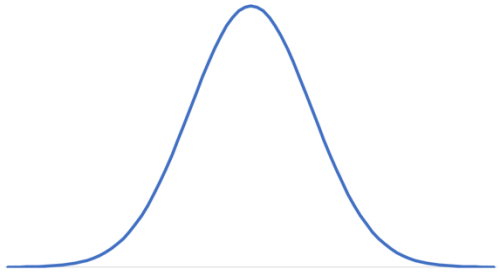
Dynamics & Physics

- ❑ [ZXDTR **NeurIPS**'25]
- ❑ [XKLZHT **The Astrophysics Journal (APJ)'**25]
- ❑ [XZ **Astronomy and Computing**'24]
- ❑ [XTHZ **The Astrophysics Journal (APJ)**'23 & ICLR-Workshop'23]

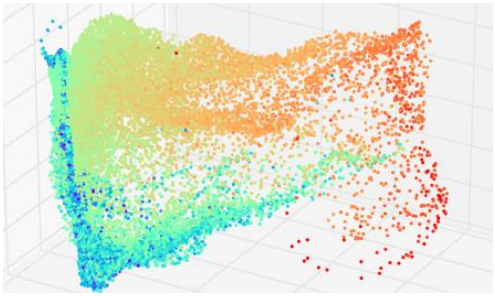
Generative Modeling

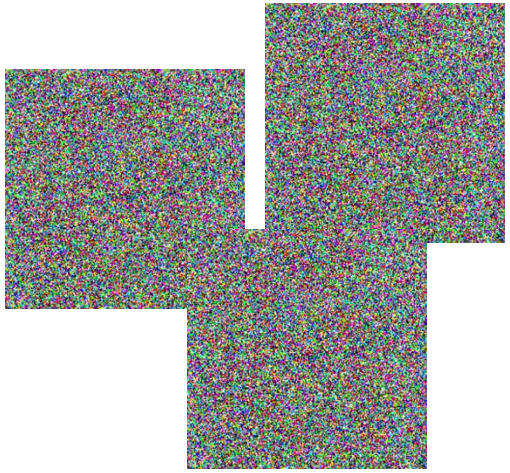
Goal of generative models:
Learning a mapping function between two distributions

$\mathbf{X}$ $\quad f: \mathbf{X} \rightarrow \mathbf{Y}$ $\quad \mathbf{Y}$



Generative
Neural Networks p_θ

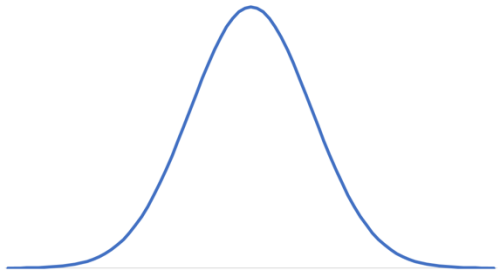




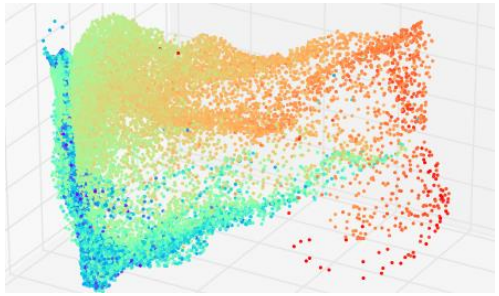
Dynamic Generative Models

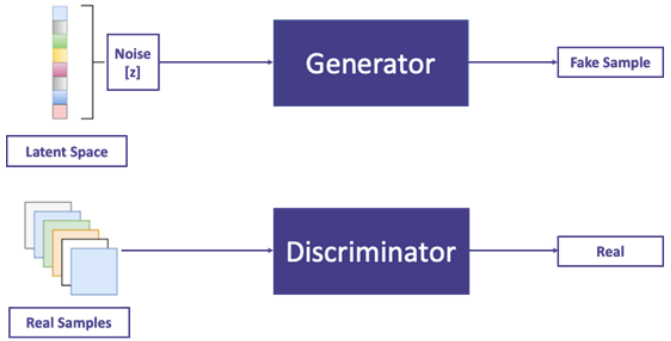
Goal of generative models:
Learning a mapping function between two distributions

$\mathbf{X}$ $\quad f: \mathbf{X} \rightarrow \mathbf{Y}$ $\quad \mathbf{Y}$

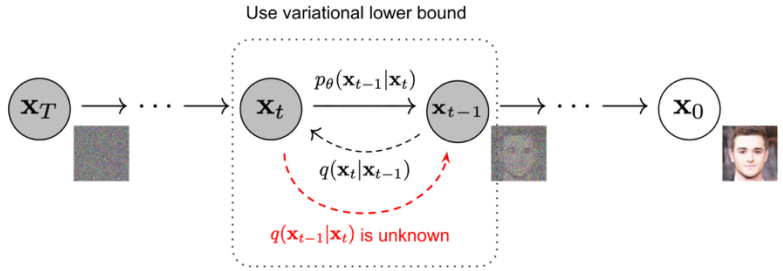


Generative
Neural Networks $\boldsymbol{p}_\theta$



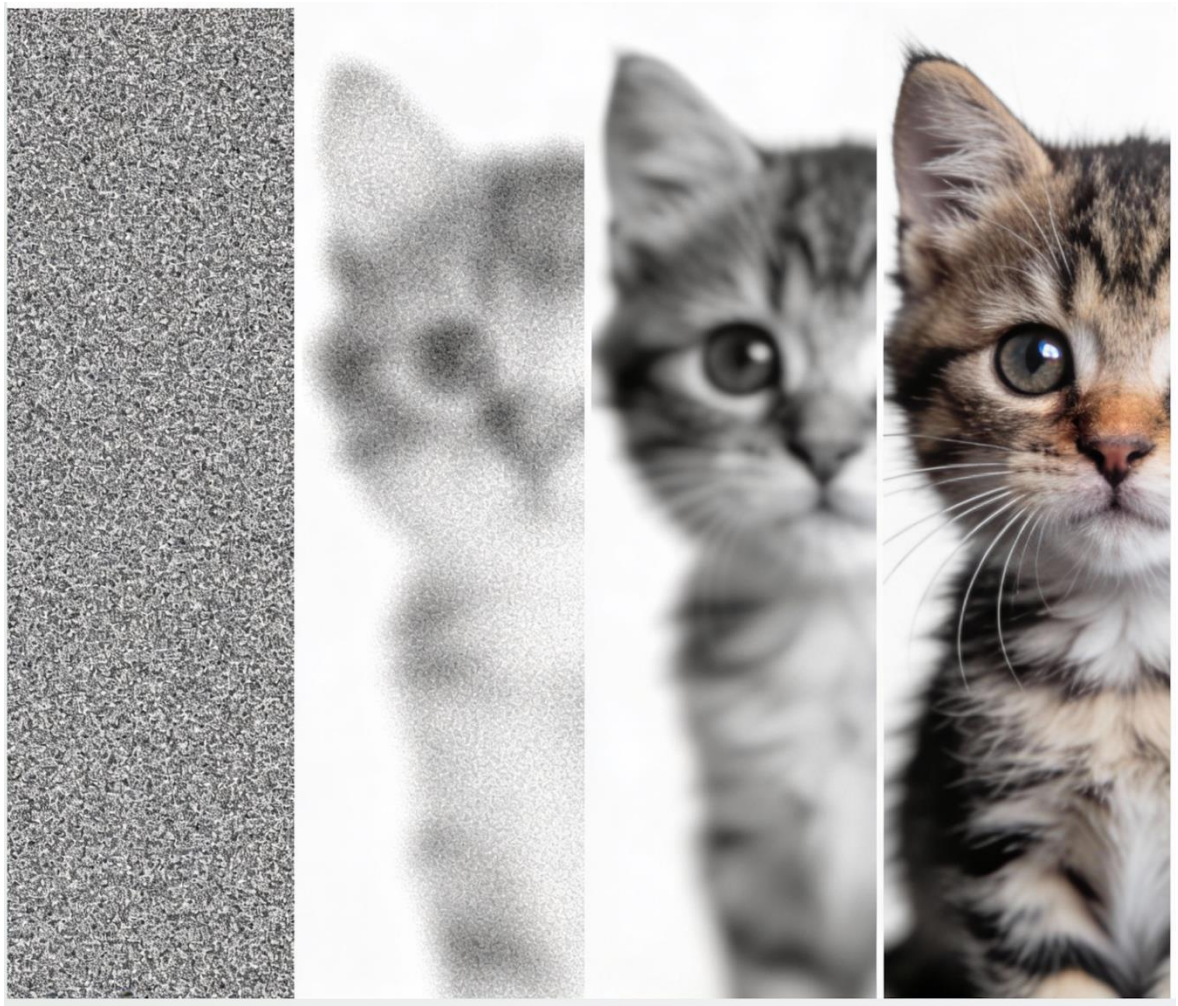


"Generative adversarial networks."
Goodfellow, Ian, et al., 2014.



"Denosing diffusion probabilistic models"
Ho, Jonathan, et al., 2020.

Dynamics along Generation Trajectory



- What happens in the middle of this generation process?
- What are the physical properties within this generation process?
- Are there any connections between marginal distributions?

How to leverage those properties to better control multimodal generations?

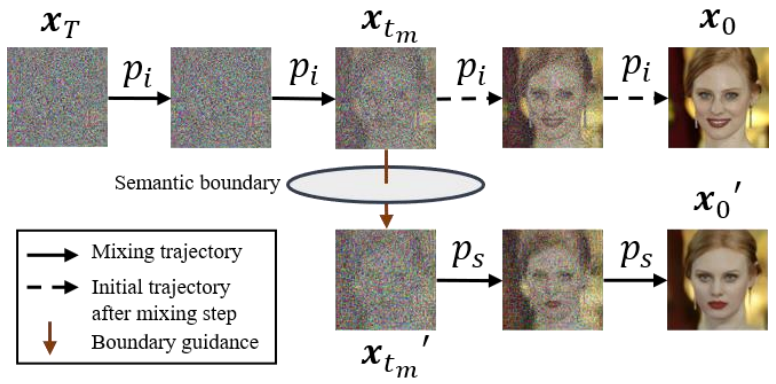
(Post) Learning v.s. Sampling for Generative Control

(Post) - Learning:
Seeks to achieve fine-grained controlling of multimodal generative models by altering the model parameters through additional guidance.

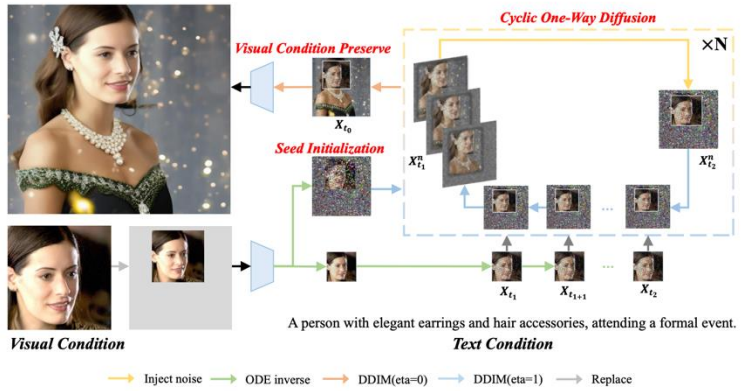
Sampling:
Seeks to achieve fine-grained controlling of multimodal generative models w/o altering the model parameters through additional guidance in inference. (Frozen Generative Models)

Why sampling over post-learning?
Pros and Cons: Preservation of the pre-trained modeling ability, better generalization, but also extra sampling overhead

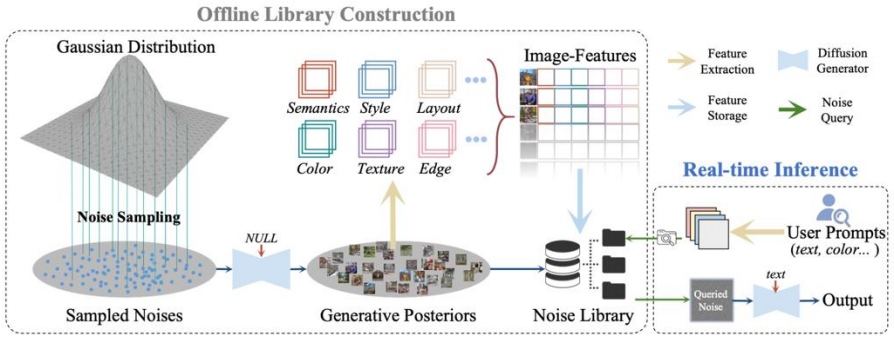
Dynamic Sampling for Interpretable Control in Multimodal Generation



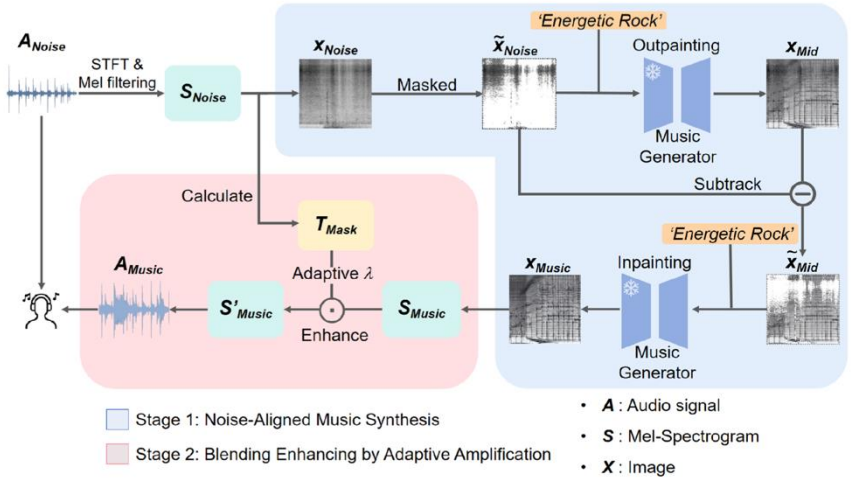
[ZWDRY **NeurIPS**'23]



[W*Y*QZ⁺W⁺ **ICLR**'24]



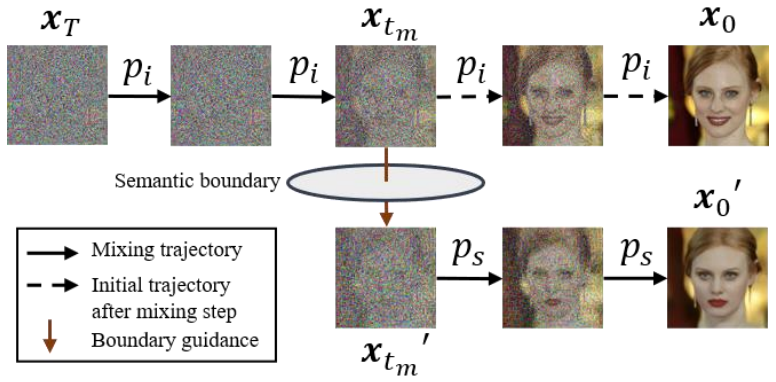
[WHZRW **ICCV'25 Highlight**]



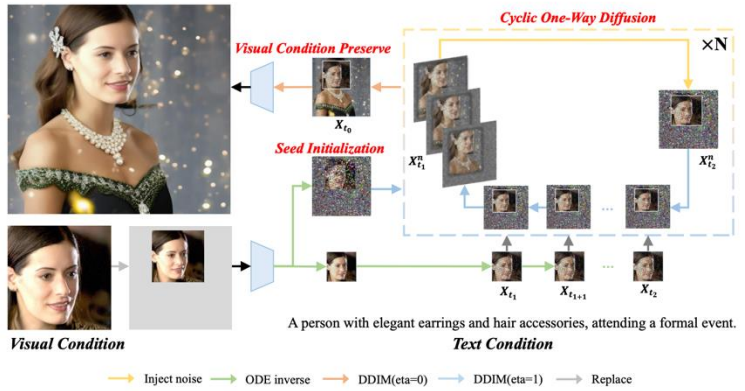
[ZMMWZ **NeurIPS'25**]

Dynamic Sampling for Interpretable Control in Multimodal Generation

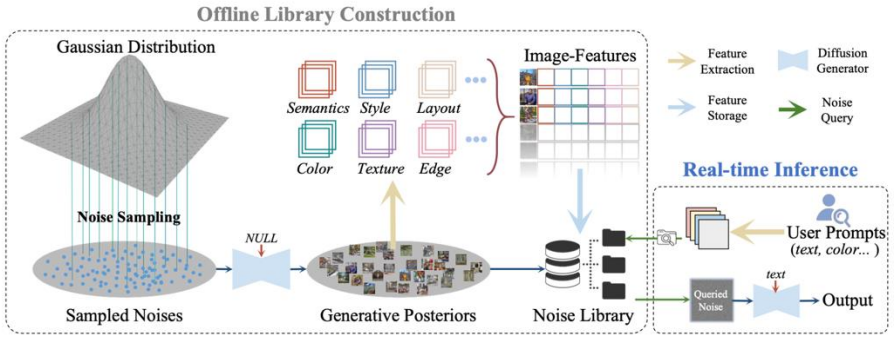
TL;DR: We prove the theoretical existence of mixing step in DMs, and propose a one-step intervention method in sampling to achieve fine-grained semantic control for data editing.



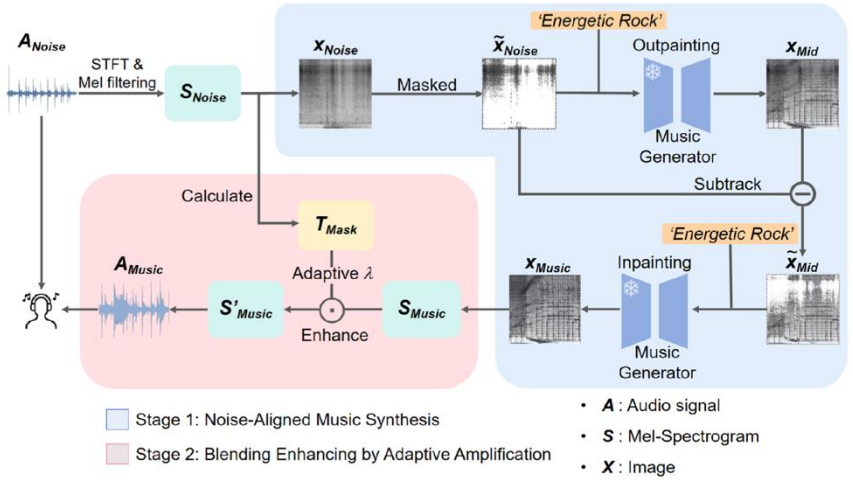
[ZWDRY **NeurIPS**'23]



[W*Y*QZ⁺W⁺ **ICLR**'24]



[WHZRW **ICCV'25 Highlight**]



[ZMMWZ **NeurIPS'25**]

Establish Markovian Study for Diffusion Models

Also “relaxation time” in thermodynamics

A transition from stochastic to (relatively) deterministic

Mixing Time in probability theory [*Markov chains and mixing times, 2017*]	Mixing Step in diffusion models
A parameter that measures the time required by a Markov chain for the distance to stationary to be small.	The critical diffusion step where the Gaussian distribution converges to the data distribution in the reverse denoising process.

***Boundary Guided Learning-Free Semantic Control with Diffusion Models*, NeurIPS**’23

Connection between Distribution Convergence and Data Semantics

Takeaway 1 :

Such mixing step theoretically exists in diffusion models, and can be empirically found in pre-trained models.

1. Define the distance measure

$$d^{(1)}(t) := \max_{x \in \mathcal{X}} ||\sigma_t(x, y) - 1||_1,$$

2. Formulation under DMs

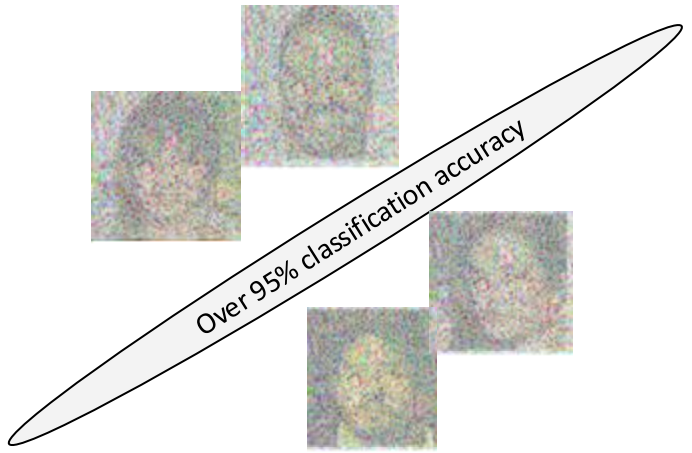
$$\sigma_t(x, y) = \frac{P^t(x, y)}{\pi(y)} = \frac{x \sim \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I)}{y \sim \mathcal{N}(0, I_d)}.$$

3. Deduce the mixing step

$$t^{(1)}_{mix}(\varepsilon) := \inf\{t \geq 0;\ d^{(1)}(t) \leq \varepsilon\}.$$

Takeaway 2 :

Such distribution properties establish close connection with common image semantics.



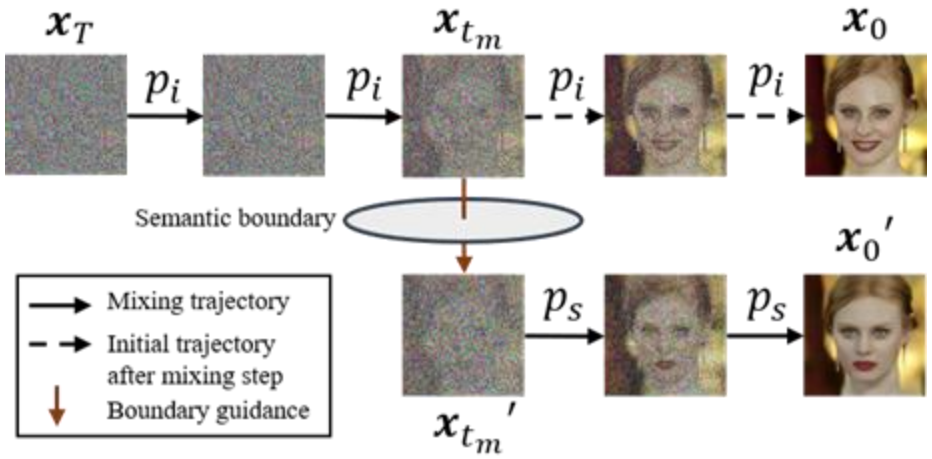
***Boundary Guided Learning-Free Semantic Control with Diffusion Models*, NeurIPS'23**

BoundaryDiffusion as a *Learning-Free*, Single-Step Controlling Method

Core methodological idea:

Adjust the latent spatial location relative to the identified hyperplane at the mixing step.

Single-step operation at the mixing step



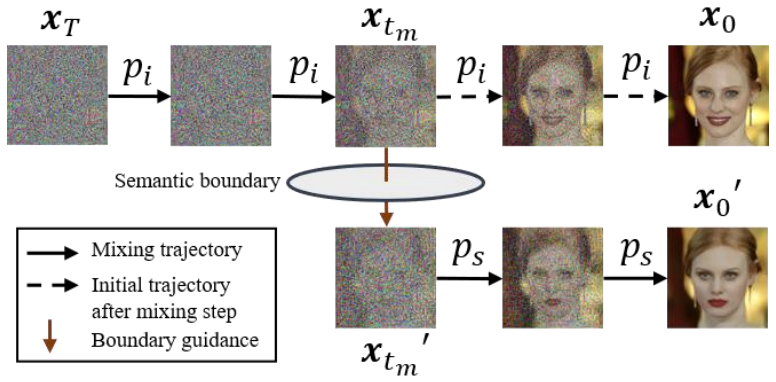
Resource budget

- *Data*: For each target attribute: about 100 images with labels
- *Time:* Inversion of raw images to the mixing step and fit a linear hyperplane (**negligible time**)
- *Compute:* Frozen model with **no tuning**

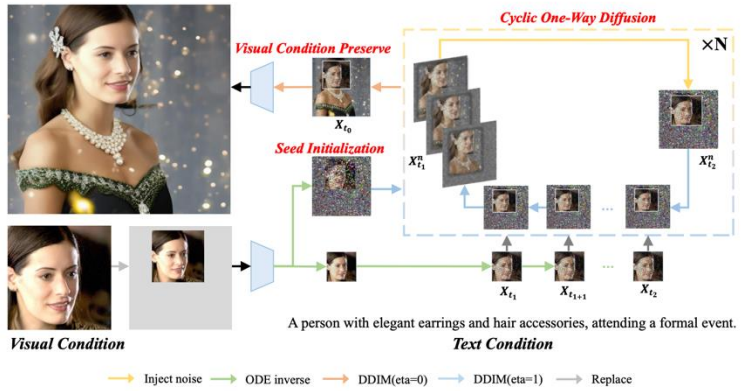
Boundary Guided Learning-Free Semantic Control with Diffusion Models, **NeurIPS**'23

Dynamic Sampling for Interpretable Control in Multimodal Generation

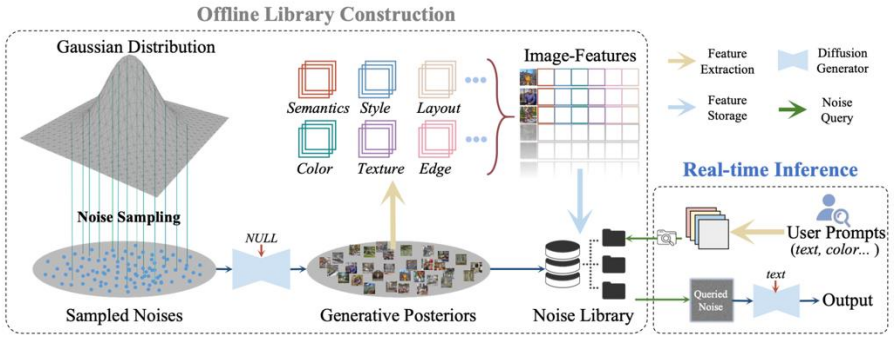
TL;DR: We disentangle the generation process into three phases based on the speed of information exchange, and propose a repetitive sampling strategy to achieve multimodal data customization.



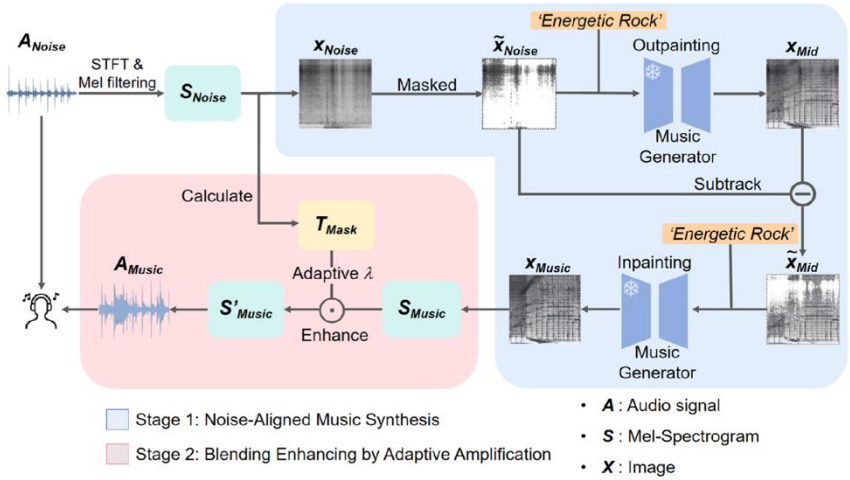
[ZWDRY **NeurIPS**'23]



[W*Y*QZ⁺W⁺ **ICLR**'24]

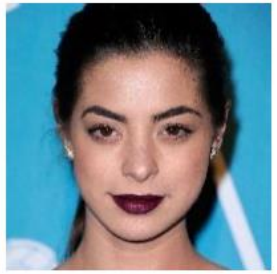


[WHZRW **ICCV'25 Highlight**]



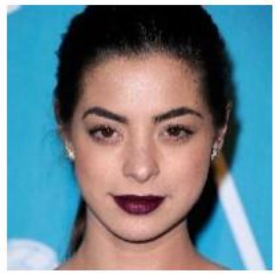
[ZMMWZ **NeurIPS'25**]

Task 1: Data Customization towards Personalized GenAI system

Input 1 - visual	Input 2 - text	Expectation
	*"A person holding a bread in kitchen."*	
	*"A kid with a flower crown."*	

Diffusion in Diffusion: Cyclic One-Way Diffusion for Text-Vision-Conditioned Generation, **ICLR**'24

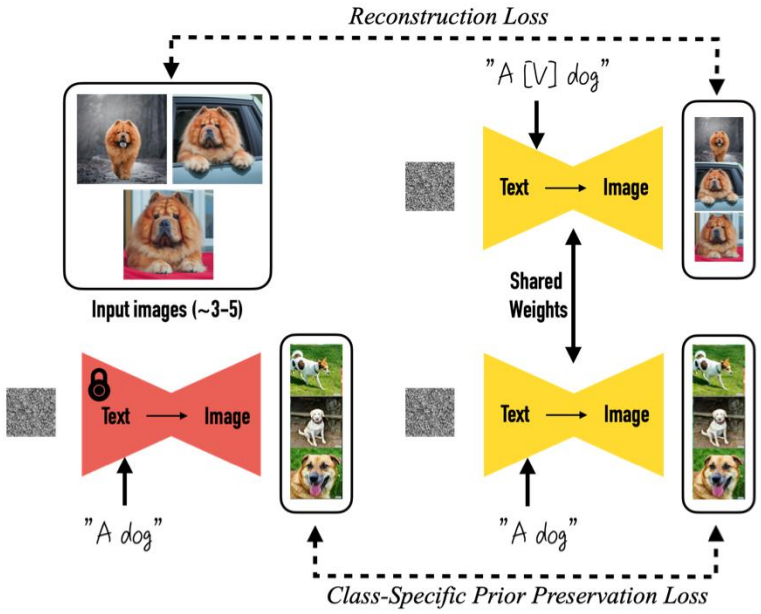
Task 1: Data Customization towards Personalized GenAI system

Input 1 - visual		Input 2 - text	Reality	Expectation
	✚	*"A person holding a bread in kitchen."*		
	✚	*"A kid with a flower crown."*		

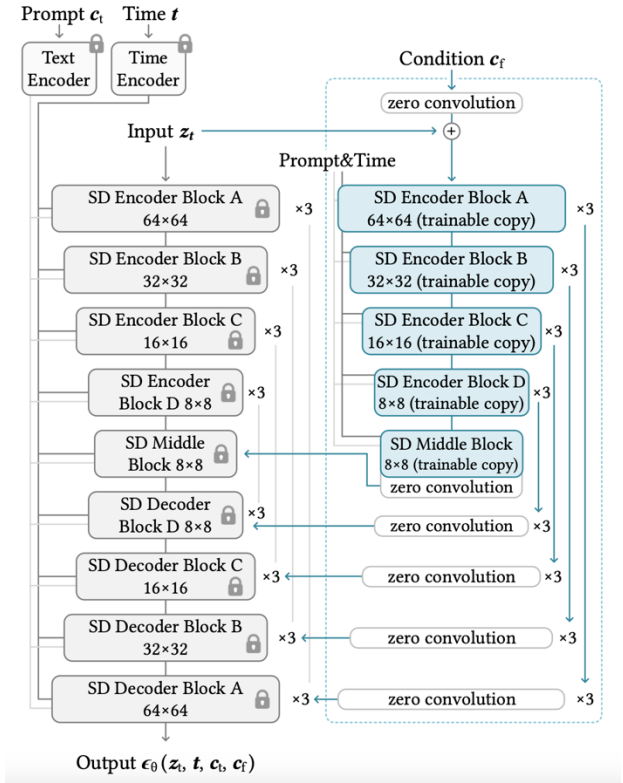
Diffusion in Diffusion: Cyclic One-Way Diffusion for Text-Vision-Conditioned Generation, **ICLR**'24

Research Question: How to Preserve the Low-level Visual Information?

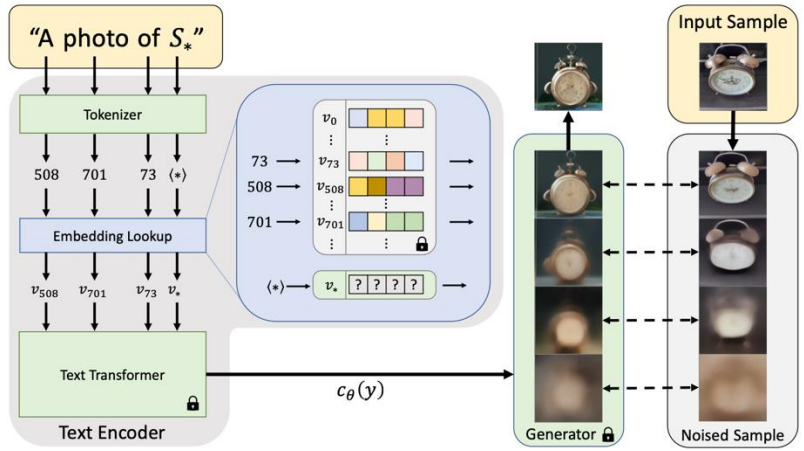
Mainstream paradigm: Tuning the pre-trained text-to-image models with extra supervisions (e.g., reconstruction)



[DreamBooth, CVPR'23]



[ControlNet, ICCV'23]



[TextInversion, ICLR'23]

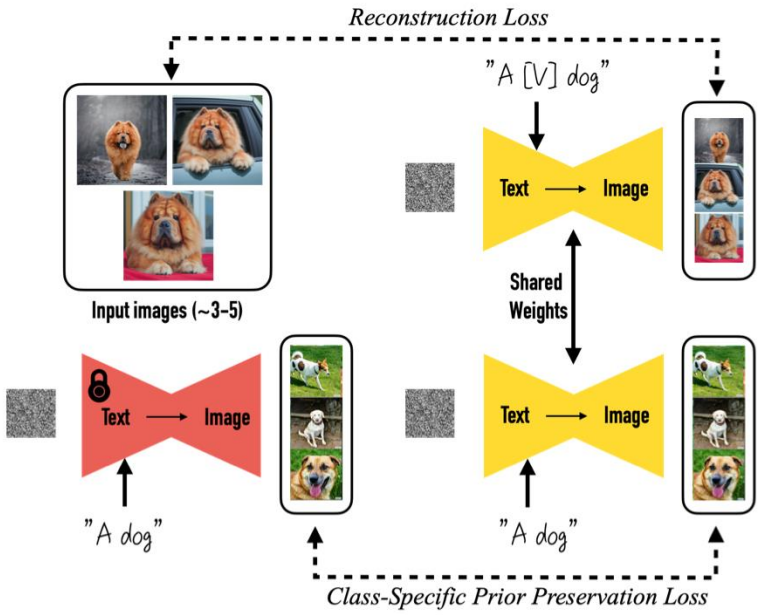
And many many more…

***Diffusion in Diffusion: Cyclic One-Way Diffusion for Text-Vision-Conditioned Generation*, ICLR'24**

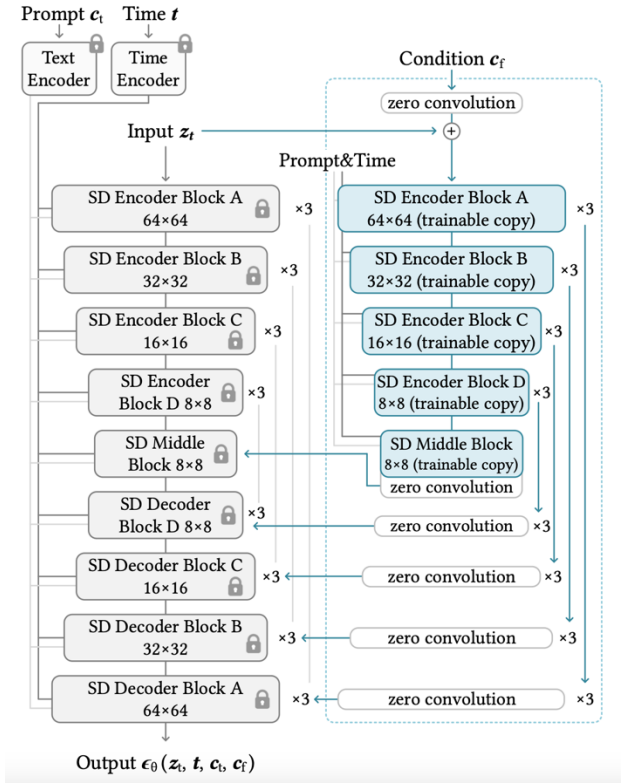
Research Question: How does Extra Condition Evolve in Latent Spaces?

~~**Research Question: How to Preserve the Low-level Visual Information?**~~

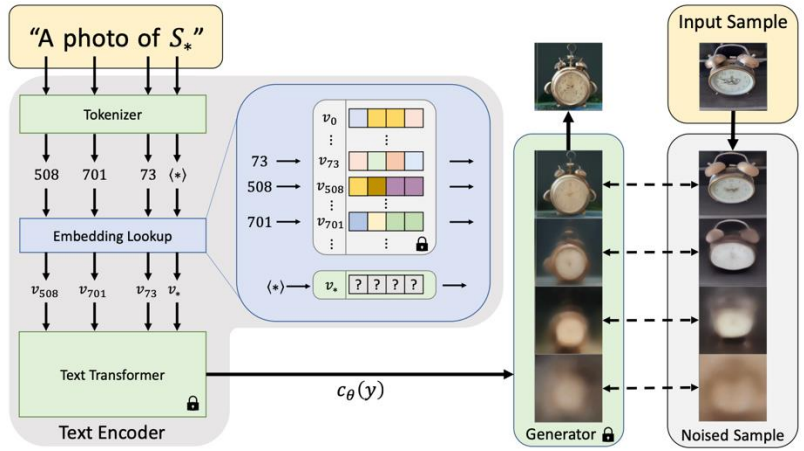
~~*Mainstream paradigm: Tuning the pre-trained text-to-image models with extra supervisions (e.g., reconstruction)*~~



[DreamBooth, CVPR'23]



[ControlNet, ICCV'23]

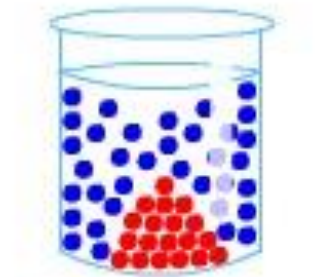


[TextInversion, ICLR'23]

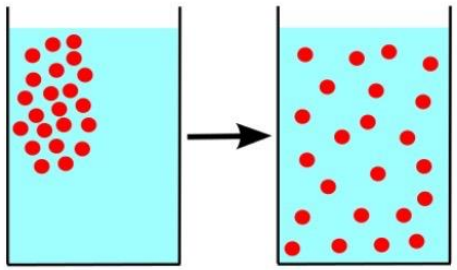
And many many more…

***Diffusion in Diffusion: Cyclic One-Way Diffusion for Text-Vision-Conditioned Generation*, ICLR'24**

Diffusion in Physics



PROCESS OF DIFFUSION



Diffusion: Particles moving from areas of high concentration to areas of low concentration.

[Physics: Temperature and Kinetic Theory]

Description:

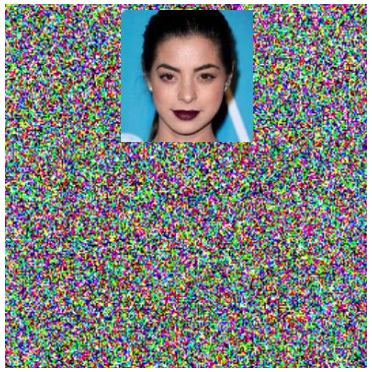
- Particles moving from areas of high concentration to areas of low concentration.

Properties:

- Driven by the concentration gradient, a higher gradient induces faster movement
- This movement is chaotic (no specific direction)

Diffusion in Diffusion: Cyclic One-Way Diffusion for Text-Vision-Conditioned Generation, **ICLR**'24

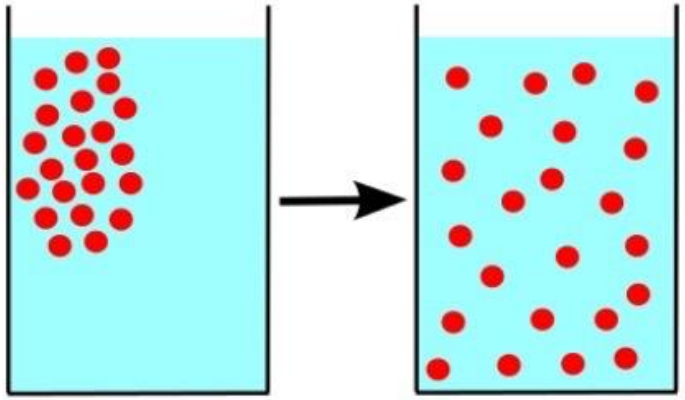
From Diffusion in Physics to Diffusion in Machine Learning



"A person holding a bread in kitchen."

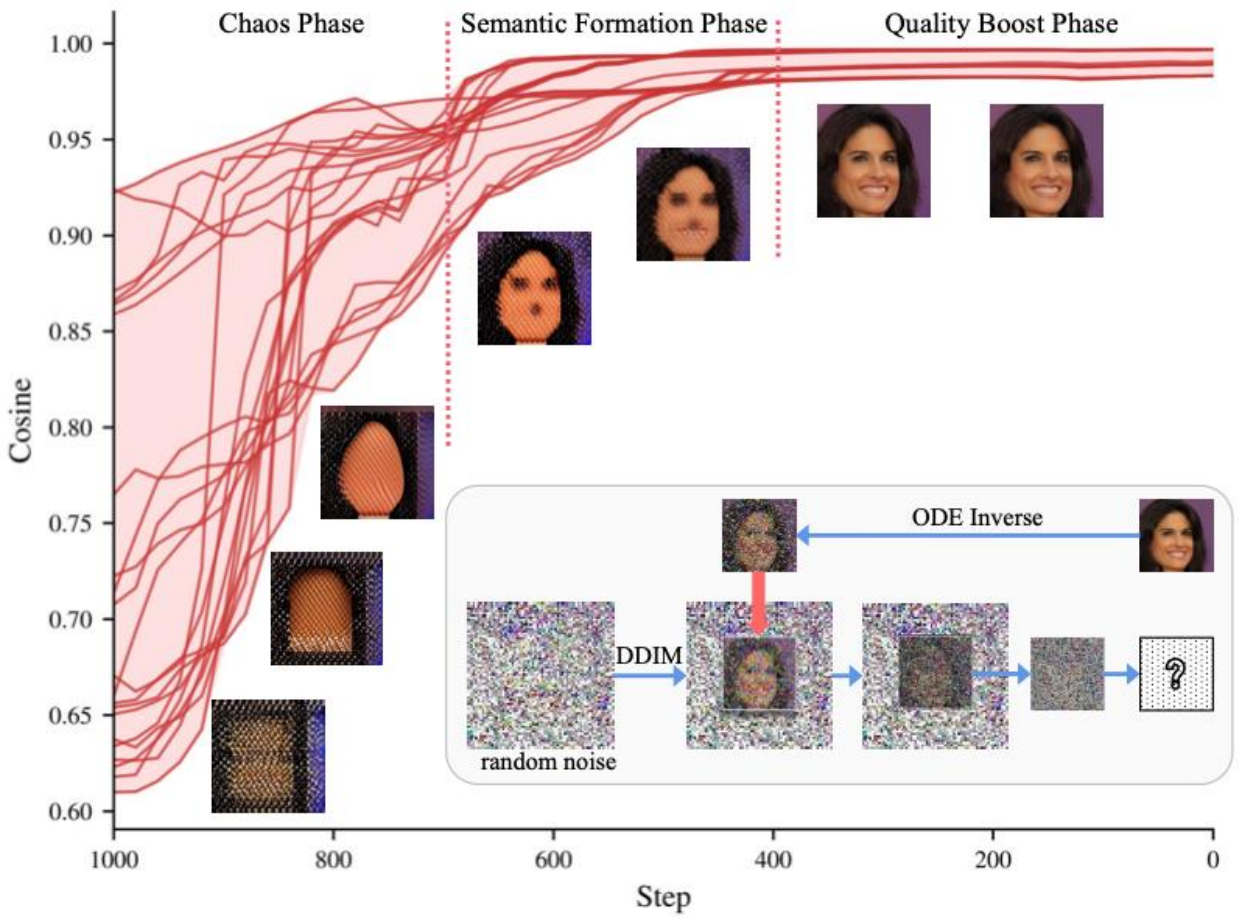
Information gradient

Diluted visual seeds



Diffusion in Diffusion: Cyclic One-Way Diffusion for Text-Vision-Conditioned Generation, **ICLR**'24

Speed of Visual Dilution along the Generative Process

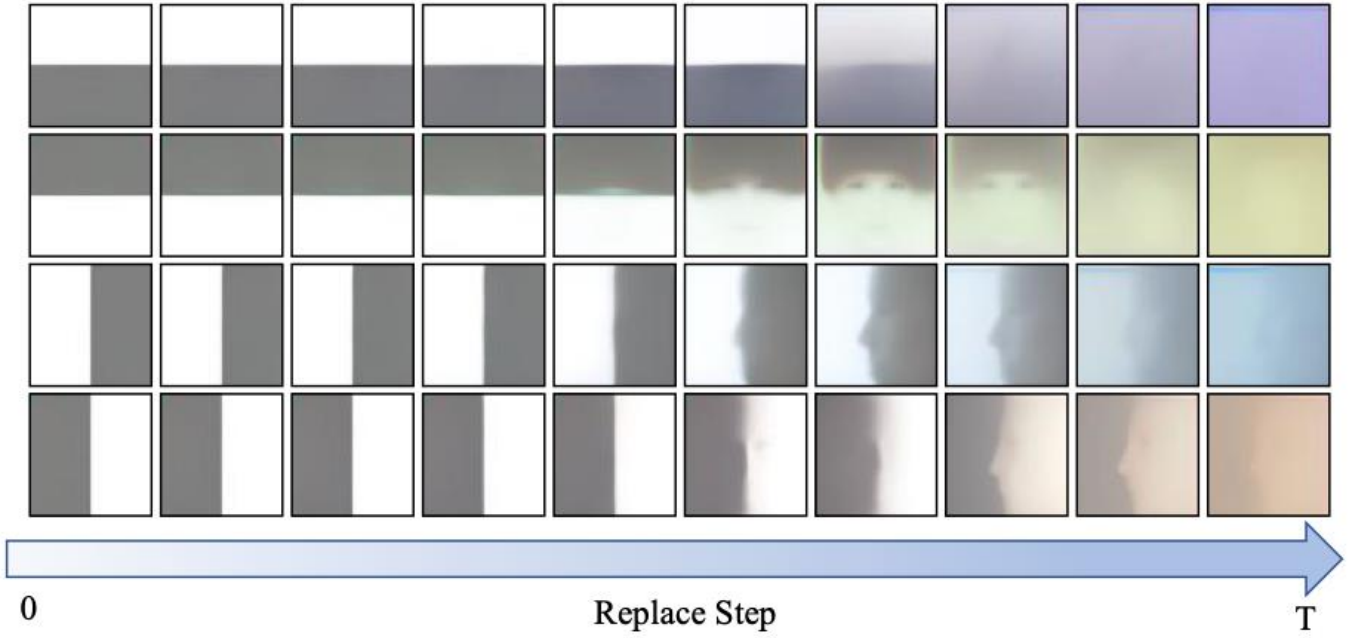


Takeaway:

The speed of information interference between visual seeds and noise mirrors physical diffusion, occurring in three distinct phases.

Diffusion in Diffusion: Cyclic One-Way Diffusion for Text-Vision-Conditioned Generation, **ICLR**'24

Direction of Visual Dilution along the Generative Process



Takeaway:

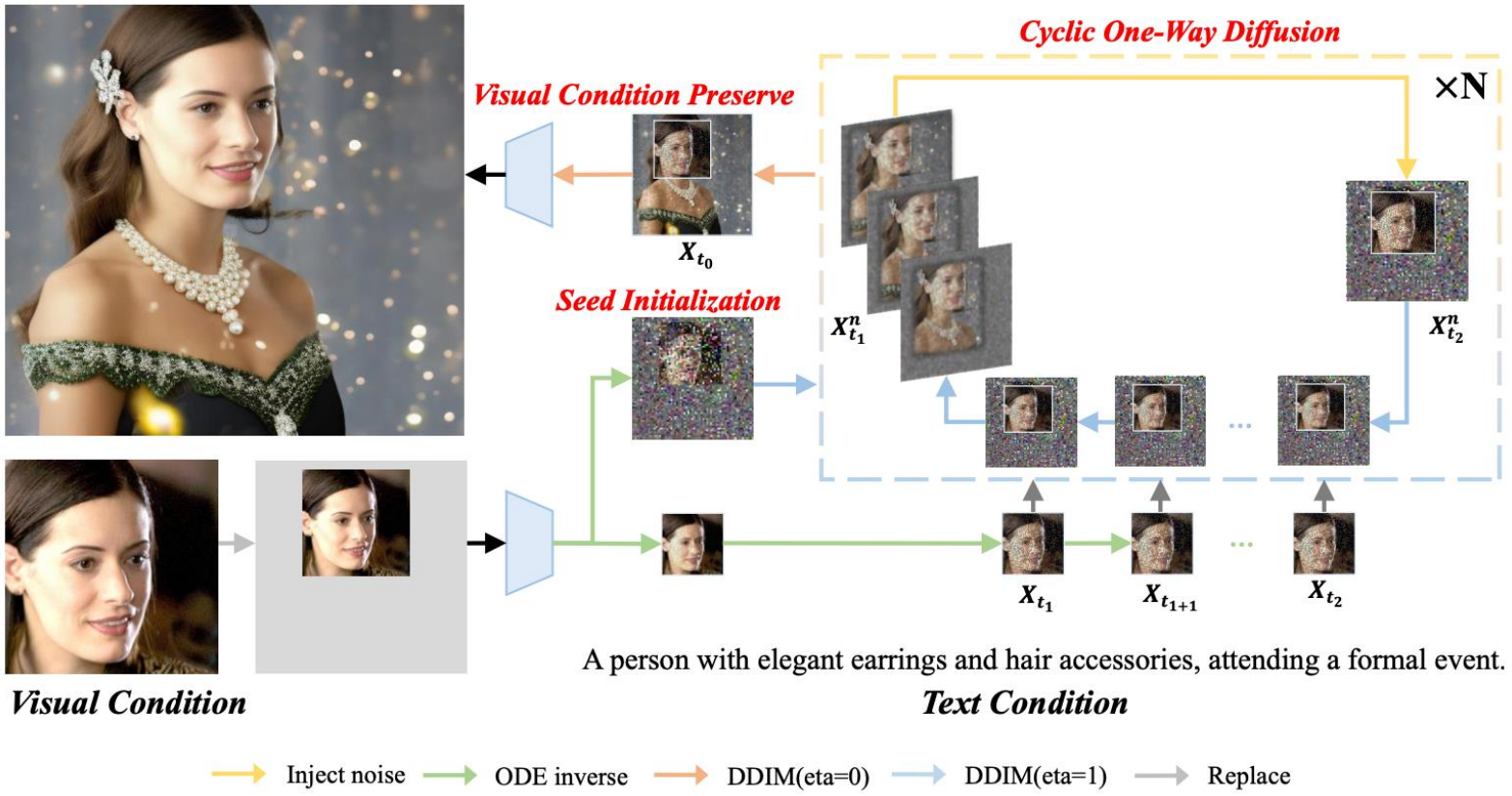
The direction of information interference is chaotic even without the concentration gradients.

Diffusion in Diffusion: Cyclic One-Way Diffusion for Text-Vision-Conditioned Generation, **ICLR**'24

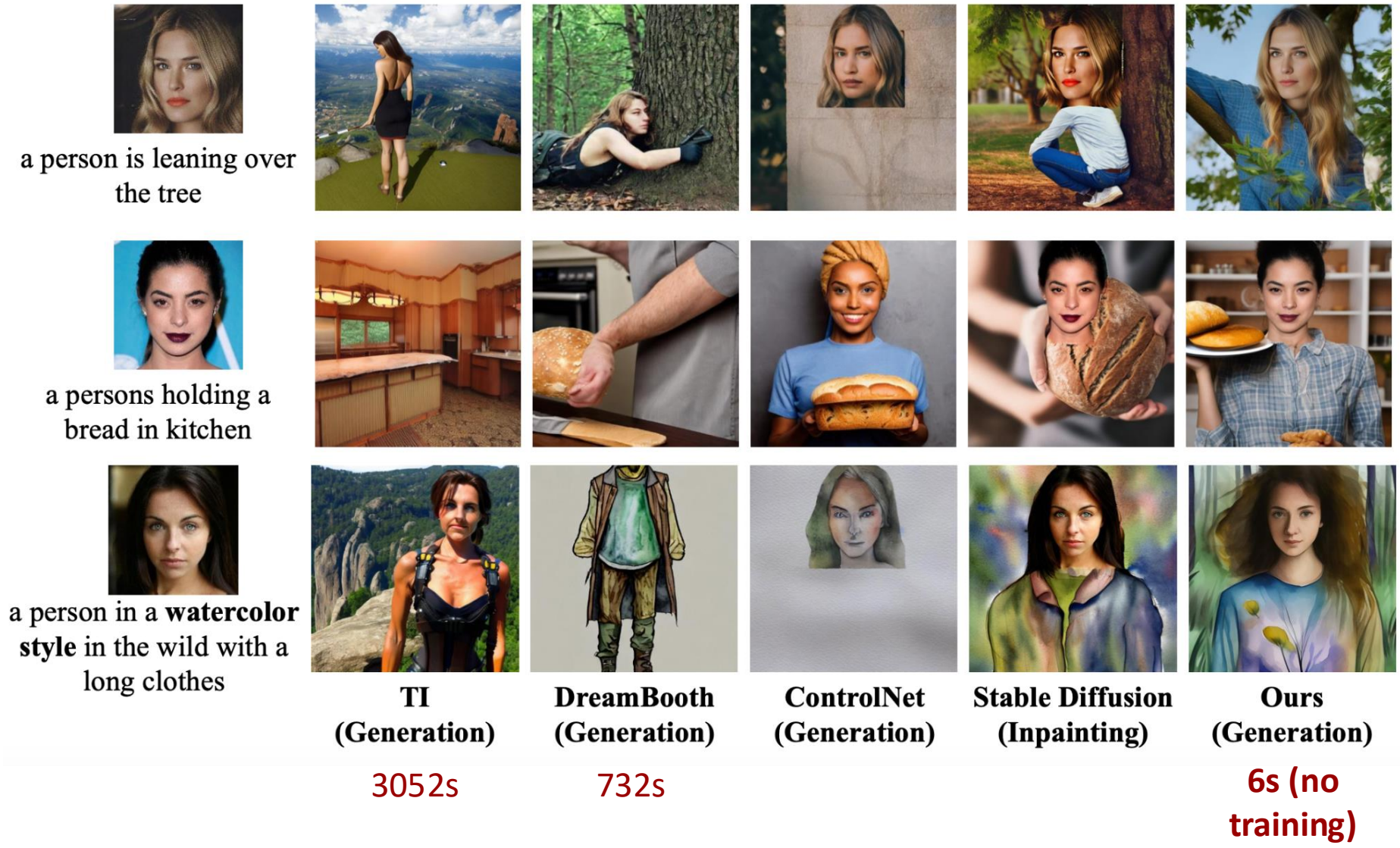
Learning-free Cyclic One-Way (COW) Diffusion for Data Customization

Core methodological idea:

Repeatedly drop the same "ink" (visual seed) into the same location in the "water bottle" (expanded image) during the semantic formation phase.



Improved Visual Preservation Compared to Learning-based SOTA



Resource budget

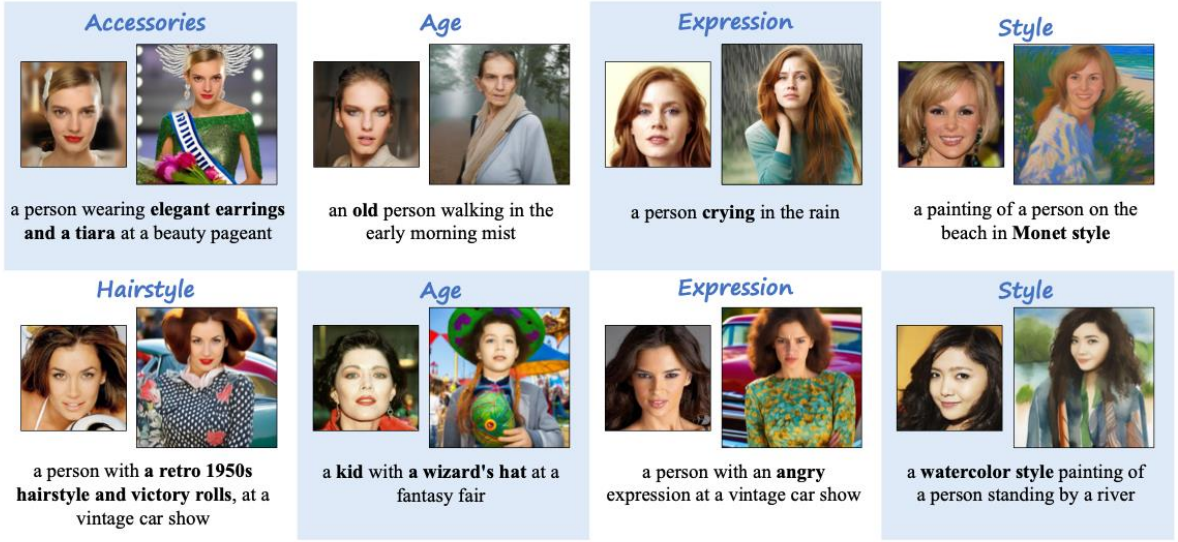
- Frozen text-to-image generative model with **no learning**
- No extra data

Conditional consistency w.r.t. To text prompt: 69.73% (58.13% higher)

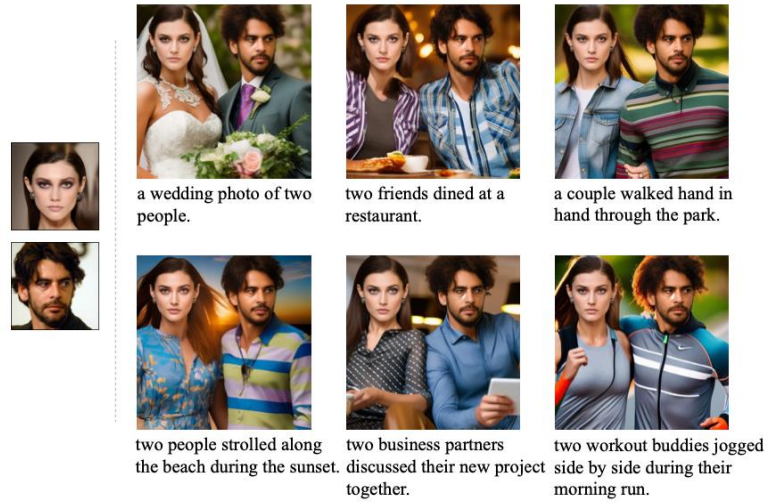
General fidelity: 51.87% (23.14% higher)

Generalization to Handle Multiple Conditions across Versatile Scenarios

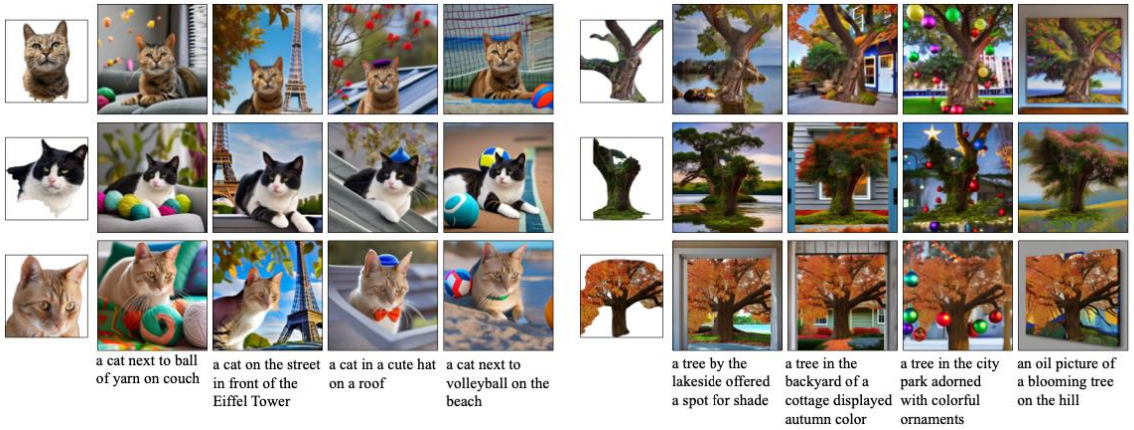
Different semantic attributes



Multiple subjects

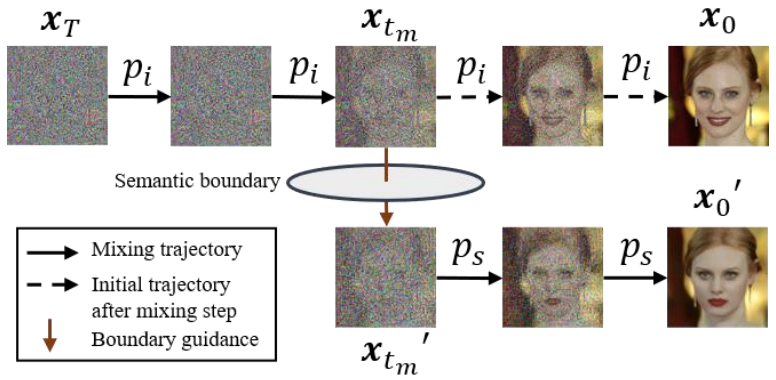


Different subjects

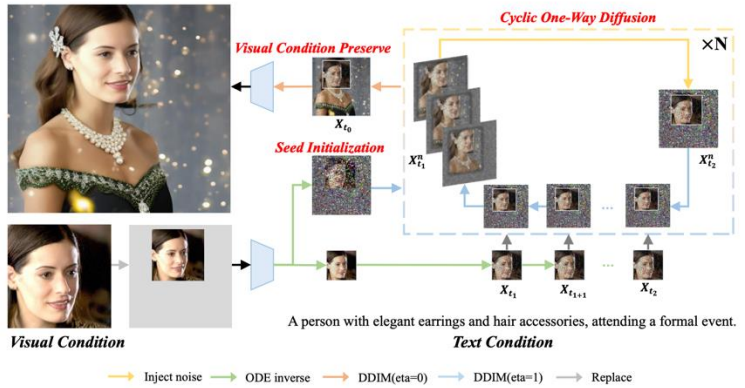


Diffusion in Diffusion: Cyclic One-Way Diffusion for Text-Vision-Conditioned Generation, **ICLR**'24

Dynamic Sampling for Interpretable Control in Multimodal Generation

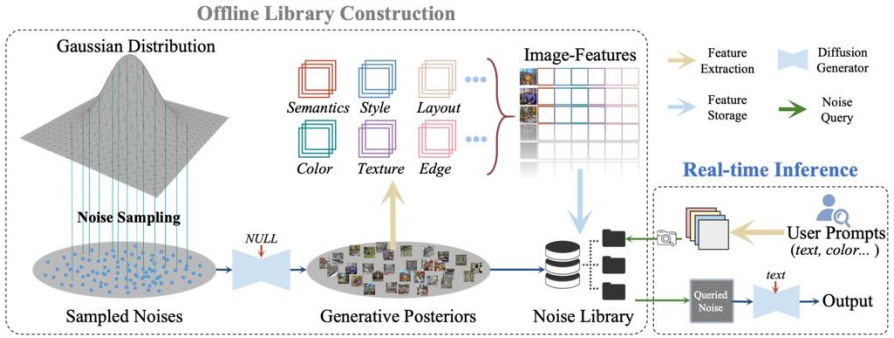


[ZWDRY **NeurIPS**'23]

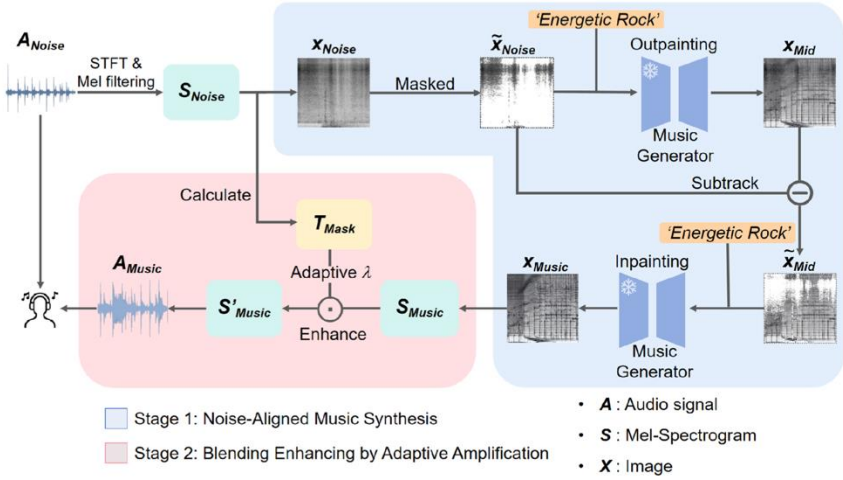


[W*Y*QZ+W+ **ICLR**'24]

TL;DR: We reveal one generic asymptotic behavior in modern diffusion T2I models between noises and generated output, and propose to achieve better alignment and control by finding the optimal initial noise.



[WHZRW **ICCV'25 Highlight**]



[ZMMWZ **NeurIPS'25**]

Task 2: Enhanced Input-Output Alignment in Text-to-Image Models

Input - text

A fluffy baby **sloth** with a **knitted hat** trying to figure out a **laptop**.

Medieval castle on a hill;

High contrast;

High Sharpness;

Expectation

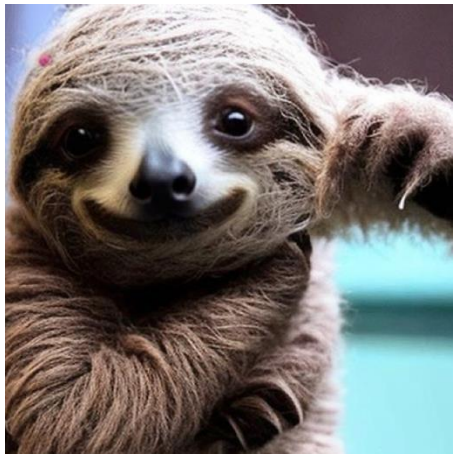
The Silent Assistant: NoiseQuery as Implicit Guidance for Goal-Driven Image Generation, ICCV'25 **Highlight**

Task 2: Enhanced Input-Output Alignment in Text-to-Image Models

Input - text

A fluffy baby **sloth** with a **knitted hat** trying to figure out a **laptop**.

Reality



Expectation

Medieval castle on a hill;

High contrast;

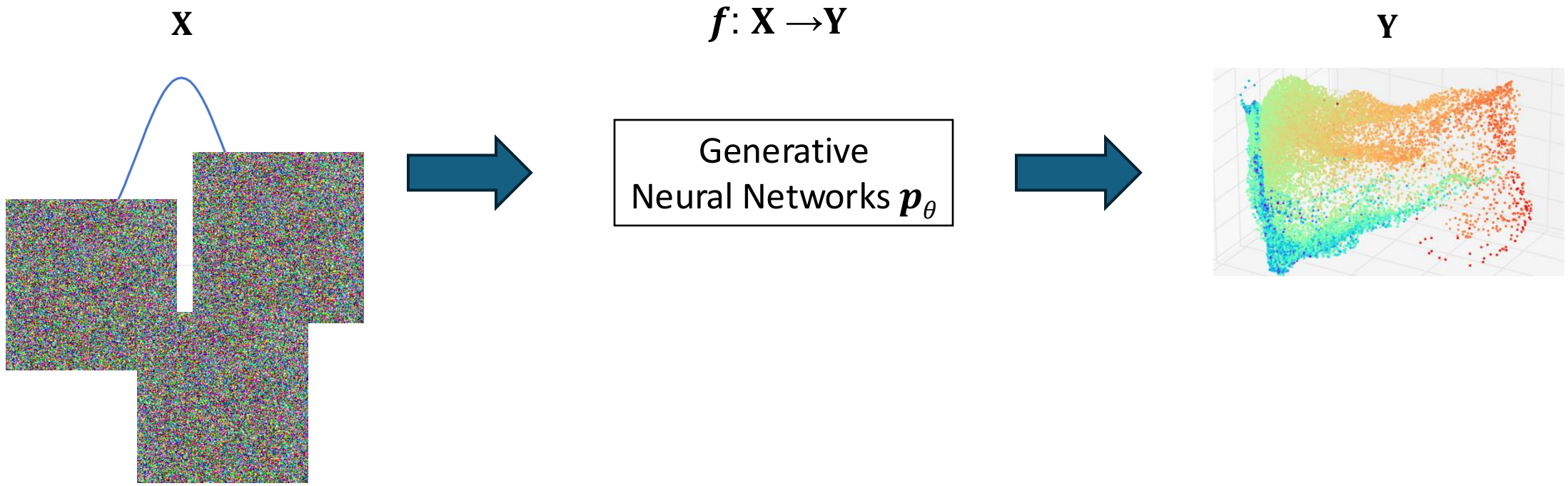
High Sharpness;

The Silent Assistant: NoiseQuery as Implicit Guidance for Goal-Driven Image Generation, ICCV'25 **Highlight**

Research Question: How to Enhance Alignment and Controllability?

Takeaway: This is fundamental research question in conditional generation, and there are many ways to address this.

Cross-modality attention, probabilistic enhancement (my ICLR'23 contrastive diffusion paper), auxiliary conditions…

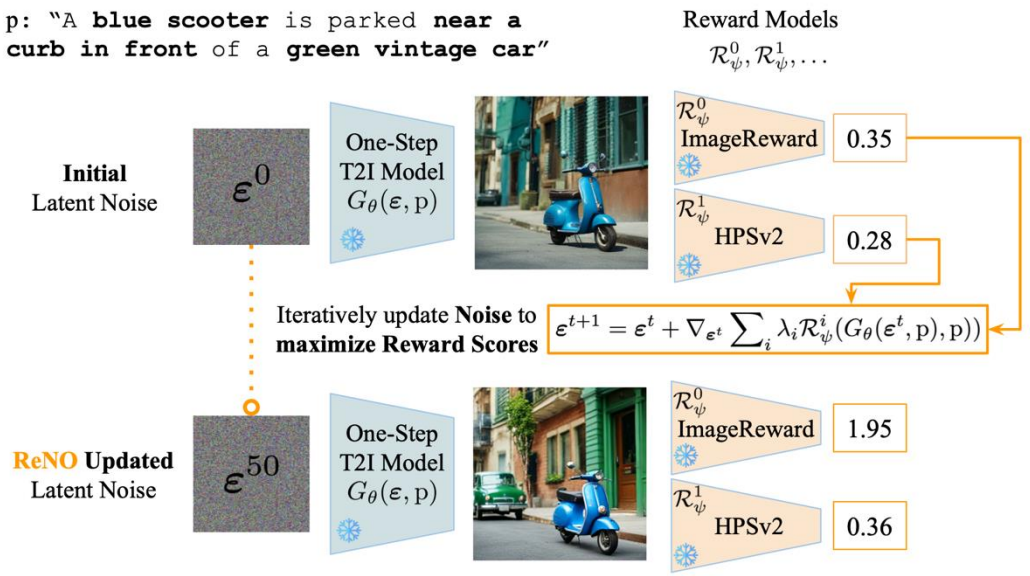


***The Silent Assistant: NoiseQuery as Implicit Guidance for Goal-Driven Image Generation*, ICCV'25 Highlight**

What Role does Initial Noise Play in T2I Diffusion Models?

Previous literature: Initial noise plays an important role in T2I generation

- ReNO (NeurIPS 2024): gradient-based noise optimization
- DPO (CVPR 2025): direct reward-based fine-tuning
- CFG++ (arXiv 2024): CFG guidance refinement
- LaVi-Bridge (ECCV 2024): text encoding enhancement via LoRA adapter

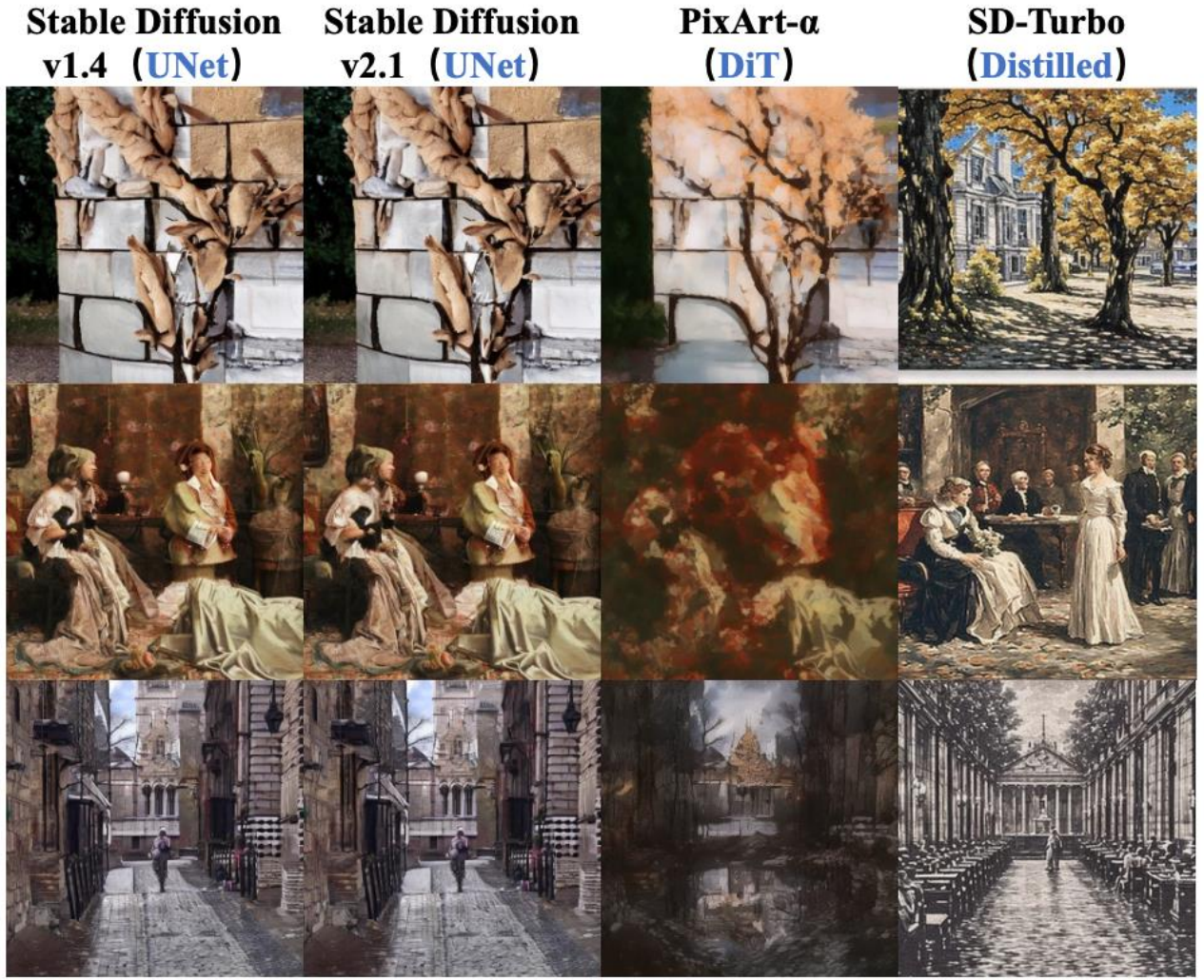


[ReNO, NeurIPS'24]

The Silent Assistant: NoiseQuery as Implicit Guidance for Goal-Driven Image Generation, **ICCV**'25 **Highlight**

Our Key Observation: Consistent Noise Impact across T2I Models

Takeaway: the impact of noise initialization persists across a wide range of T2I diffusion models, regardless of their backbone architectures (Unet or DiT)



The Silent Assistant: NoiseQuery as Implicit Guidance for Goal-Driven Image Generation, ICCV'25 **Highlight**

How to Interpret the Source of those Observations?

Original diffusion formulation: forward process

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}),$$

Re-parameterization tricks: $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\epsilon \sim \mathcal{N}(0, I)$

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,$$

Particularly, in practical implementations, people like to take few-step samplers for better inference efficiency, which even further enhances this shortcut connection

The Silent Assistant: NoiseQuery as Implicit Guidance for Goal-Driven Image Generation, ICCV'25 **Highlight**

Direct Visualization of Such Noise Impact

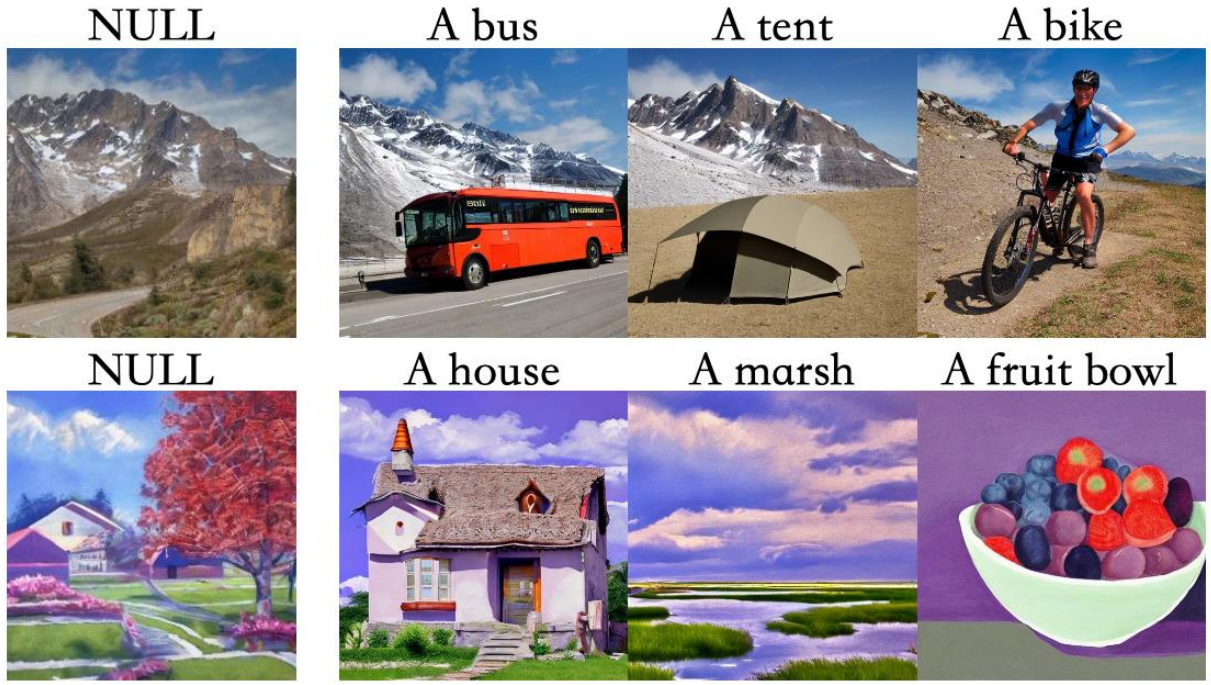


Figure 3. Using the same initial noise, the generative posteriors (left) exhibit similarity with the text-conditioned images (right).

Generative posteriors:
The output from unconditional generation with NULL text.

NoiseQuery: The Silent Assistant for Enhanced Noise Initialization

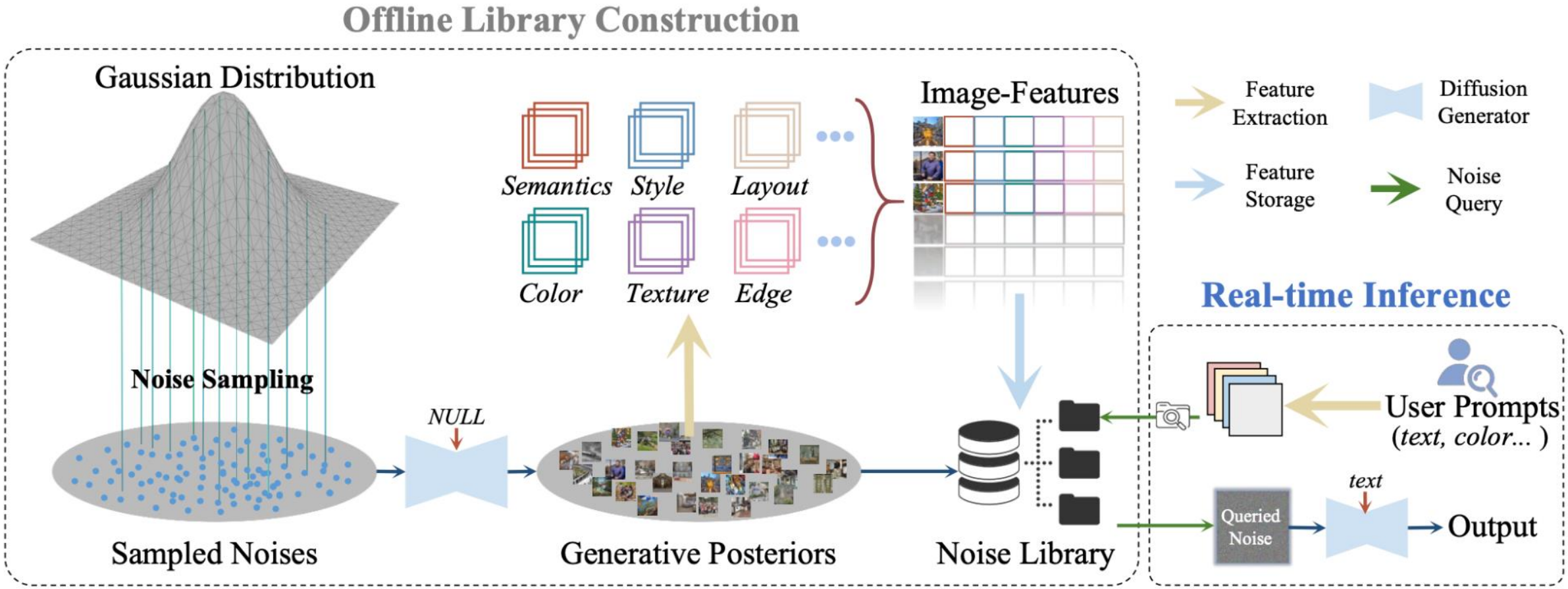


Figure 4. Illustration of our *NoiseQuery* pipeline which involves offline library construction and real-time inference.

The Silent Assistant: NoiseQuery as Implicit Guidance for Goal-Driven Image Generation, **ICCV**'25 **Highlight**

NoiseQuery: Technical Breakdown – Offline Library Construction

Offline Library Construction

foreach $\epsilon_i \in \mathcal{N}_{set}$ **do**

$\mathcal{I}_i^{\text{uncond}} = \mathcal{M}(\epsilon_i, c = \emptyset)$;

Extract features $\mathcal{F}_i$ (as detailed in Tab. 1) from $\mathcal{I}_i^{\text{uncond}}$;

Store $(\epsilon_i, \mathcal{F}_i)$ in the noise library.

end

Generation Goals	Feature Type	Match Function
Semantics	CLIP [37], BLIP [26]	Cosine Similarity
Style	Gram Matrix [14]	MSE
Color	RGB, HSV, LAB	Absolute Difference
Texture	GLCM [17]	Euclidean Distance
Shape	Hu Moments [20]	Euclidean Distance
Sharpness	High Frequency Energy (HFE)	Absolute Difference

Table 1. Feature types and matching functions.

The Silent Assistant: NoiseQuery as Implicit Guidance for Goal-Driven Image Generation, ICCV'25 **Highlight**

NoiseQuery: Technical Breakdown – Real-Time Inference

Real-time Inference

Extract desired features $\mathcal{F}_O$ from objective $\mathcal{O}$;

$\epsilon^* = \arg\max_{\epsilon_i \in \mathcal{N}_{set}} \{S(\mathcal{F}_i, \mathcal{F}_O) \mid \epsilon_i \in \mathcal{N}_{set}\}$.

Library Size	0.5k	1k	2k	5k	10k	50k	100k
Matching Function Cost ($\times 10^{-4}$ s)	1.39	1.39	1.39	1.39	1.40	1.40	1.51
Argmax Selection Cost ($\times 10^{-4}$ s)	0.25	0.25	0.25	0.25	0.37	6.16	13.25
CLIP Score	31.51	31.53	31.57	31.59	31.66	31.73	31.74

Table 3. Retrieval time breakdown across various library sizes.

The Silent Assistant: NoiseQuery as Implicit Guidance for Goal-Driven Image Generation, ICCV'25 Highlight

Improvement on High-Level Semantics

Base Model	Method	DrawBench [42]				MSCOCO [30]				Time Cost
		ImageReward	PickScore	HPS v2	CLIPScore	ImageReward	PickScore	HPS v2	CLIPScore	
SD 1.5	Base Model	0.04	21.11	24.57	30.90	0.15	21.41	25.65	31.08	1.334 s
	+ NoiseQuery	**0.08**	**21.16**	**25.02**	**31.41**	**0.27**	**21.48**	**26.07**	**31.47**	1.336 s
	+ Diffusion-DPO [48]	0.09	21.29	25.02	31.19	0.25	21.64	26.31	31.26	1.350 s
	+ Diffusion-DPO [48] + NoiseQuery	**0.17**	**21.33**	**25.25**	**31.41**	**0.35**	**21.68**	**26.60**	**31.55**	1.352 s
SD 2.1	Base Model	0.12	21.33	24.93	31.13	0.36	21.72	26.58	31.40	1.301 s
	+ NoiseQuery	**0.26**	**21.46**	**25.39**	**31.68**	**0.44**	**21.76**	**26.82**	**31.50**	1.303 s
	+ CFG++ [8]	0.12	21.33	24.83	31.13	0.37	21.72	26.66	31.31	3.724 s
	+ CFG++ [8] + NoiseQuery	**0.27**	**21.43**	**25.55**	**31.61**	**0.47**	**21.76**	**26.97**	**31.67**	3.726 s
SD-Turbo	Base Model	0.26	21.78	25.23	31.29	0.47	22.07	26.22	31.51	0.072 s
	+ NoiseQuery	**0.41**	**21.87**	**25.66**	**31.58**	**0.50**	**22.17**	**26.82**	**31.76**	0.074 s
	+ ReNO [10]	1.67	23.40	32.48	32.55	-	-	-	-	23.56 s
	+ ReNO [10] + NoiseQuery	**1.71**	**23.52**	**32.92**	**32.78**	-	-	-	-	23.56 s
PixArt-α	Base Model	0.70	22.08	28.27	30.83	0.78	22.24	29.33	31.48	4.327 s
	+ NoiseQuery	**0.82**	**22.11**	**28.45**	**31.27**	**0.79**	**22.33**	**29.56**	**31.64**	4.328 s
	+ LaVi-Bridge [60]	0.63	22.08	28.35	30.92	0.75	22.31	29.49	31.86	5.092 s
	+ LaVi-Bridge [60] + NoiseQuery	**0.72**	**22.24**	**28.61**	**31.35**	**0.78**	**22.35**	**29.67**	**32.01**	5.094 s

Table 2. Evaluation of objective metrics on different datasets and models. Higher is better for all metrics. Our approach enhances the base model's performance and complements a wide range of T2I enhancement methods, including DPO [48] (reward-based fine-tuning), CFG++ [8] (guidance refinement), ReNO [10] (gradient-based initial noise optimization), and LaVi-Bridge [60] (text encoding enhancement via LORA/adapters). '-': ReNO results on MSCOCO are excluded due to excessive inference time.

The Silent Assistant: NoiseQuery as Implicit Guidance for Goal-Driven Image Generation, **ICCV**'25 **Highlight**

Improvement on High-Level Semantics

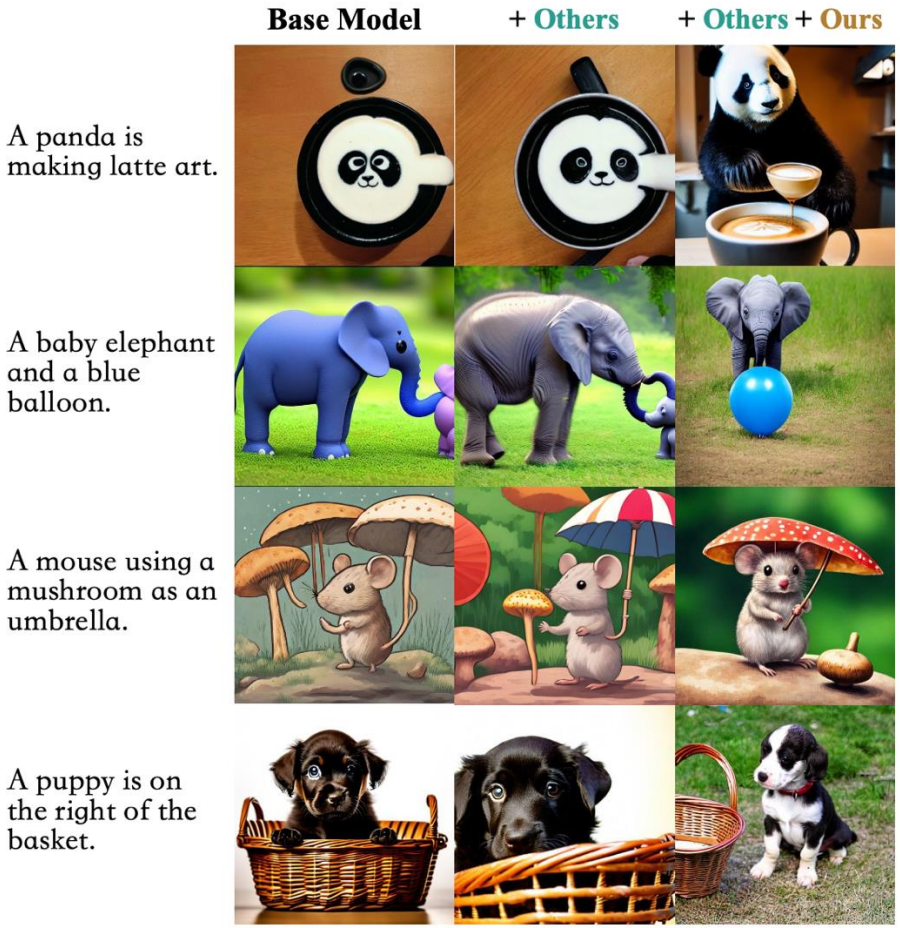


Figure 5. Generated images of the base model, base model with T2I enhancement methods, and base model with enhancement and *NoiseQuery*. The four rows, from top to bottom, respectively correspond to the method groups listed in Table 2: Diffusion-DPO [48], CFG++ [8], ReNO [10], and LaVi-Bridge [60].

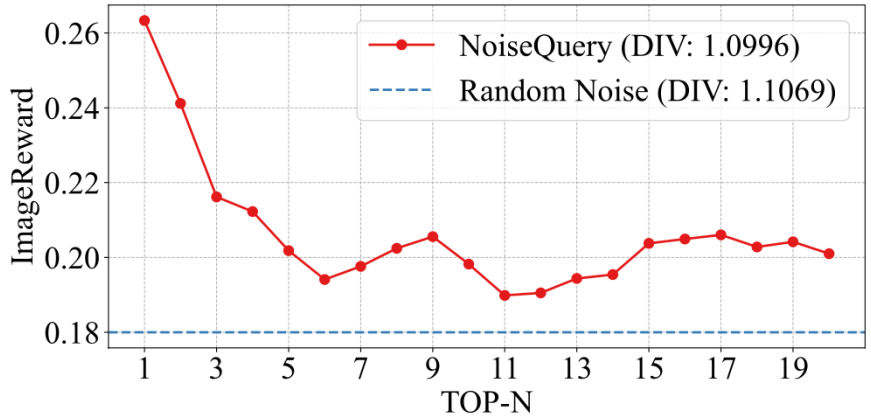


Figure 6. Queried noise v.s. random noise.

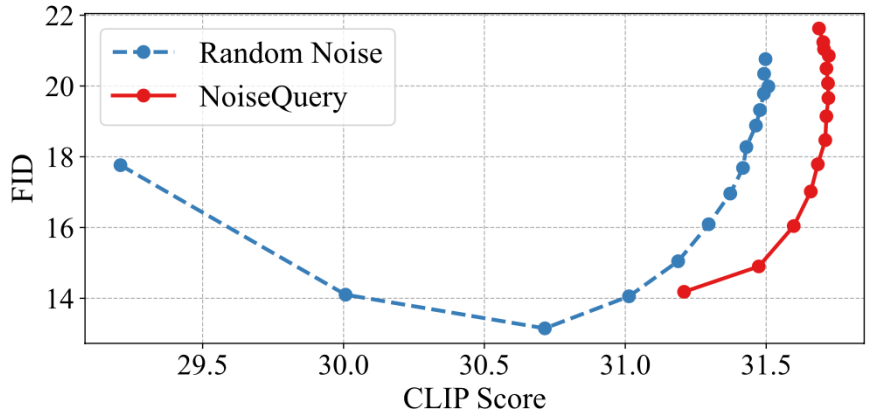
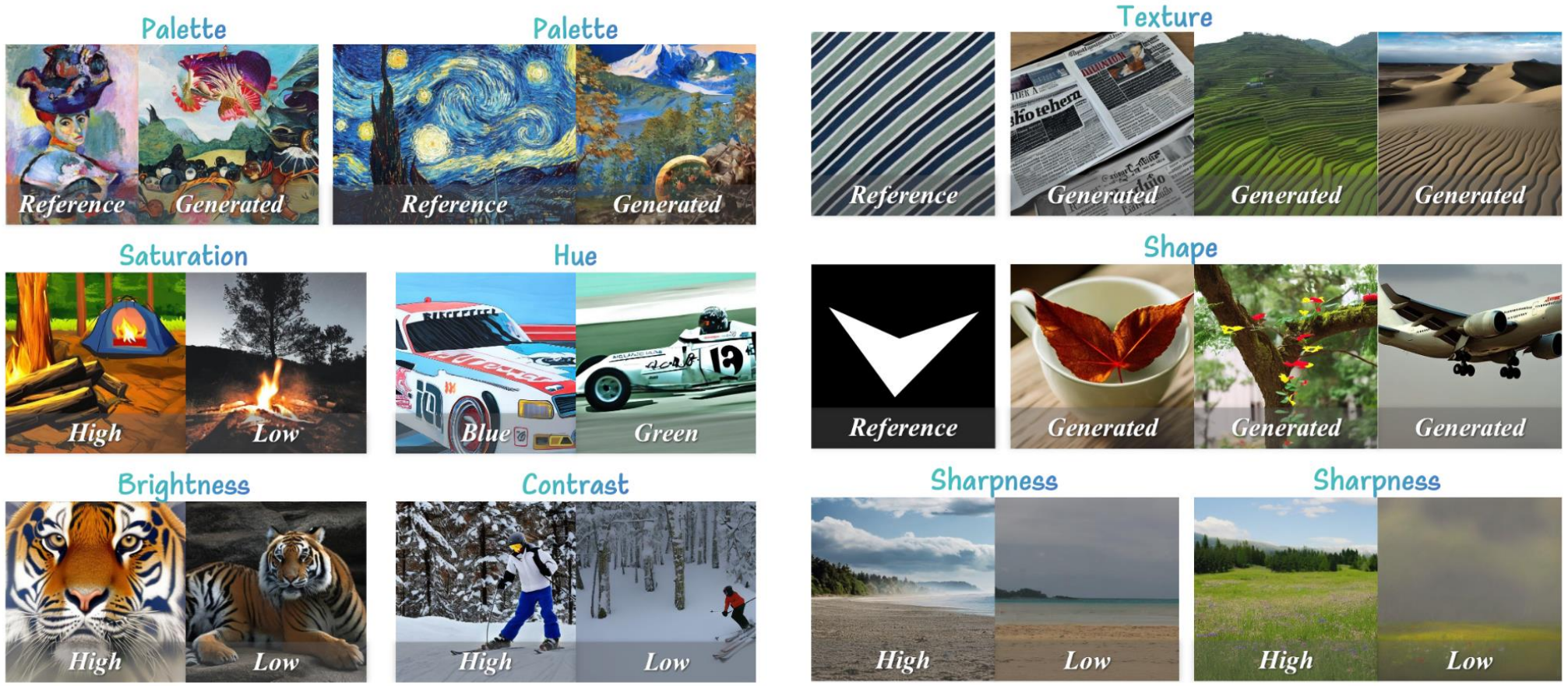


Figure 7. CLIP and FID evaluations across CFG scales on MSCOCO [30], with the CFG scale increasing from left to right.

***The Silent Assistant: NoiseQuery as Implicit Guidance for Goal-Driven Image Generation*, ICCV'25 Highlight**

Controllability on Low-Level Visual Properties



(a) Color properties guidance.

(b) Structural features guidance.

Figure 8. Multiple low-level visual properties control through *NoiseQuery*. **All images use minimal text prompts containing only semantic concepts (e.g., object labels), lacking explicit visual specifications.**

The Silent Assistant: NoiseQuery as Implicit Guidance for Goal-Driven Image Generation, **ICCV**'25 **Highlight**

Controllability on Low-Level Visual Properties

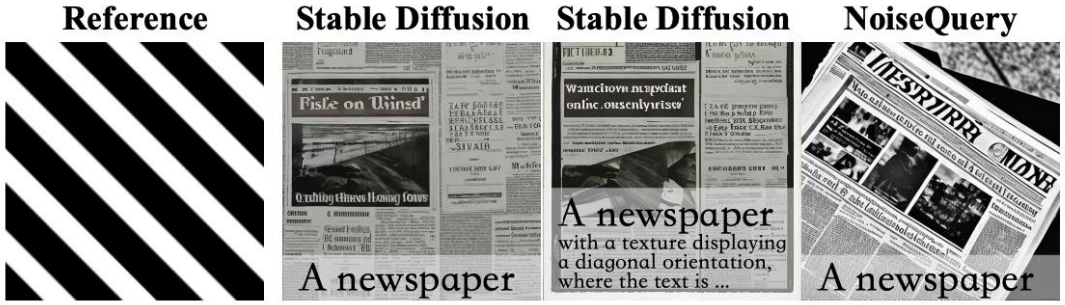
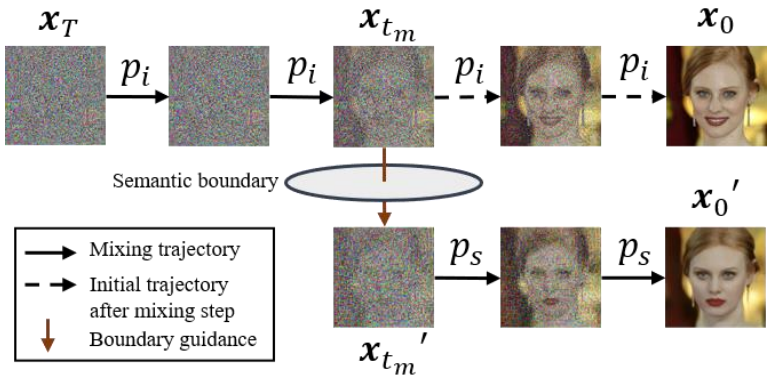


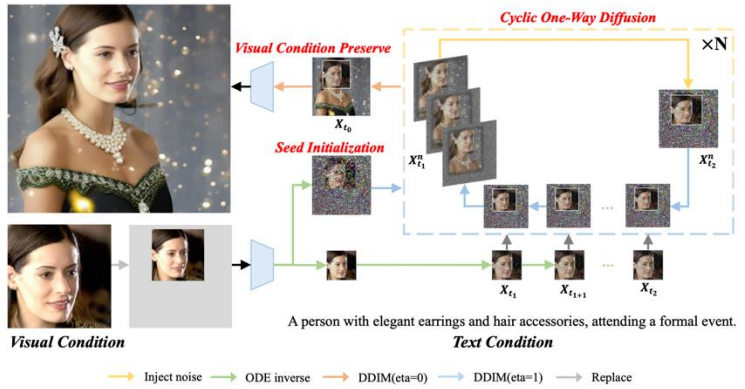
Figure 9. Text prompts often fail to align generated images with user low-level attribute preferences, while initial noise plays a dominant role in controlling them.

Figure 10. Unlike the original Stable Diffusion , which produces images with medium brightness, our *NoiseQuery* (offset) expands the range to include both very bright and very dark samples.

Dynamic Sampling for Interpretable Control in Multimodal Generation

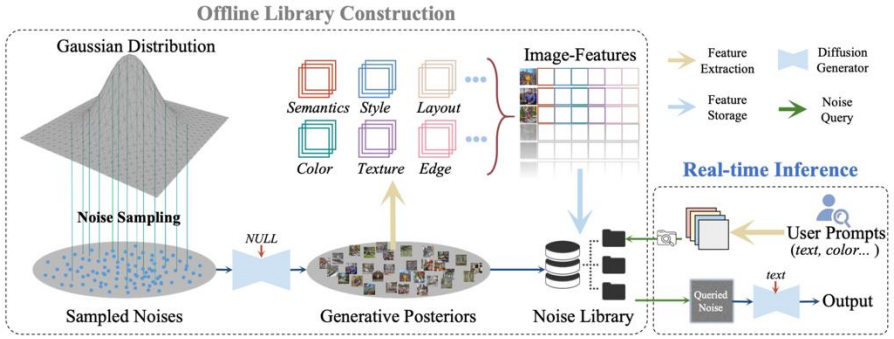


[ZWDRY **NeurIPS**'23]

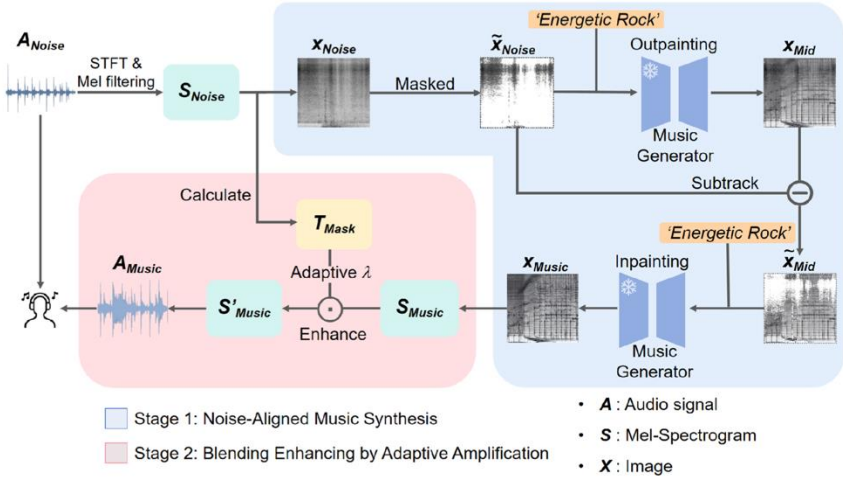


[W*Y*Q**Z**+W+ **ICLR**'24]

TL;DR: We extend the structural sampling control from text-to-visual to text-to-music generation, propose to achieve acoustic masking that generates customized music and blends noises.



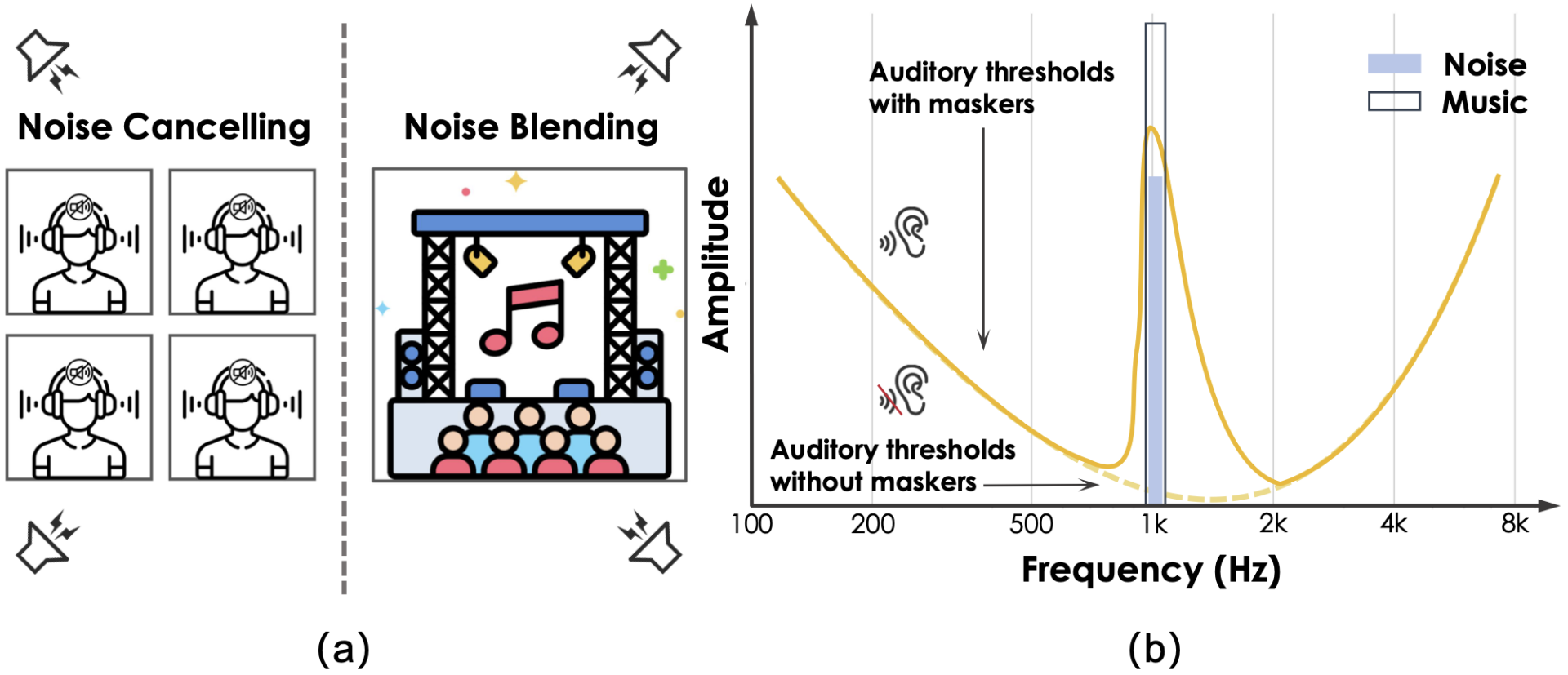
[WHZRW **ICCV'25 Highlight**]



[ZMMW**Z** **NeurIPS'25**]

Task 3: Noise Masking through Text Guided Music Generation

Auditory Masking: the perception of one sound is affected by the presence of another sound.



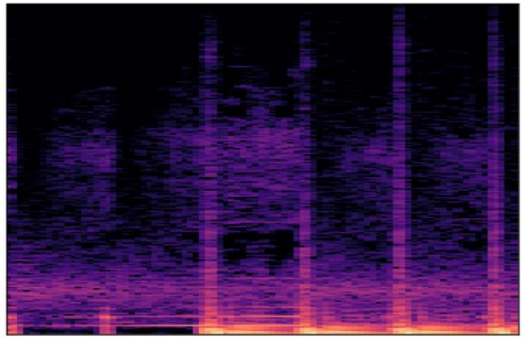
BNMusic: Blending Environmental Noises into Personalized Music, NeurIPS'25

Novel Application in Music Generation

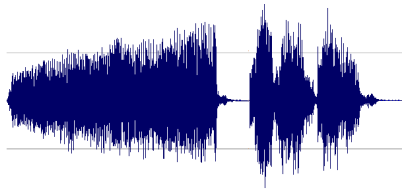
Novel task and objective:
blend the environmental noises into a music piece specified by user prompt.

"Energetic Music."

Tuned Diffusion Models on Mel-spectrogram Plots



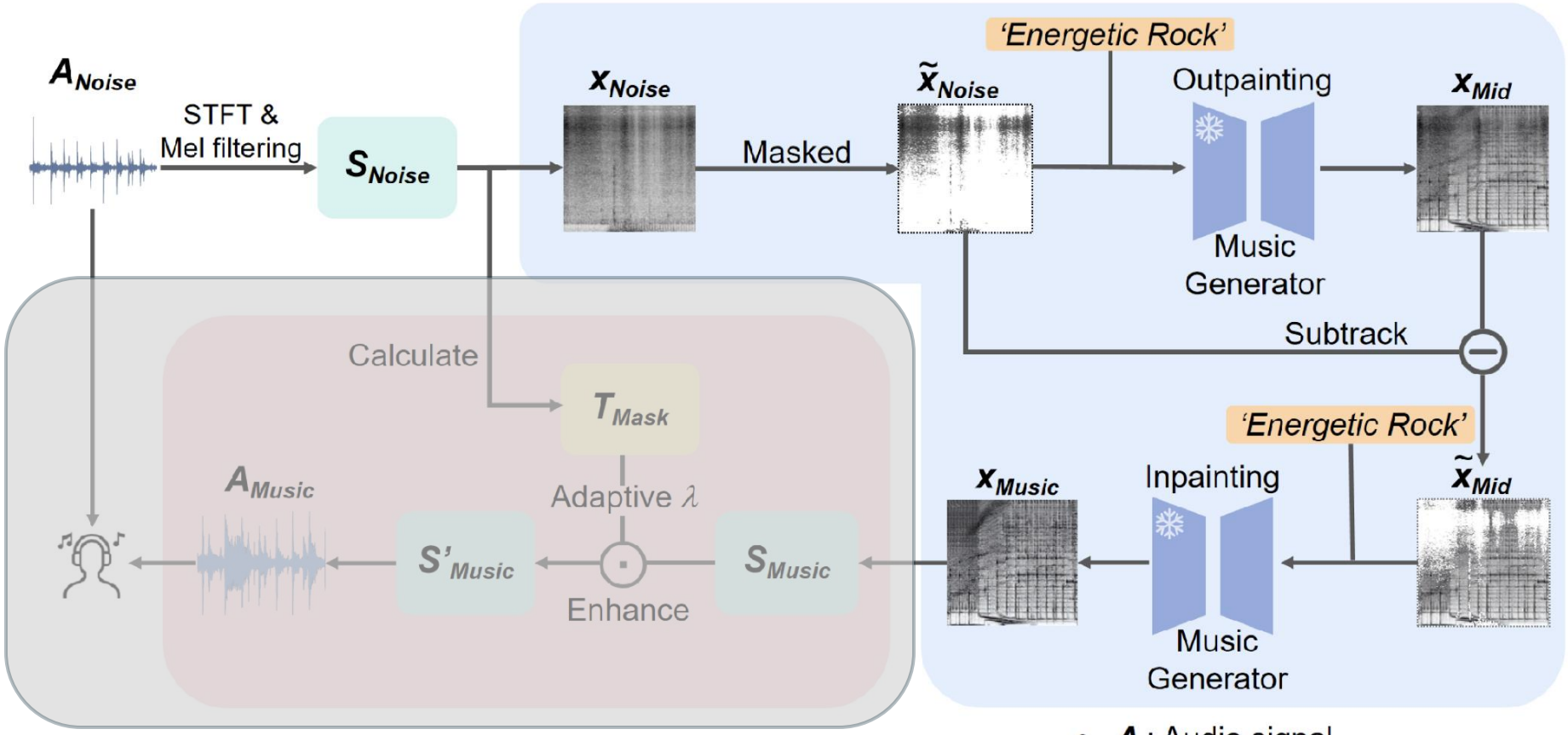
Tuned Decoder



Key Technical Challenges:
- Cover the main frequency component in noises;
- Control the volume.

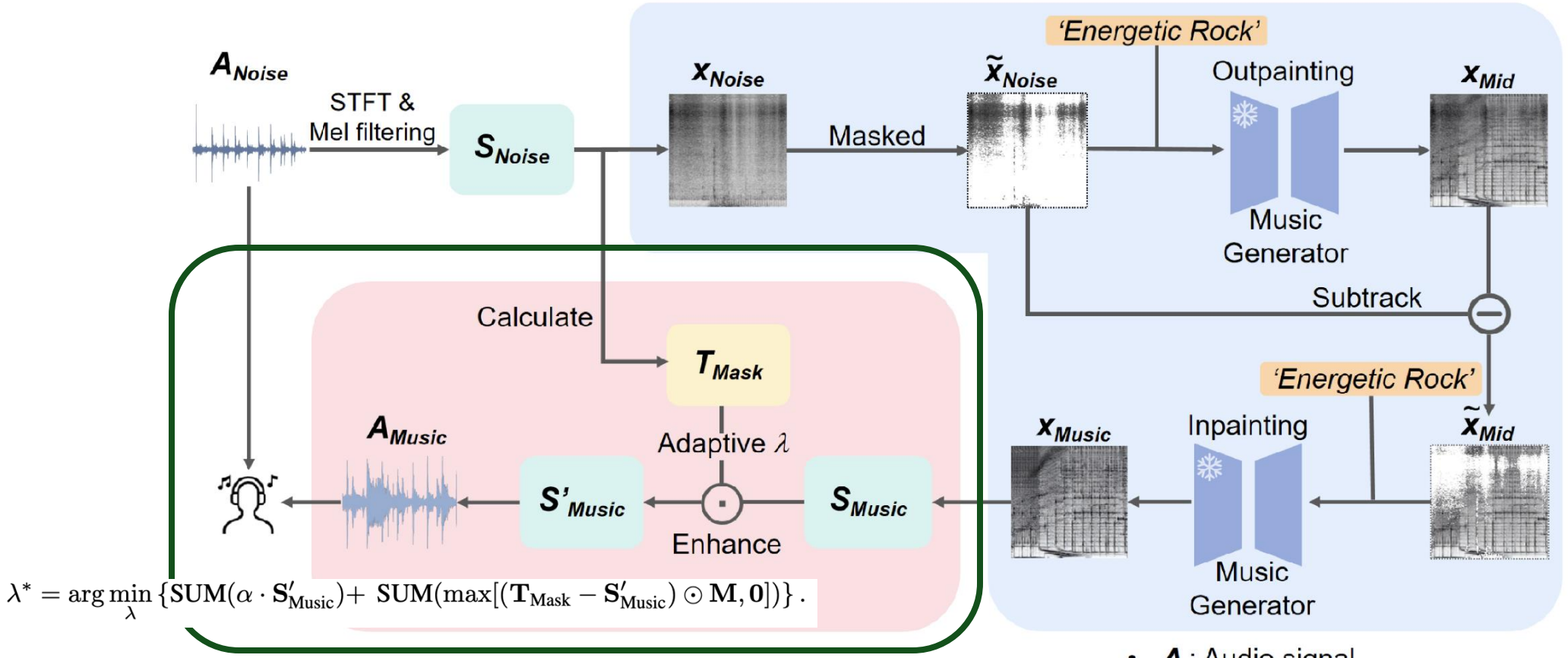
BNMusic: Blending Environmental Noises into Personalized Music, **NeurIPS**'25

Stage 1: Noise Aligned Music Synthesis



BNMusic: Blending Environmental Noises into Personalized Music, **NeurIPS**'25

Stage 2: Blending Enhancement by Adaptive Amplification

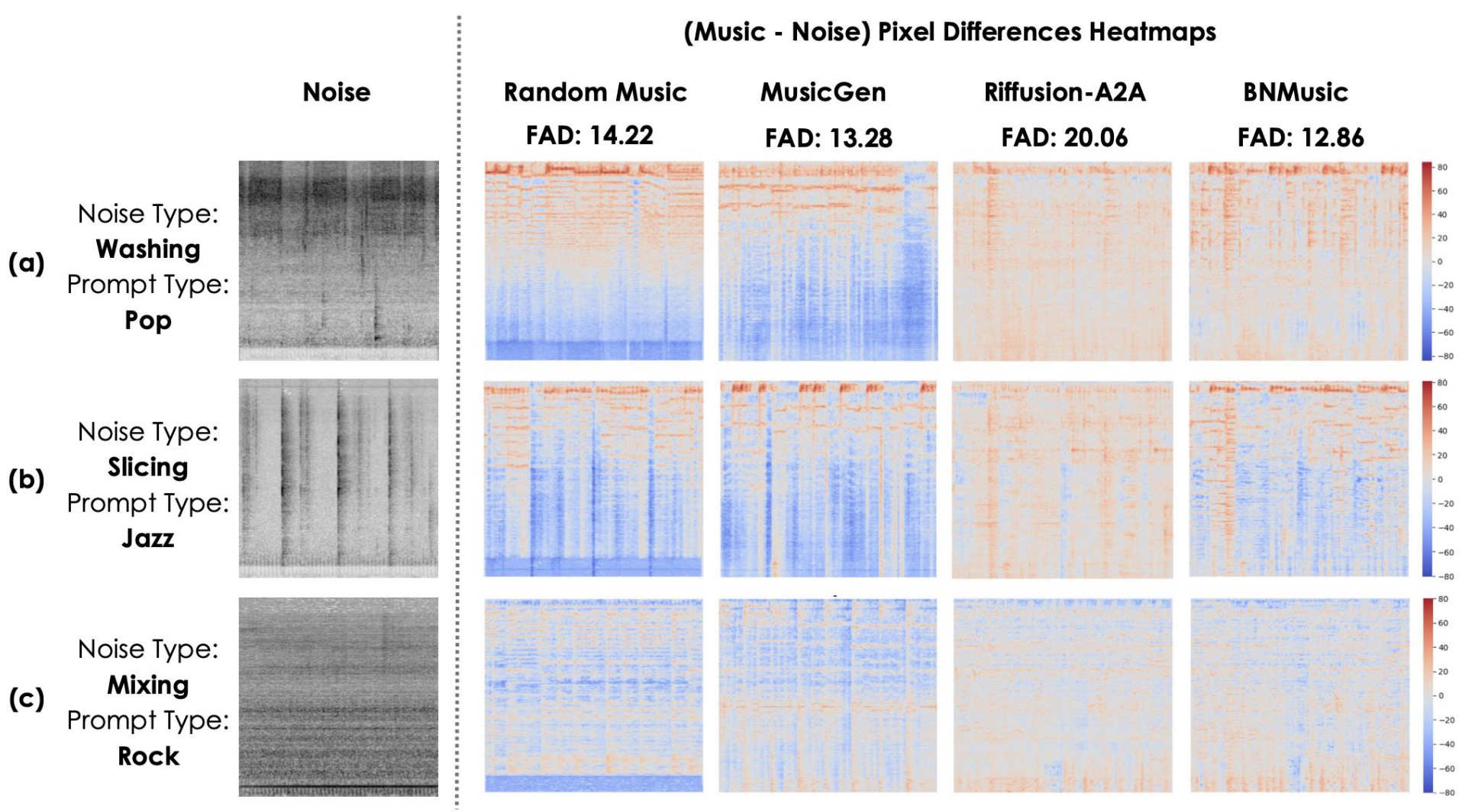


BNMusic: Blending Environmental Noises into Personalized Music, **NeurIPS**'25

Objective and Subjective Evaluation

Table 1: **Objective evaluati**
Kullback-Leibler (KL) Dive
and Direct Outputs (right)—
our BNMusic method consis
generalizability of our approa

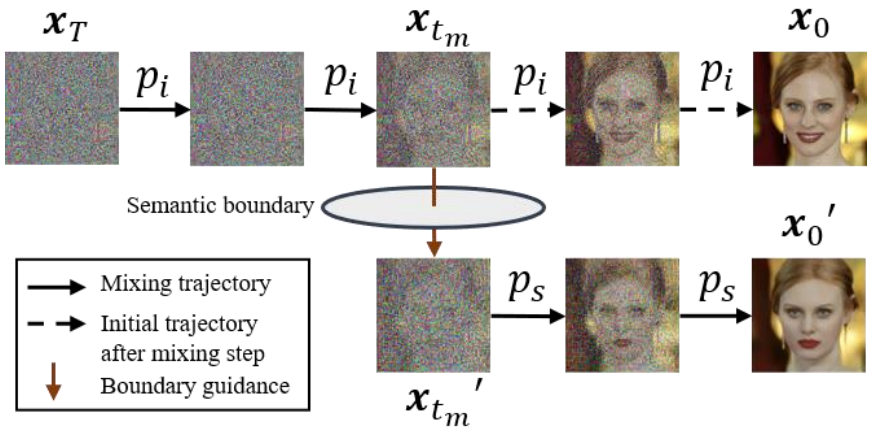
Methods	EPIC-S
	FAD↓
Noise Only	34.17
Random Music	14.22
MusicGen [3]	13.28
Riff A2A [6]	20.06
BNMusic (Ours)	**12.86**



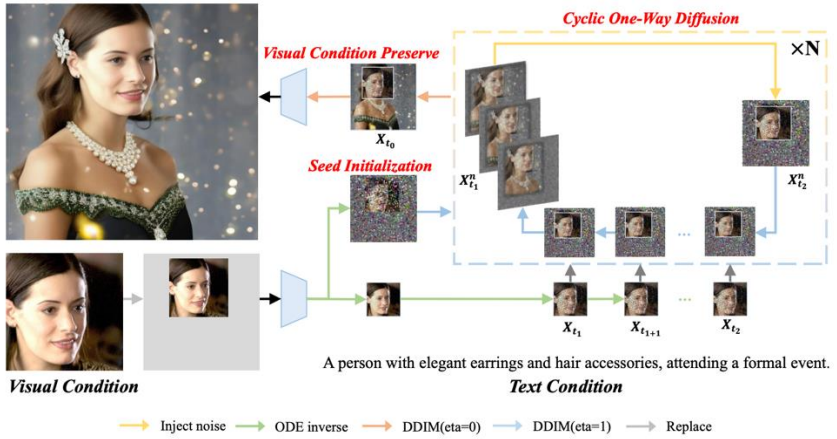
https://d-fas.github.io/BNMusic_page/

BNMusic: Blending Environmental Noises into Personalized Music, **NeurIPS**'25

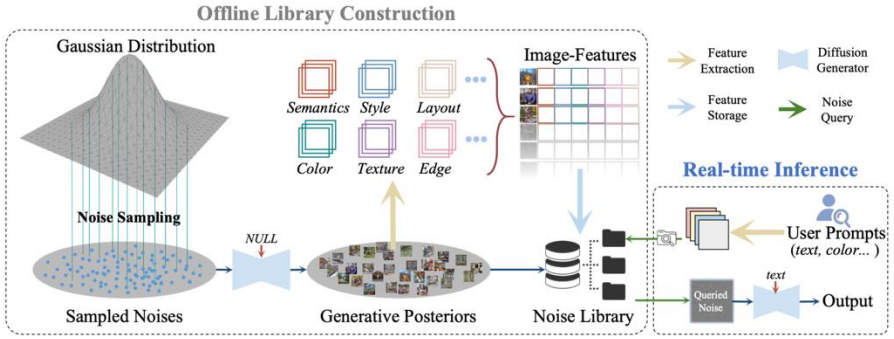
Dynamic Sampling for Interpretable Control in Multimodal Generation



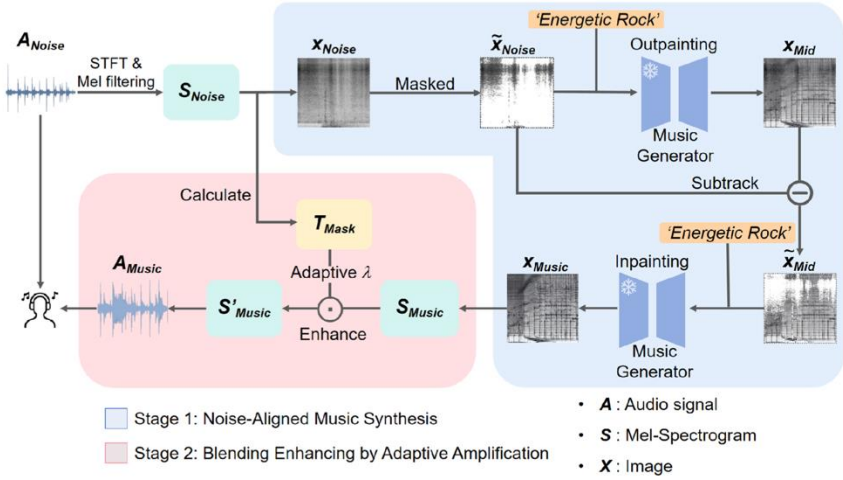
[ZWDRY **NeurIPS**'23]



[W*Y*QZ⁺W⁺ **ICLR**'24]



[WHZRW **ICCV'25 Highlight**]



[ZMMWZ **NeurIPS'25**]

Dynamic Modeling for Dynamic Physical Systems in Astronomy

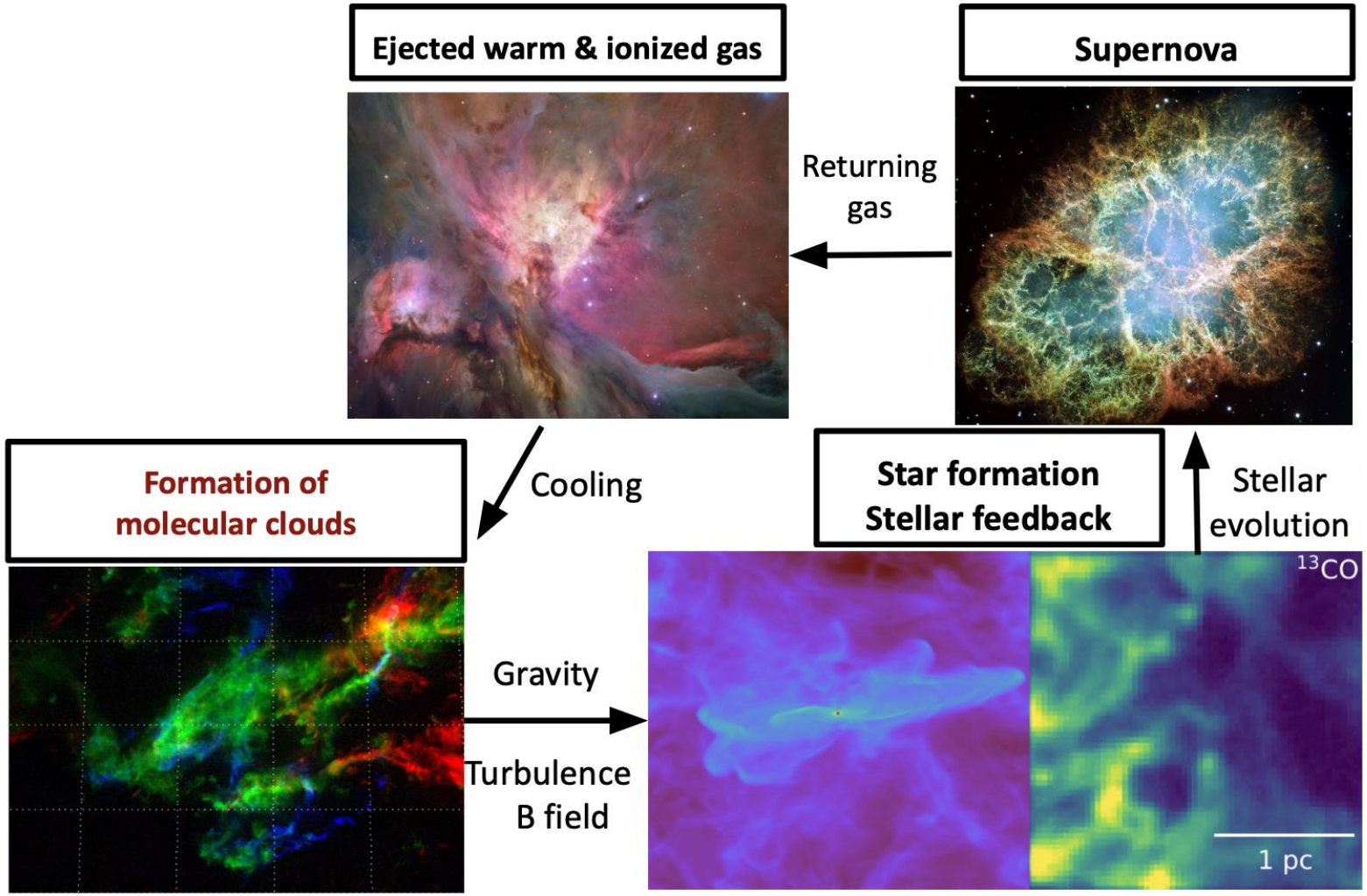
My other line of research works:

[ZXDTR *NeurIPS*'25] Dynamic Diffusion Schrödinger Bridge in Astrophysical Observational Inversions

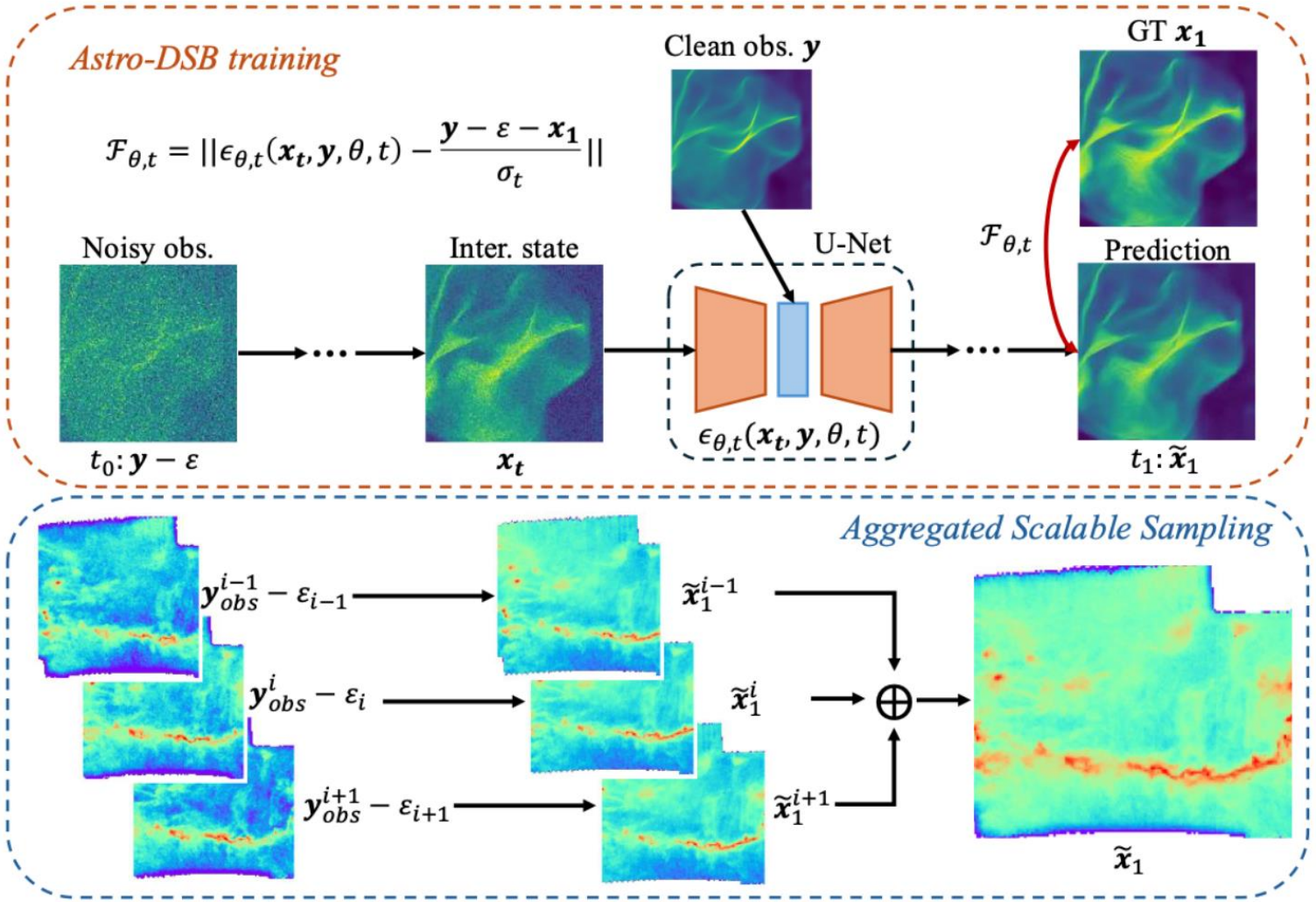
[XKLZHT *ApJ'25*] Exploring Magnetic Fields in Molecular Clouds through Denoising Diffusion Probabilistic Models

[XZ *Astronomy and Computing'24*] Surveying Image Segmentation Approaches in Astronomy

[XTHZ *ApJ'23*] Denoising Diffusion Probabilistic Models to Predict the Density of Molecular Clouds



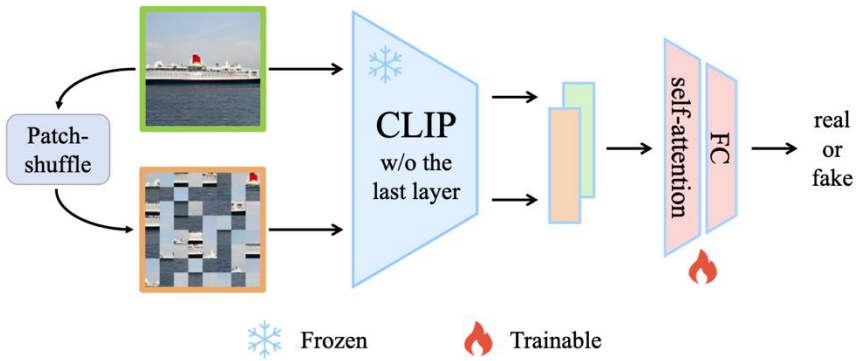
Astro-DSB for Astrophysical Inversion Predictions



A Diffusion Schrodinger Bridge (DSB) based modeling framework for predicting physical states based on partial observations.

My Other Recent ML/CV Works, Outreach and Activities

Generalized deepfake detection by learning from discrepancy at scale, better OOD detetction

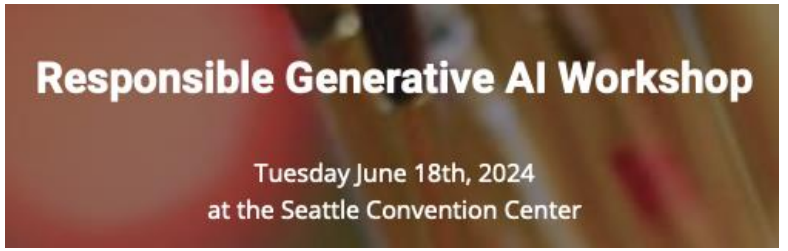


[Y*Q*ZRW CVPR'25] D^3: Scaling Up Deepfake Detection by Learning from Discrepancy

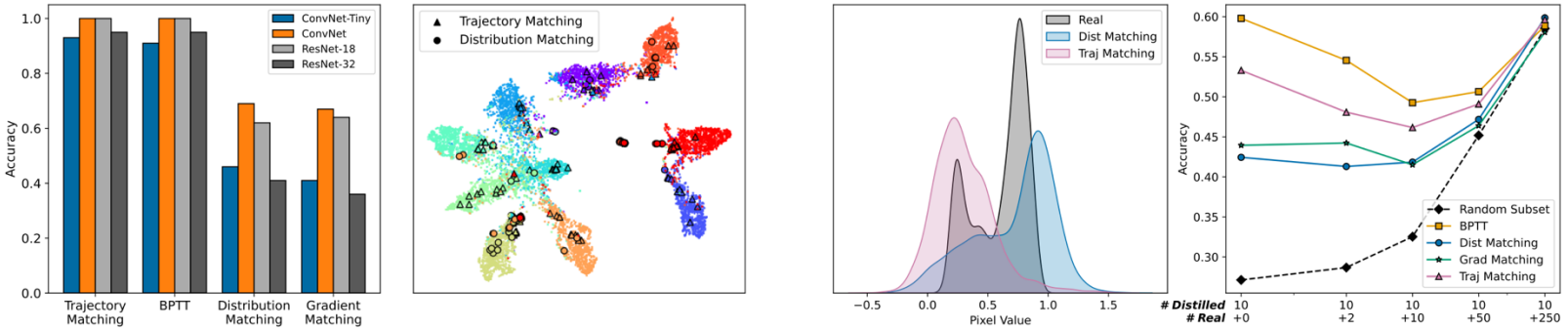
CV4Science workshop @ CVPR'25 '26

Computer Vision for Science

Responsible GenAI workshop @ CVPR'24 '25



Analytic work on dataset distillation algorithms for better interpretability.



[YZDR ICML'24] What is Dataset Distillation Learning ?

MIT EECS Rising Stars 2024

My Research Focus:

Bridge the First-Principal Driven and Data-Driven Modeling

Integrating designed dynamics in generative modeling design

Leveraging the dynamics in practical applications (multimodal and physics)

Understanding the dynamics of generative models and physical systems