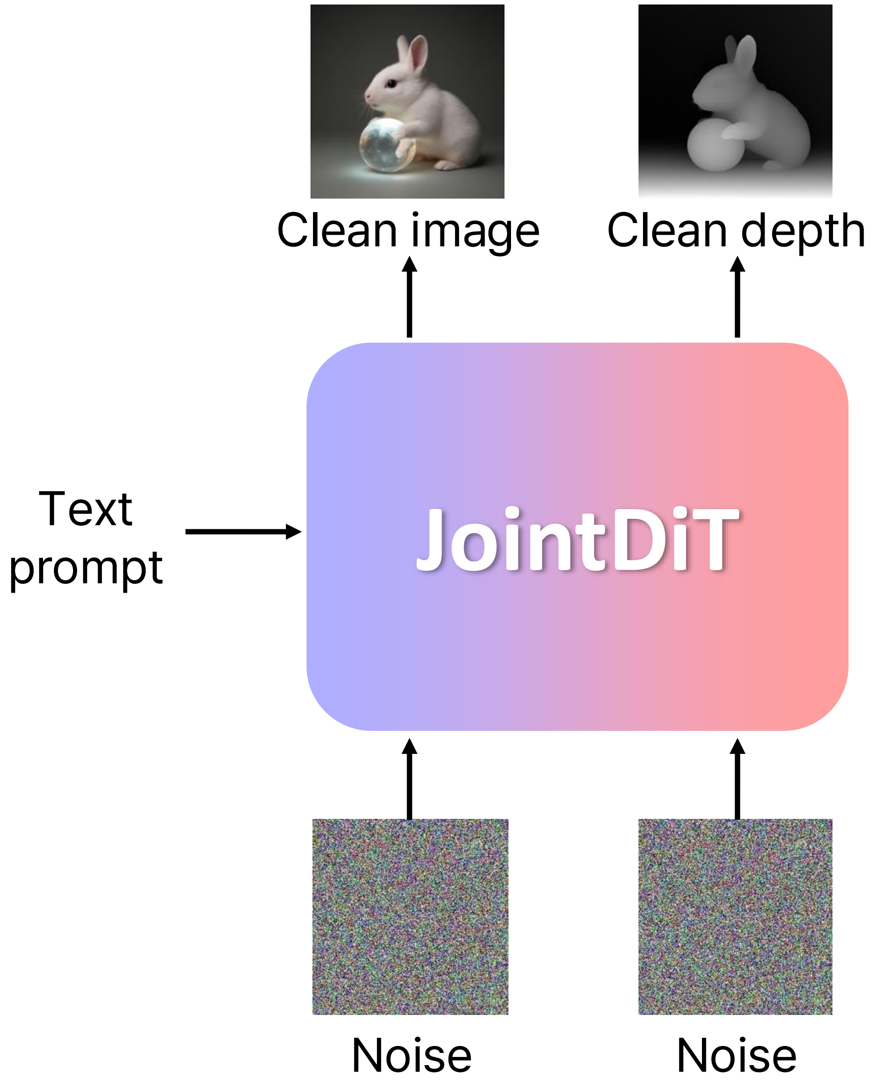
JointDiT: Enhancing RGB-Depth Joint Modeling with Diffusion Transformers

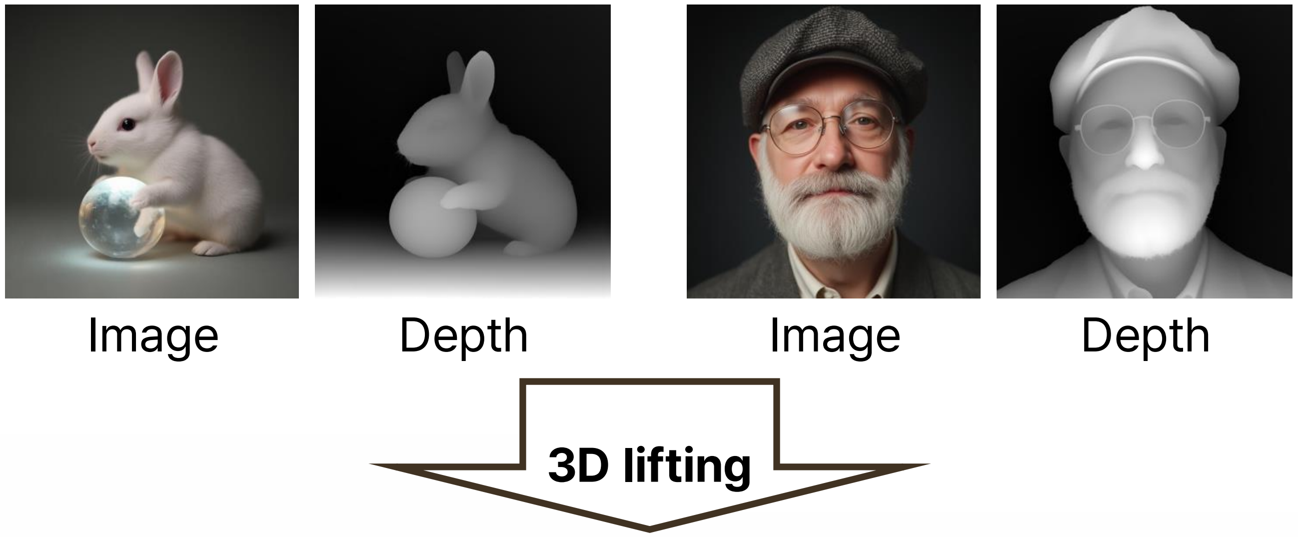
Kwon Byung-Ki[1,2] Qi Dai[2] Lee Hyoseok[1] Chong Luo[2] Tae-Hyun Oh[3]

[1]POSTECH [2]Microsoft Research Asia [3]KAIST

A Unified Model for Image and Depth Vision Tasks

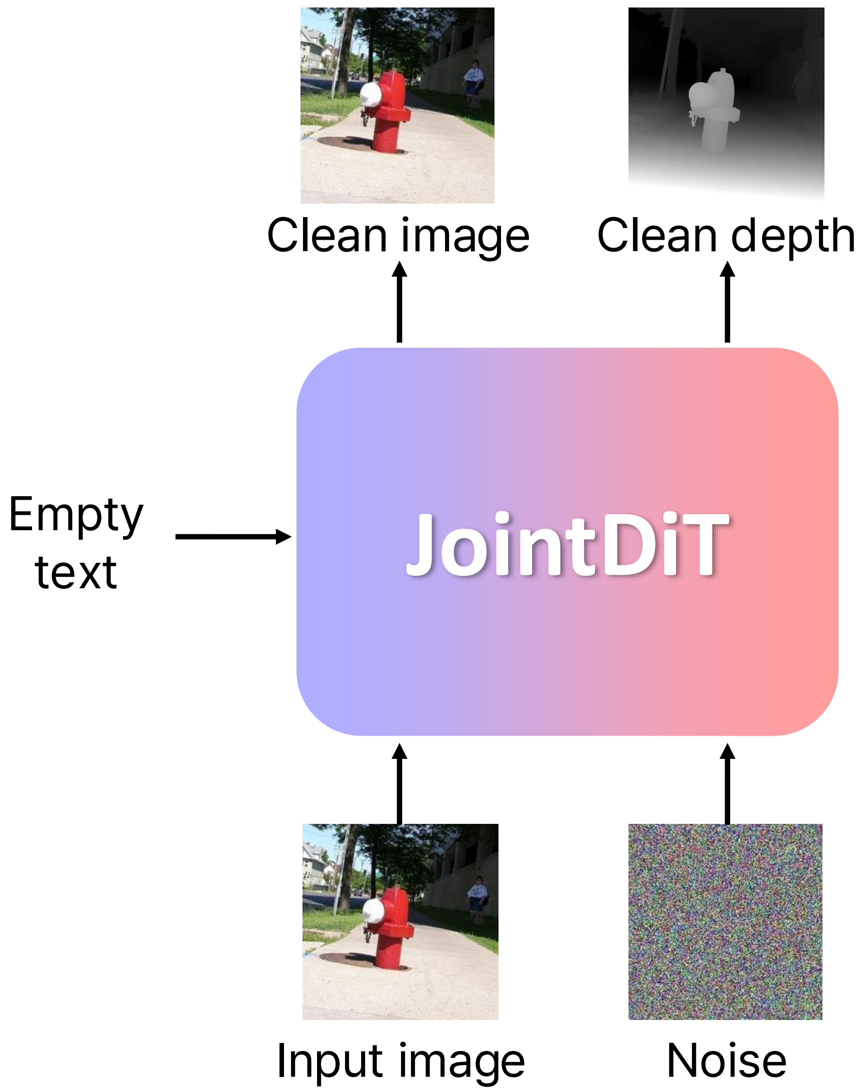
1. Joint generation





A Unified Model for Image and Depth Vision Tasks

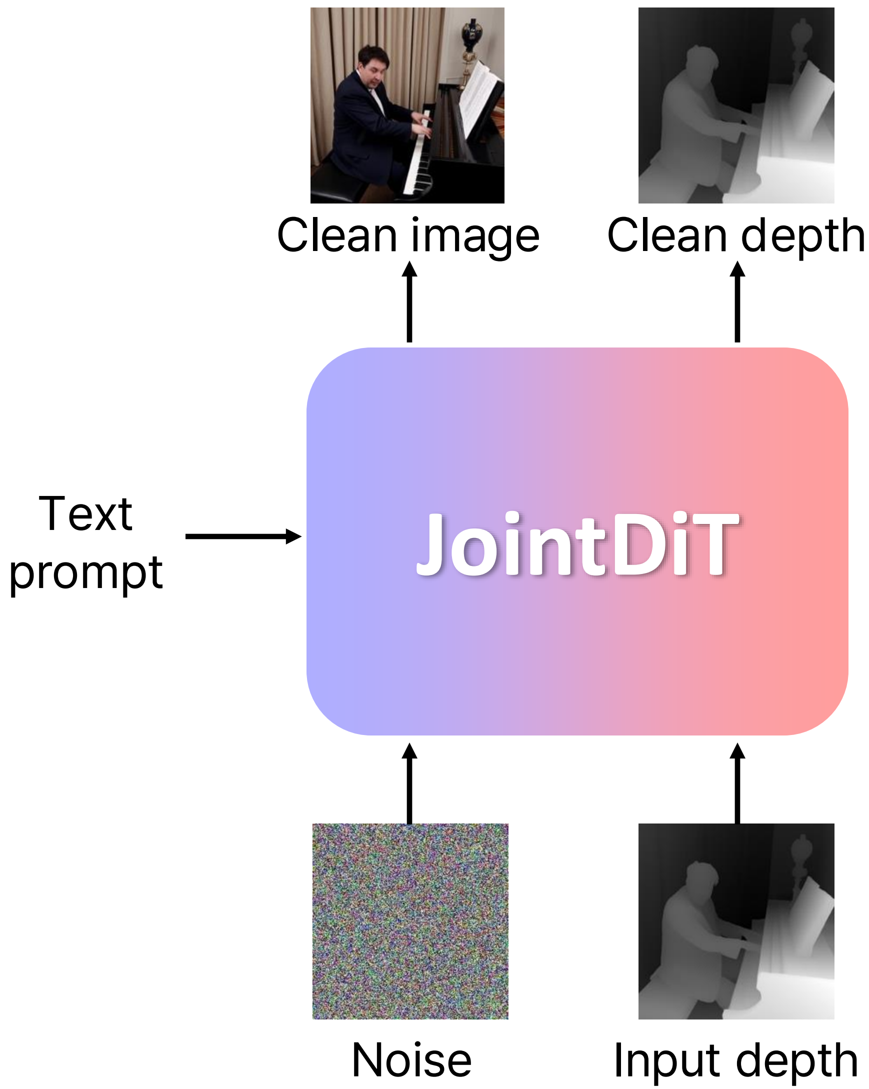
2. Depth estimation



Input image　　Depth estimation

A Unified Model for Image and Depth Vision Tasks

3. Depth-conditioned image generation



"A man in a suit is playing the piano"

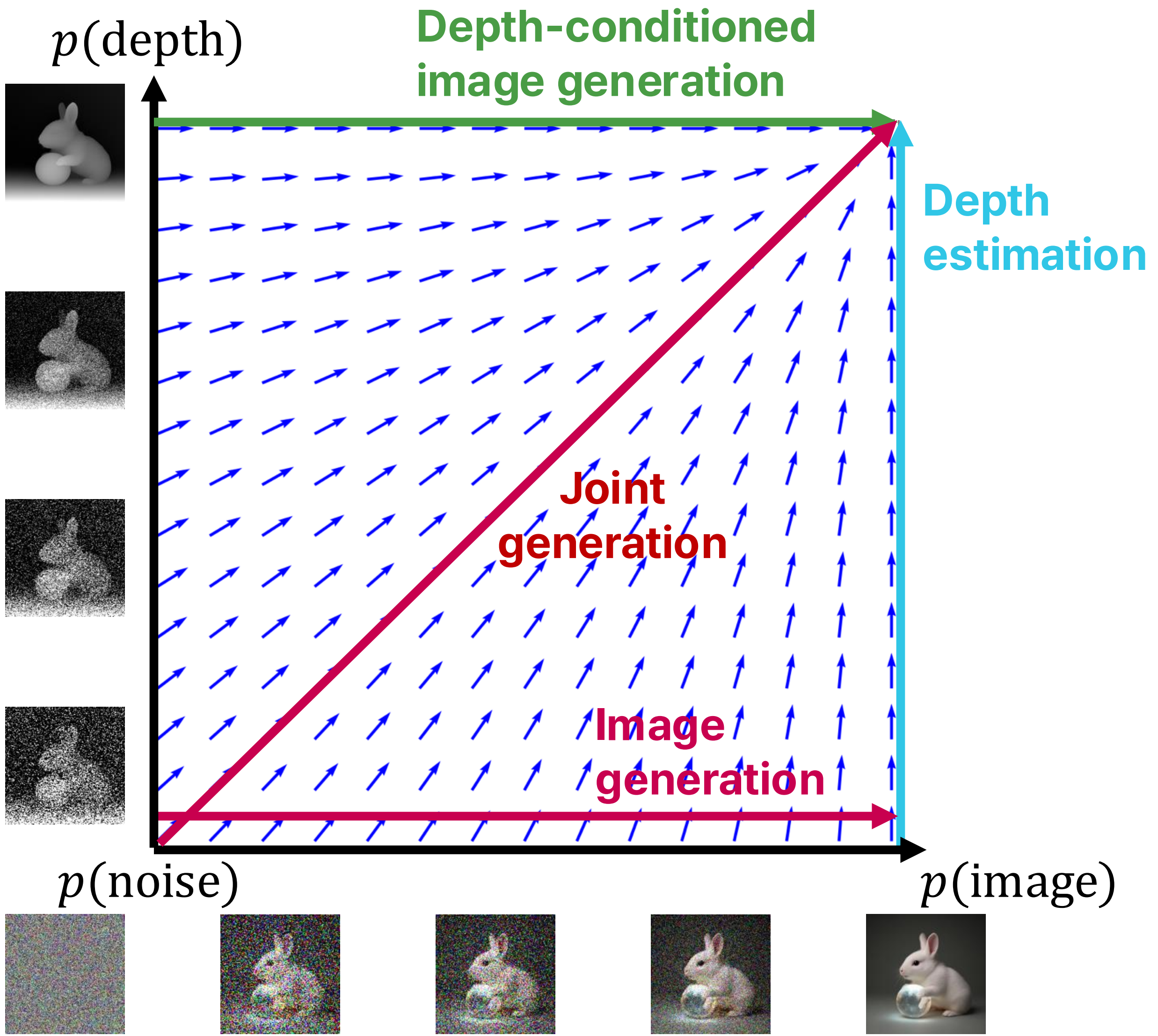
"A gray fox sitting on the ground near a road"

Input depth

Generated image

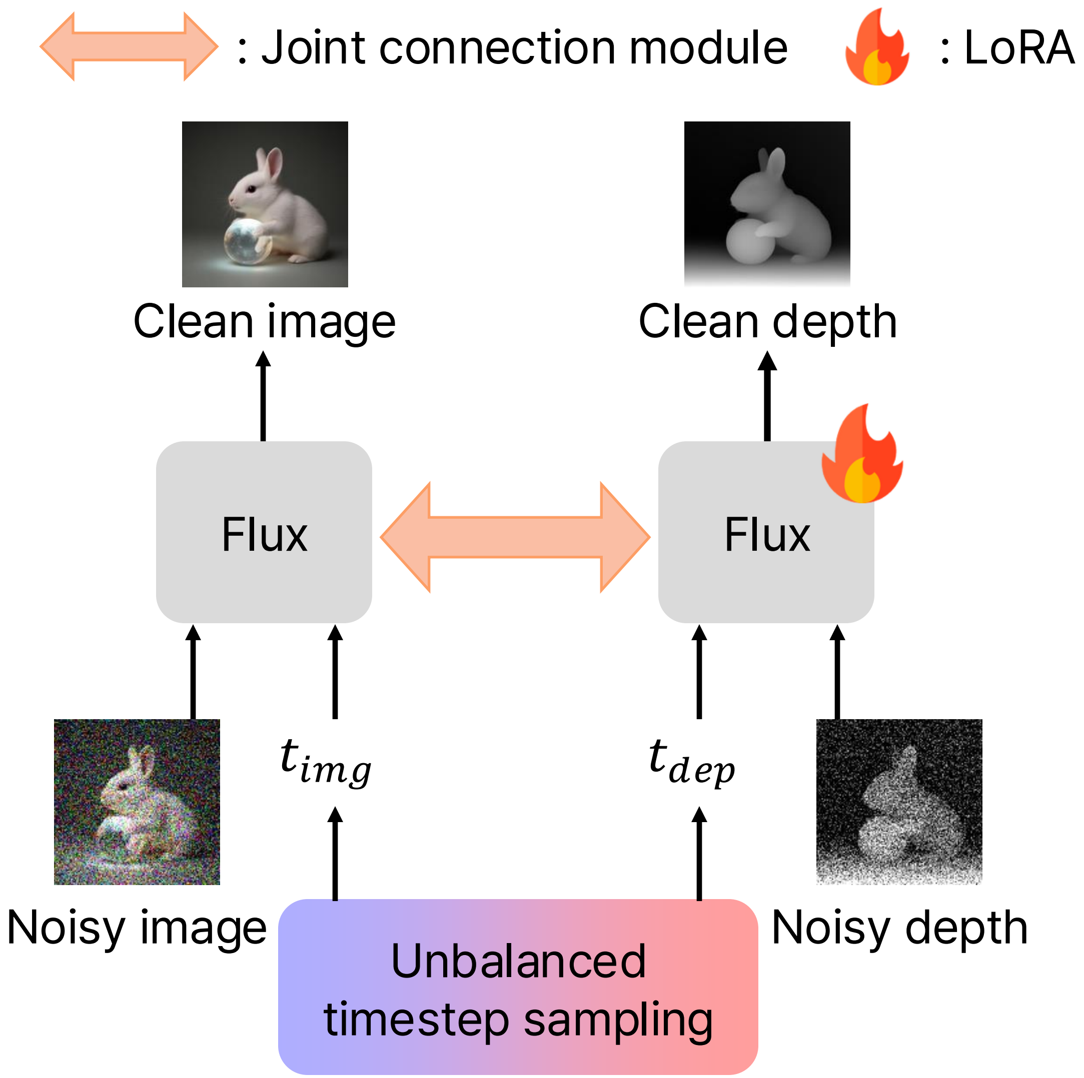
Joint Distribution Modeling for Generative Tasks

- Joint distribution modeling can cover various generative tasks



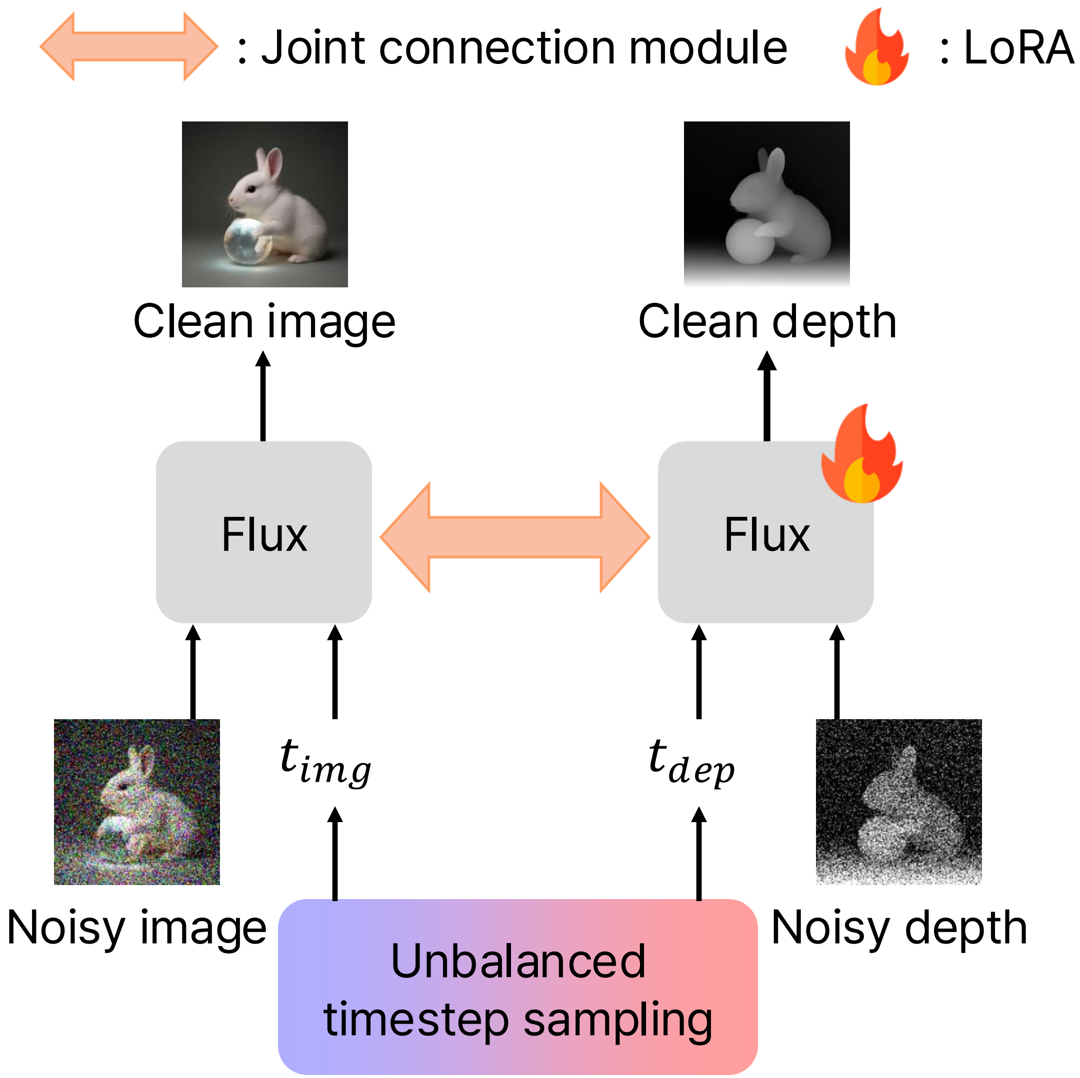
Dedicated Pipeline for Separate Noise Level Training

- Building on Flux, we introduce depth branch and joint connection module

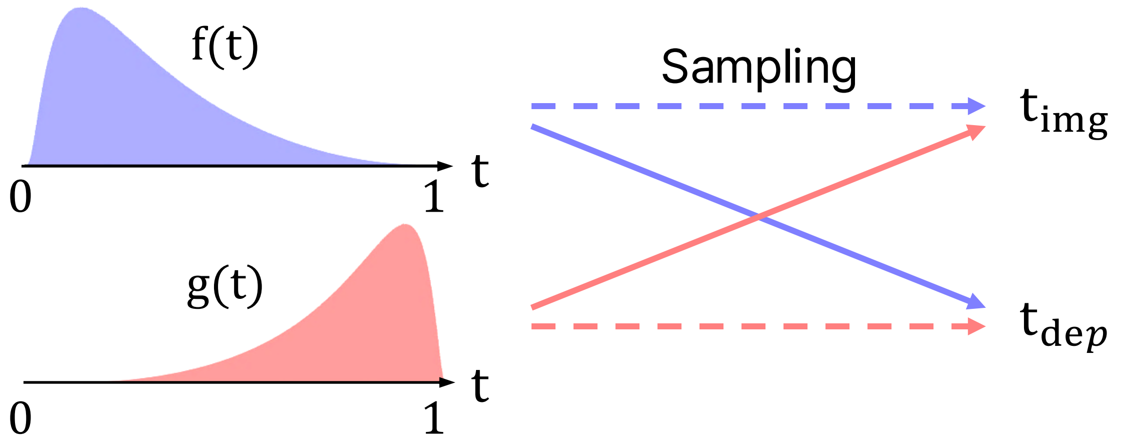


Dedicated Pipeline for Separate Noise Level Training

- Unbalanced timestep sampling → Balancing the combination of noise levels

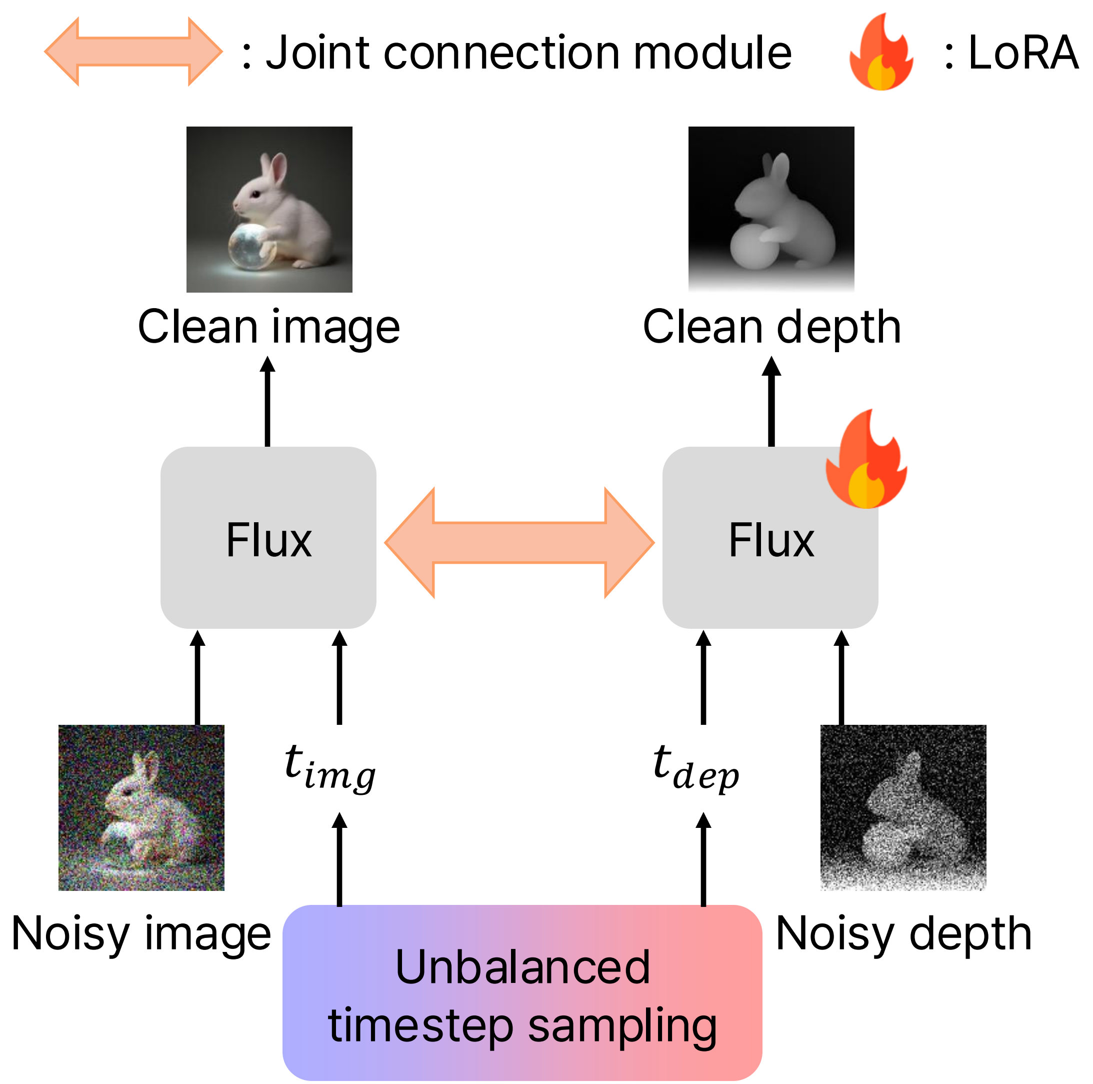


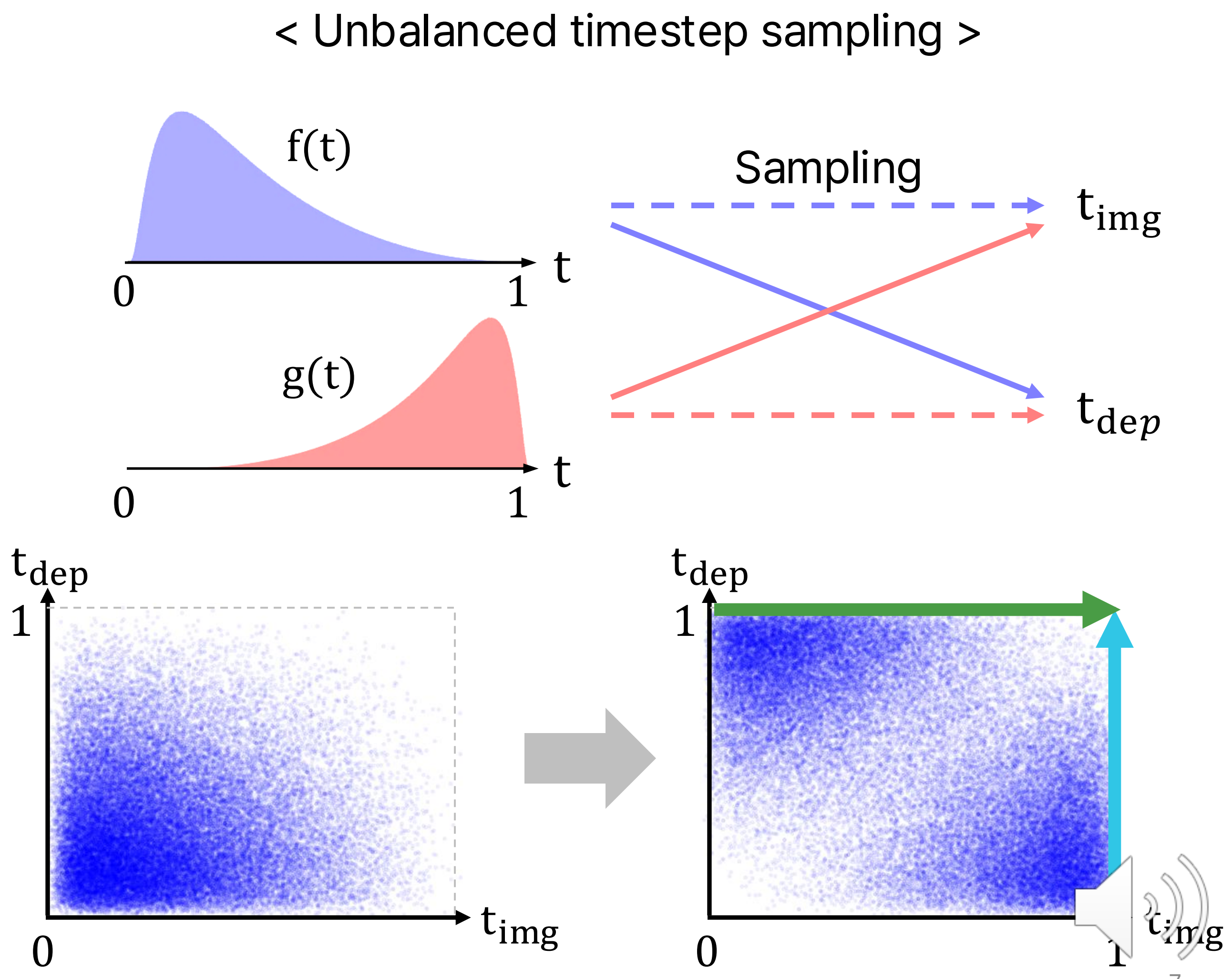
< Unbalanced timestep sampling >



Dedicated Pipeline for Separate Noise Level Training

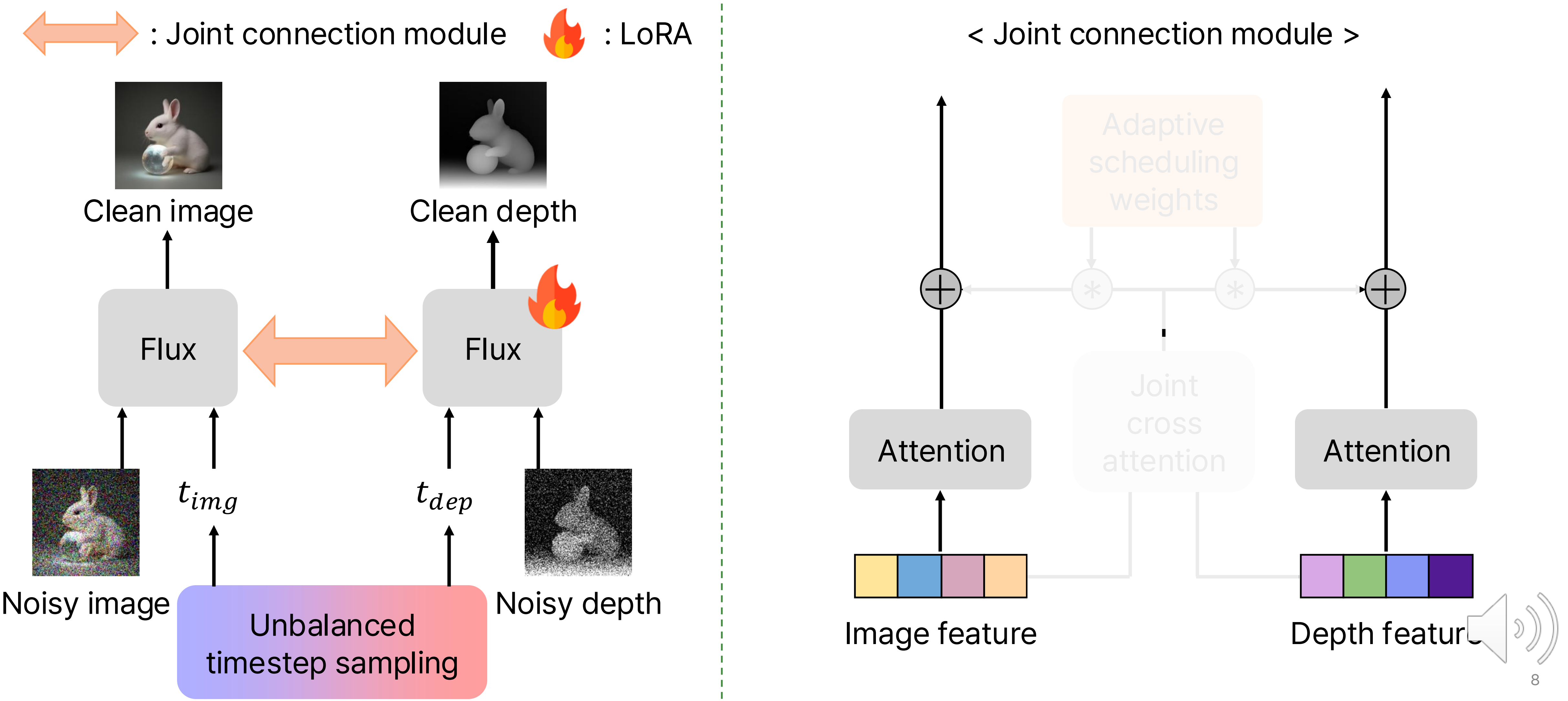
- Unbalanced timestep sampling → Balancing the combination of noise levels





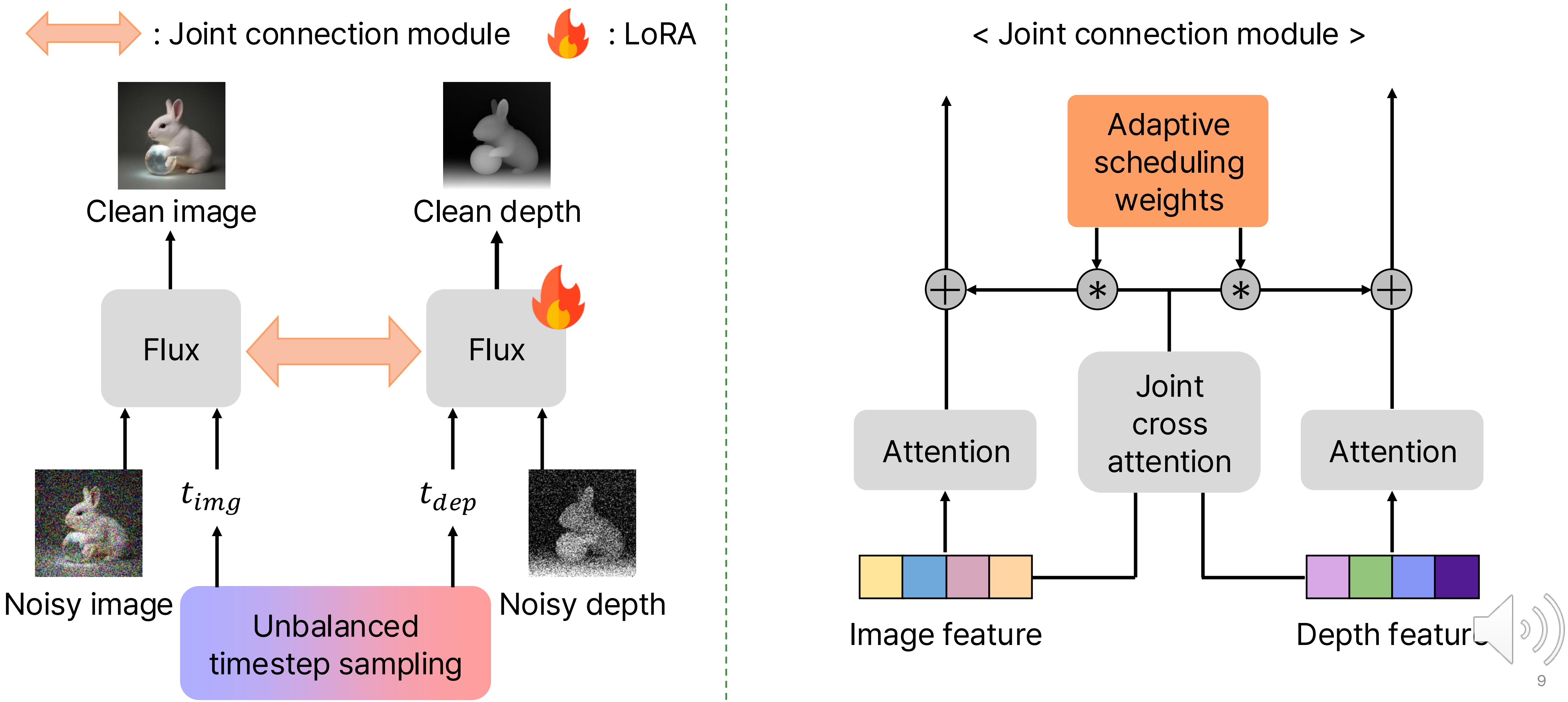
Dedicated Pipeline for Separate Noise Level Training

- Adaptive scheduling weights → Guiding noisier modality with cleaner modality



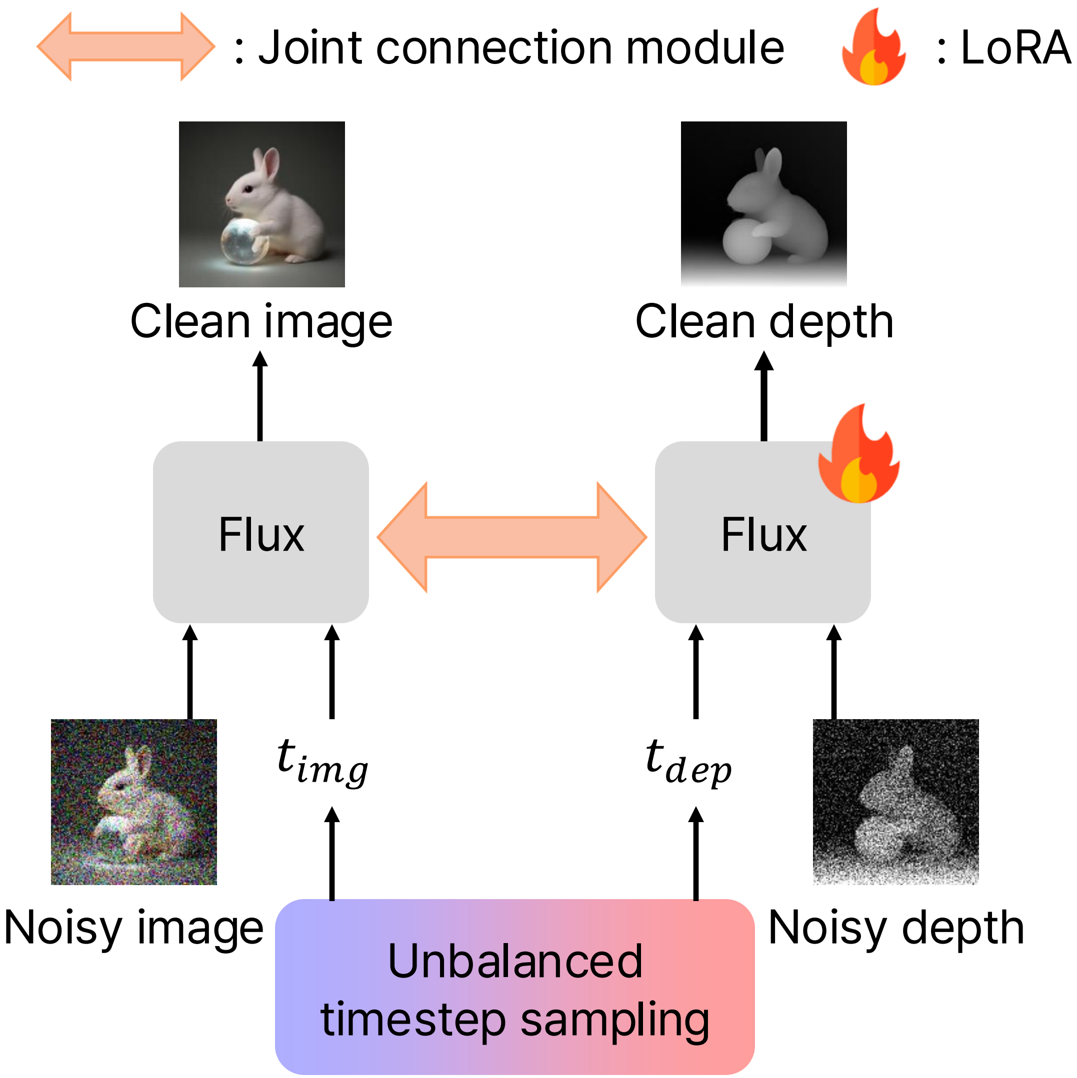
Dedicated Pipeline for Separate Noise Level Training

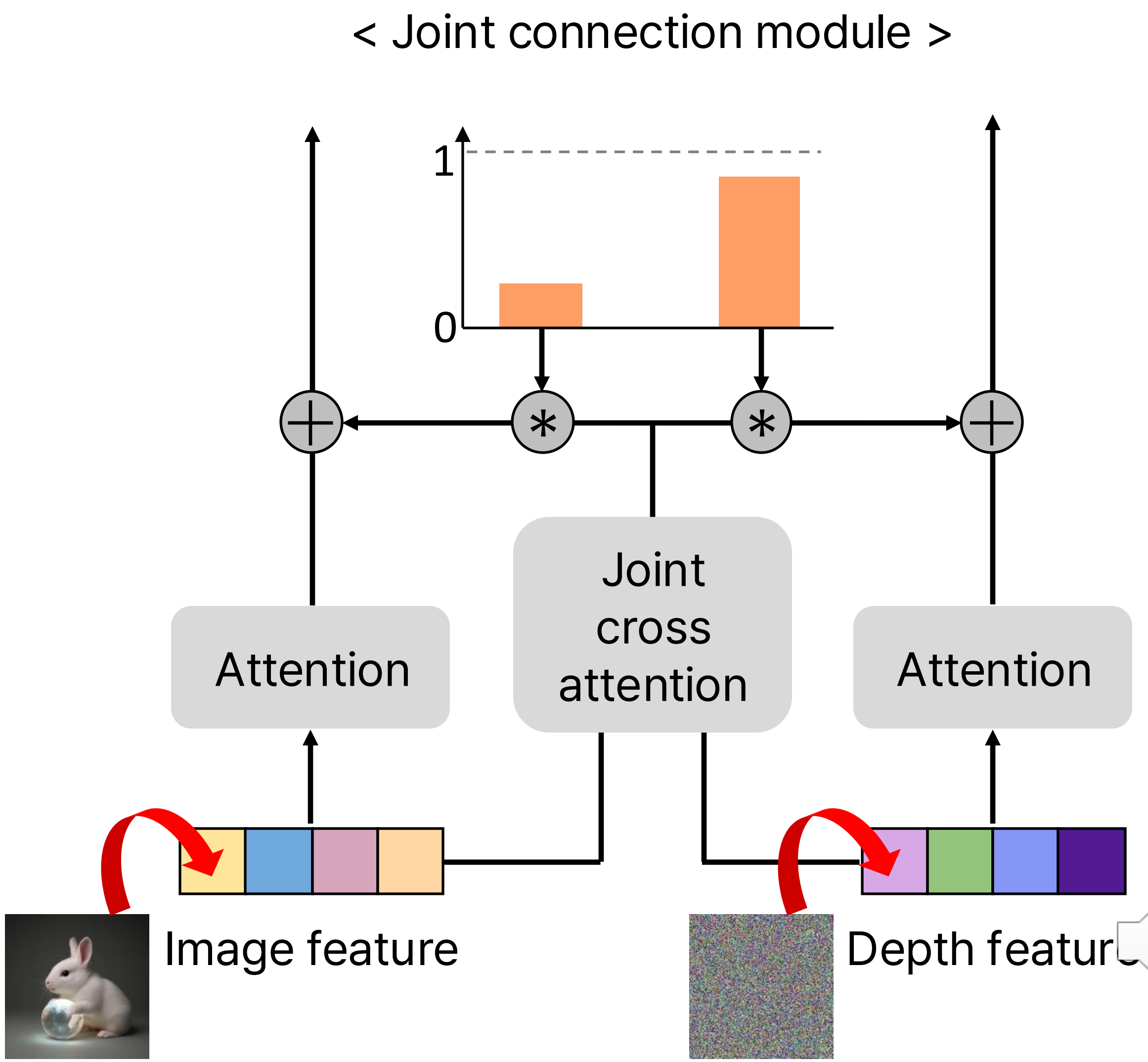
- Adaptive scheduling weights → Guiding noisier modality with cleaner modality



Dedicated Pipeline for Separate Noise Level Training

- Adaptive scheduling weights → Guiding noisier modality with cleaner modality

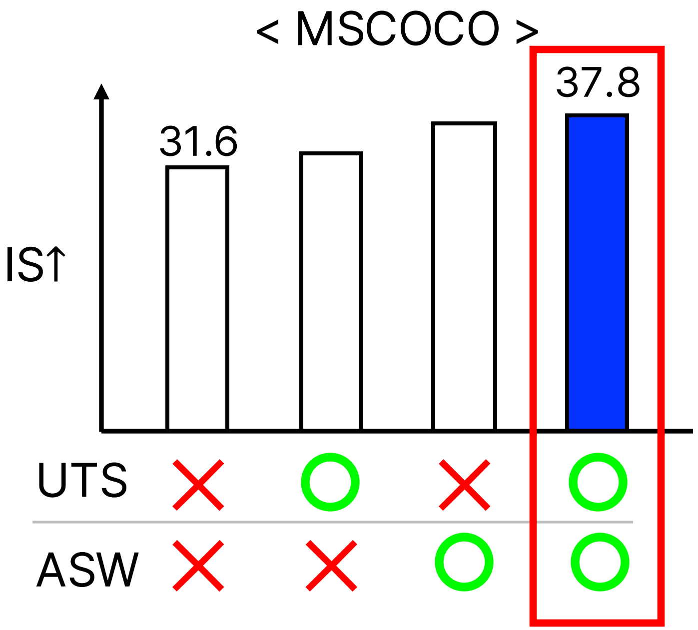




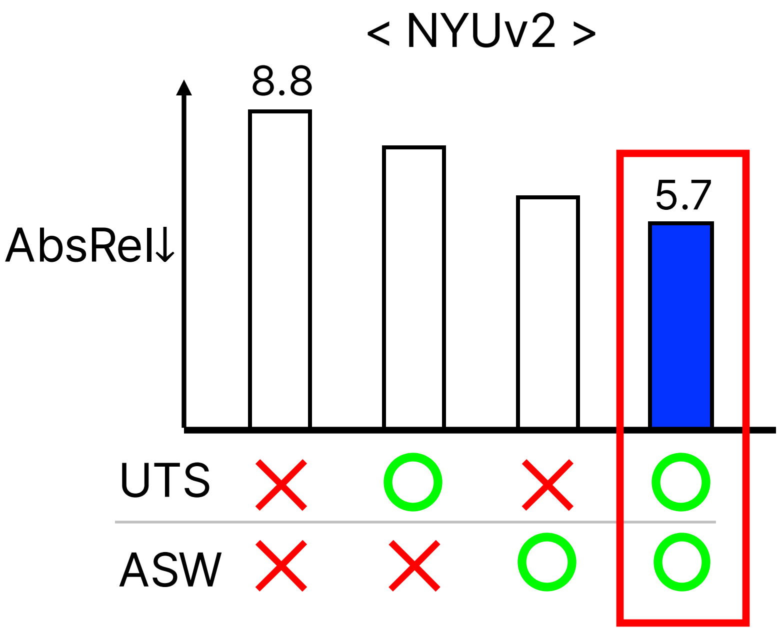
Effects of Proposed Methods on Generative Tasks

- Unbalanced Timestep Sampling (UTS) and Adaptive Scheduling Weights (ASW)
  → Significant improvement in image and depth joint generative tasks

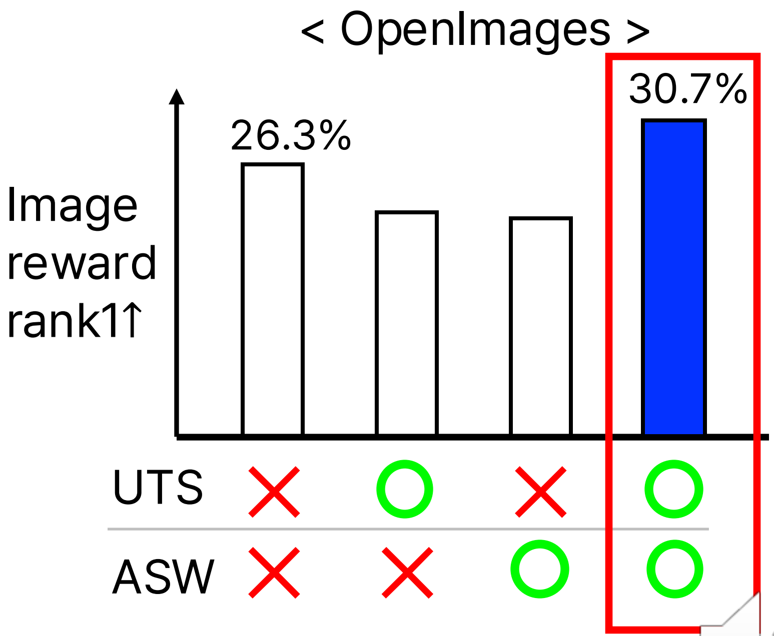
Joint generation



Depth estimation



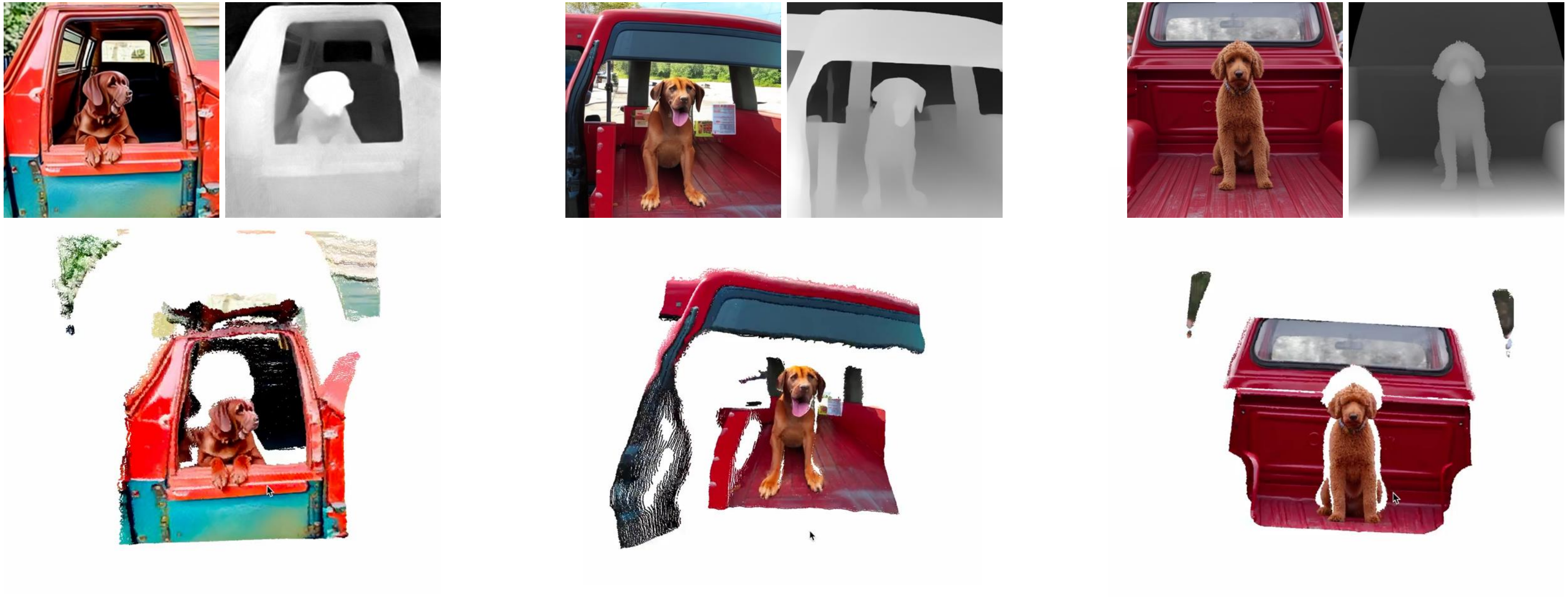
Depth-conditioned image generation



Experiment - Joint Generation

- Outstanding joint generation capability compared to LDM3D [1] and JointNet [2]

“A big brown dog sitting in the back of a red truck”

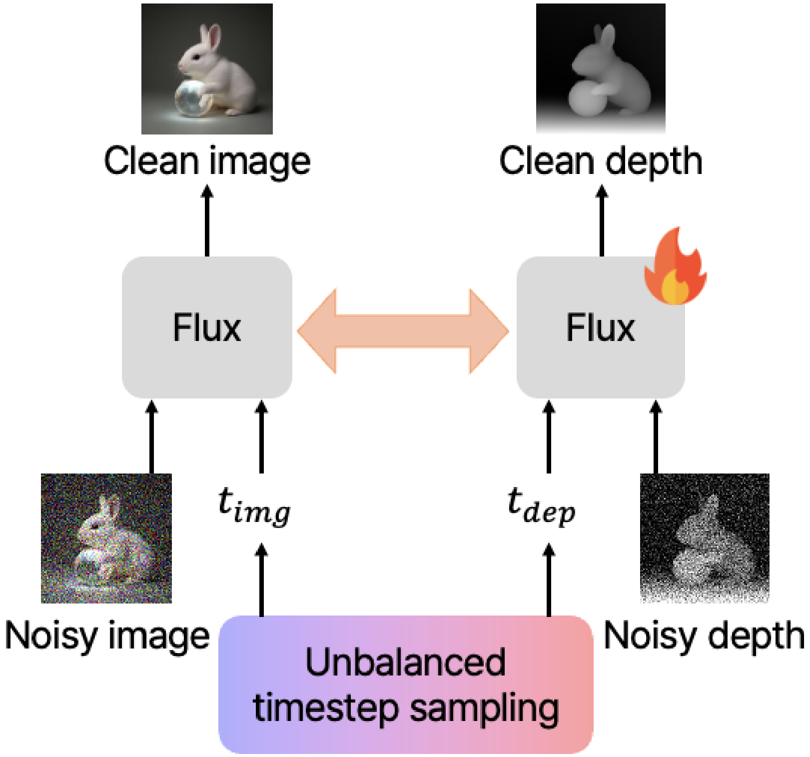


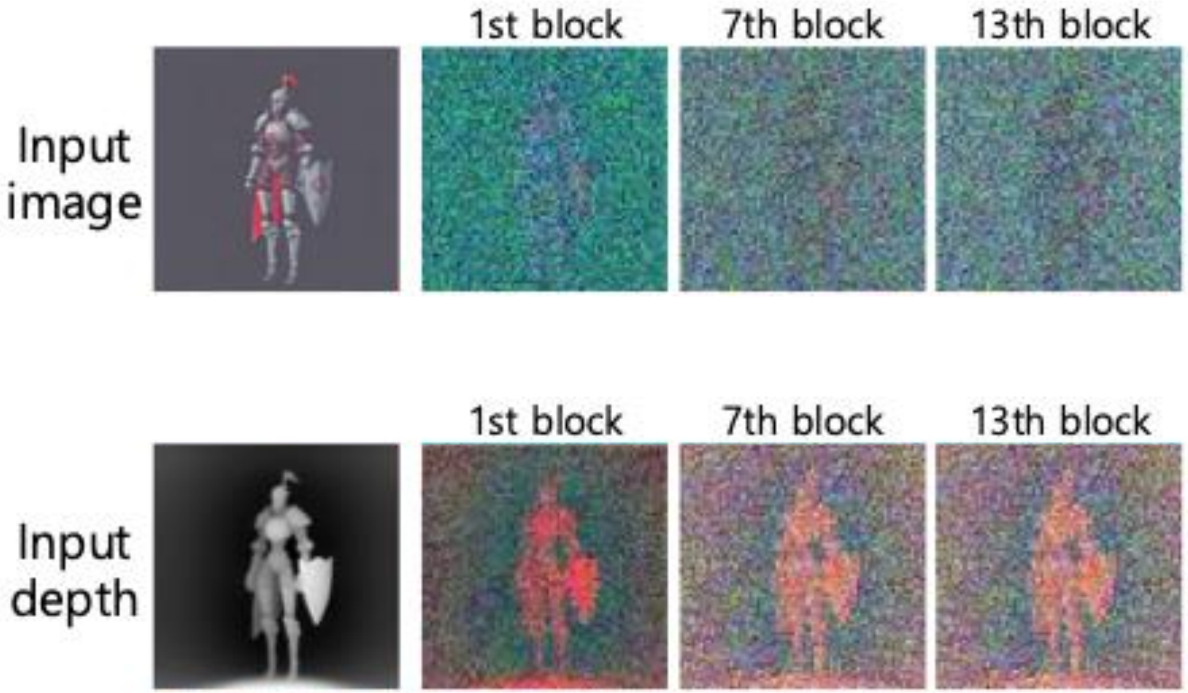
LDM3D [1] JointNet [2] Ours

[1] Stan, Gabriela Ben Melech, et al. "Ldm3d: Latent diffusion model for 3d," arXiv 2023.
[2] Zhang, Jingyang, et al. "Jointnet: Extending text-to-image diffusion for dense distribution modeling," ICLR 2024.

Experiment – Feature analysis

- Feature visualization of RGB and Depth branches following Tumanyan *et al.* [3]





- RGB branch focuses on high-frequency semantic patterns
- Depth branch focuses on coarse geometric structures of the scene

[3] Tumanyan et al. "Plug-and-play diffusion features for text-driven image-to-image translation," CVPR 2023.

Experiment - Depth Estimation

- Superior performance compared to generative joint generation methods
- Comparable accuracy to generative depth estimation methods

The evaluation metric is Absolute Mean Relative Error (AbsRel),. Lower is better.

Types	Methods	NYUv2	ScanNet	DIODE
Generative depth estimation	Marigold [4]	5.5	6.4	30.8
	Geowizard [5]	5.2	6.1	29.7
Generative joint generation	JointNet [2]	13.7	14.7	35.0
	UniCon [6]	7.9	9.2	–
	Ours	5.7	6.6	27.3
	Ours + finetune	**5.0**	**5.6**	**26.6**

[2] Zhang, Jingyang, et al. "Jointnet: Extending text-to-image diffusion for dense distribution modeling," ICLR 2024.
[4] Ke, Bingxin, et al. "Repurposing diffusion-based image generators for monocular depth estimation," CVPR 2024.
[5] Fu, Xiao, et al. "Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image," ECCV 2024.
[6] Li, Xirui, et al. "A simple approach to unifying diffusion-based conditional generation," ICLR 2025.

Experiment - Depth-Conditioned Image Generation

- More text-aligned and realistic results compared to JointNet [2] and UniCon [6]

Three dogs playing with a green frisbee

A gray fox sitting on the ground near a road

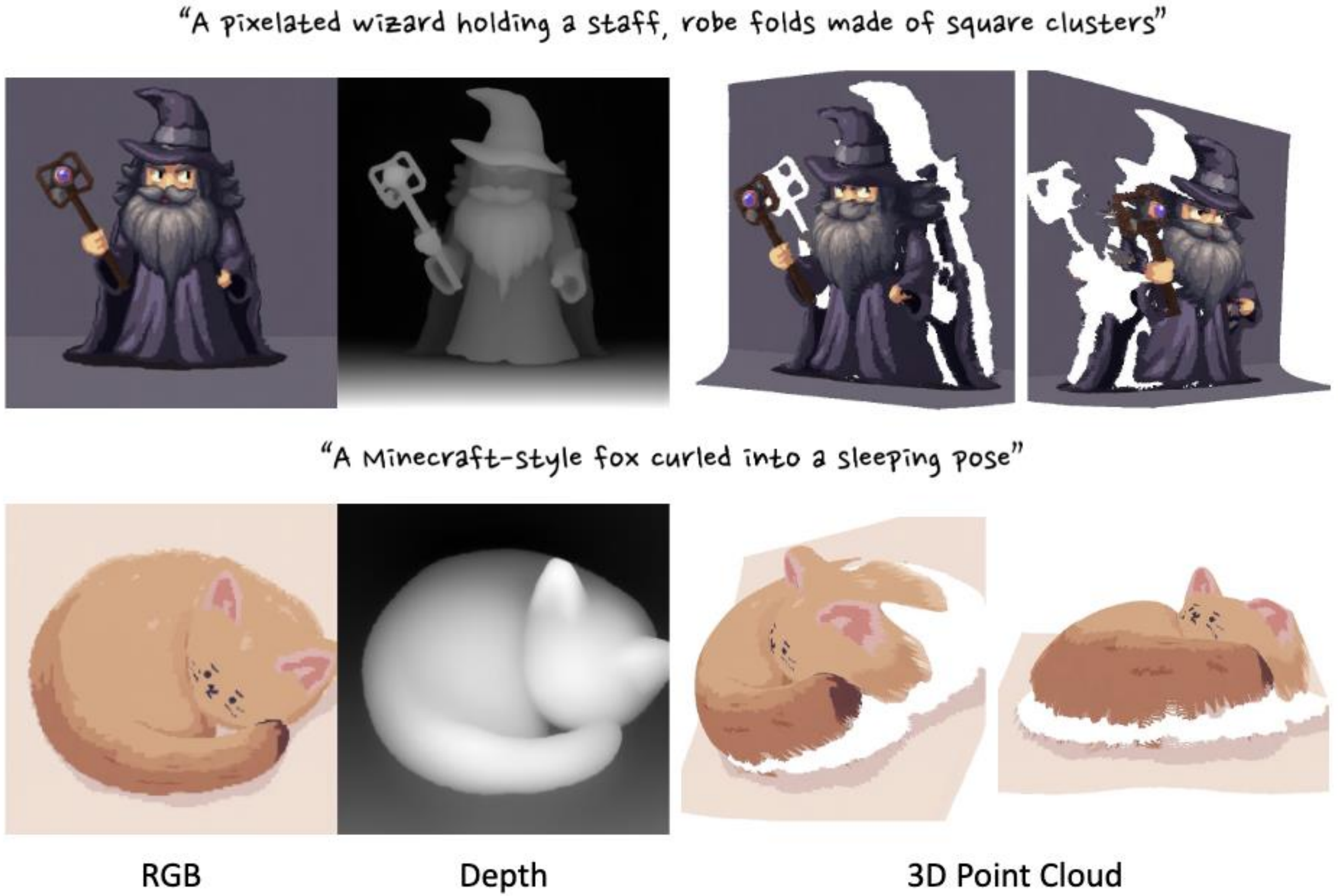
Depth condition | JointNet [2] | UniCon [6] | Ours | Original image

[2] Zhang, Jingyang, et al. "Jointnet: Extending text-to-image diffusion for dense distribution modeling," ICLR 2024.
[6] Li, Xirui, et al. "A simple approach to unifying diffusion-based conditional generation," ICLR 2025.

Additional Capability: Joint Panoramic Generation

Expansive view of an ancient Roman city with grand marble buildings, a massive colosseum, peoples, and lively markets..

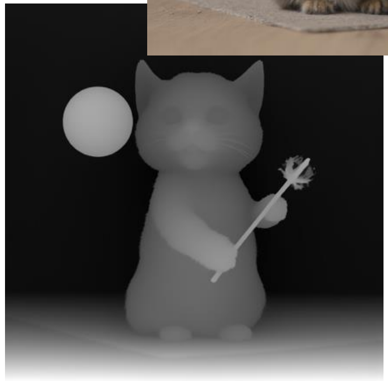
Additional Capability: 3D-aware Cartoon Content Generation

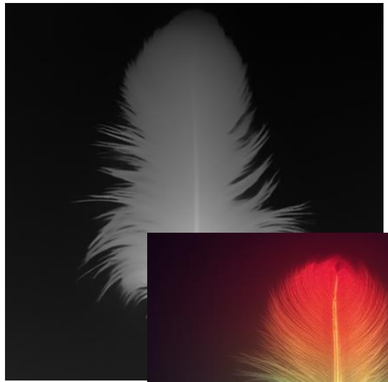


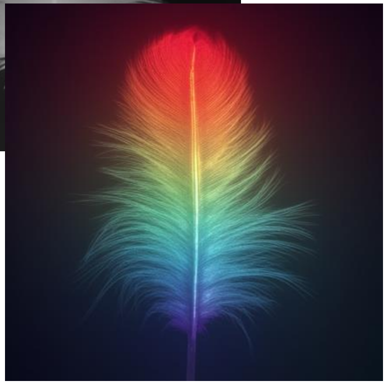
Conclusion

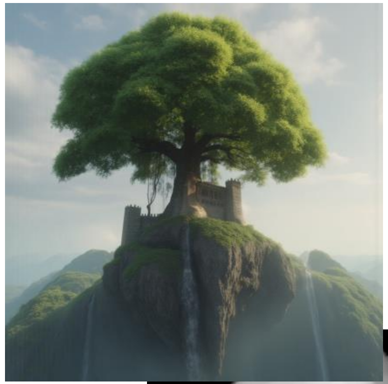
- We present JointDiT, a model for robust joint image-depth distribution modeling, enabling diverse tasks by controlling each branch's timestep:
    1. Joint generation
    2. Depth estimation
    3. Depth-conditioned image generation

- We propose the adaptive scheduling weights and unbalanced timestep sampling
→ Effective above three tasks!









Thank you

Project page