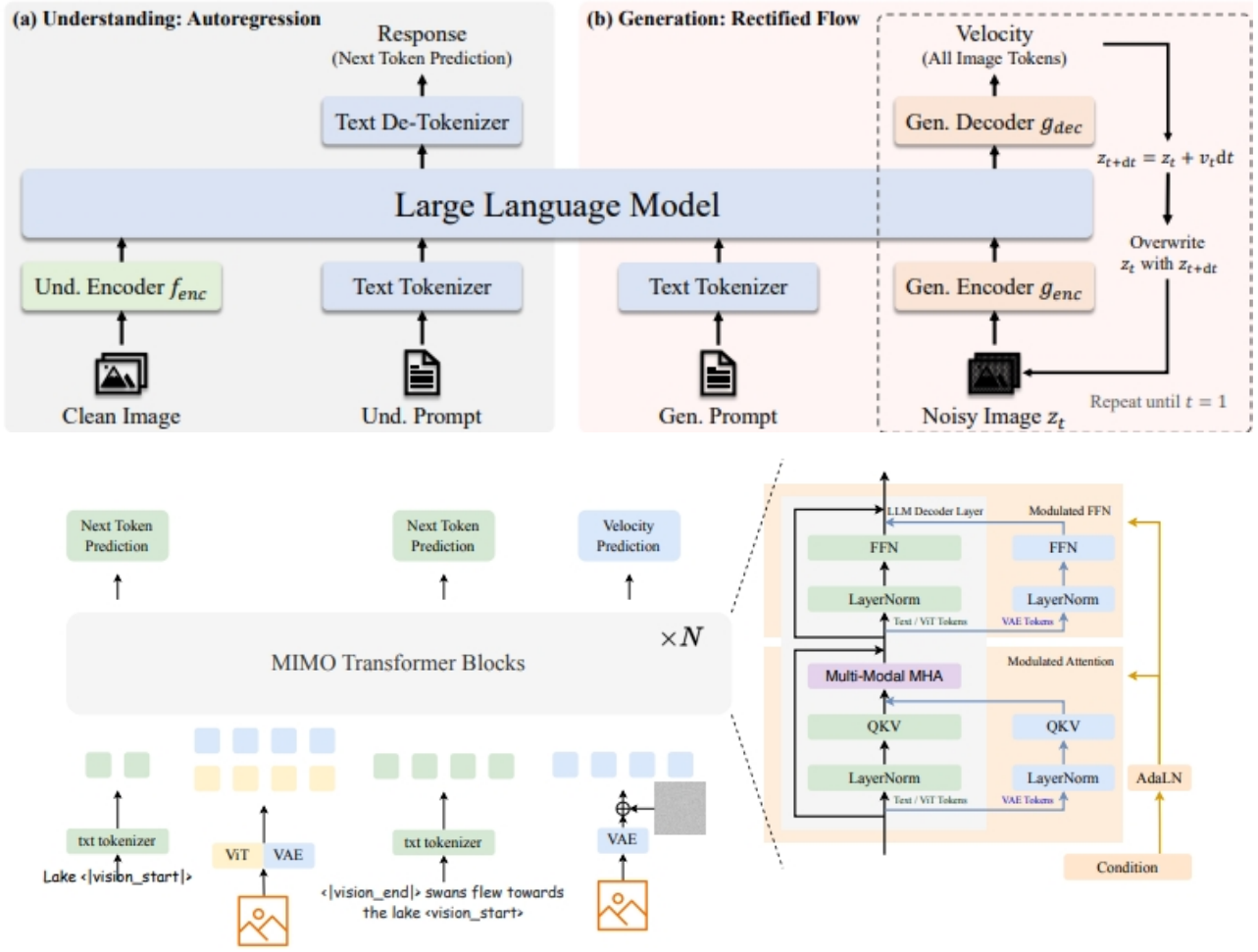
WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens



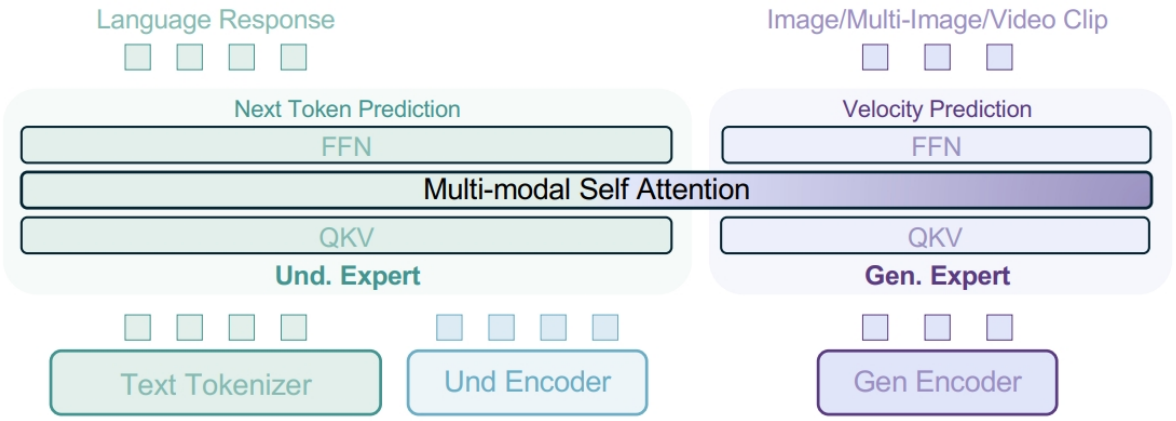
The Challenge: Early studies reveal severe task interference when using shared parameters for both visual understanding and generation.

The Shift: Following JanusFlow and mogao, the community focuses on efficiently co-existing or bridging generation and understanding within unified multimodal large models.

- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.

Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

■ Technical Paradigms: Mainstream approaches are gradually converging on the **Bagel** and MetaQuery families.



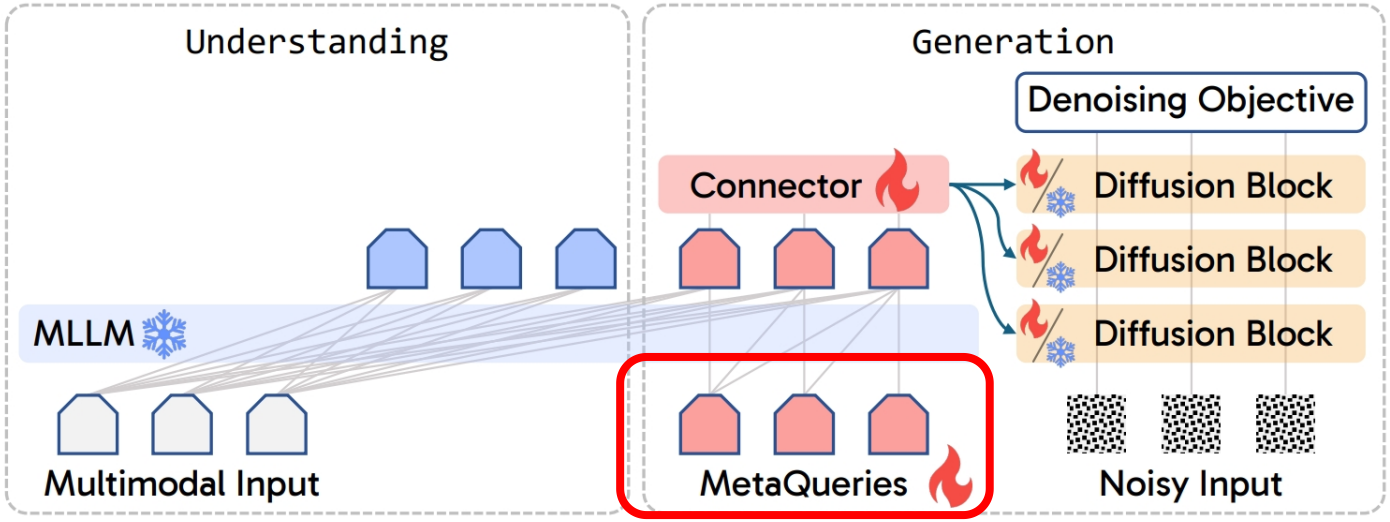
	Alignment	PT	CT	SFT
Hyperparameters				
Learning rate	1×10^{-3}	1.0×10^{-4}	1.0×10^{-4}	2.5×10^{-5}
LR scheduler	Cosine	Constant	Constant	Constant
Weight decay	0.0	0.0	0.0	0.0
Gradient norm clip	1.0	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1.0 \times 10^{-15}$)			
Loss weight (CE : MSE)	-	0.25 : 1	0.25 : 1	0.25 : 1
Warm-up steps	250	2500	2500	500
Training steps	5K	200K	100k	15K
EMA ratio	-	0.9999	0.9999	0.995
Sequence length per rank (min, max)	(32K, 36K)	(32K, 36K)	(40K, 45K)	(40K, 45K)
# Training seen tokens	4.9B	2.5T	2.6T	72.7B
Max context window	16K	16k	40k	40k
Gen resolution (min short side, max long side)	-	(256, 512)	(512, 1024)	(512, 1024)
Und resolution (min short side, max long side)	(378, 378)	(224, 980)	(378, 980)	(378, 980)
Diffusion timestep shift	-	1.0	4.0	4.0

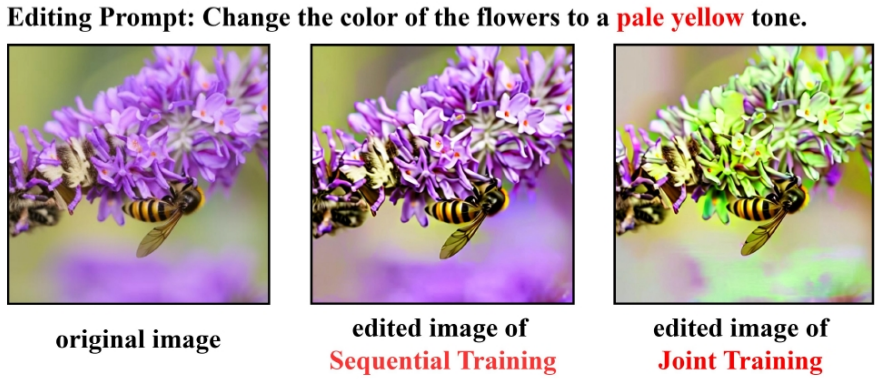
Yields stable training and outstanding performance, but incurs massive computational overhead due to the need for large-scale data and extensive training steps.

- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.

Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

■ Technical Paradigms: Mainstream approaches are gradually converging on the Bagel and MetaQuery families.





Task-rigid and struggles with continual learning.
(Requires modeling as a distribution instead of a point representation to prevent shortcut learning.)

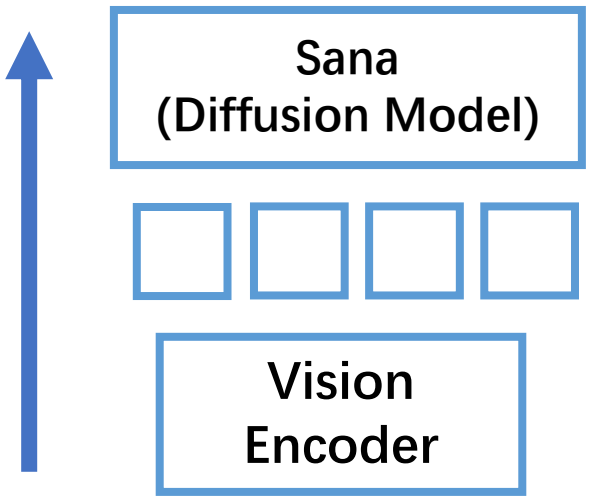
Features efficient training and architectural flexibility, but suffers from instability and a critical flaw of task generalization collapse within Learnable Queries.

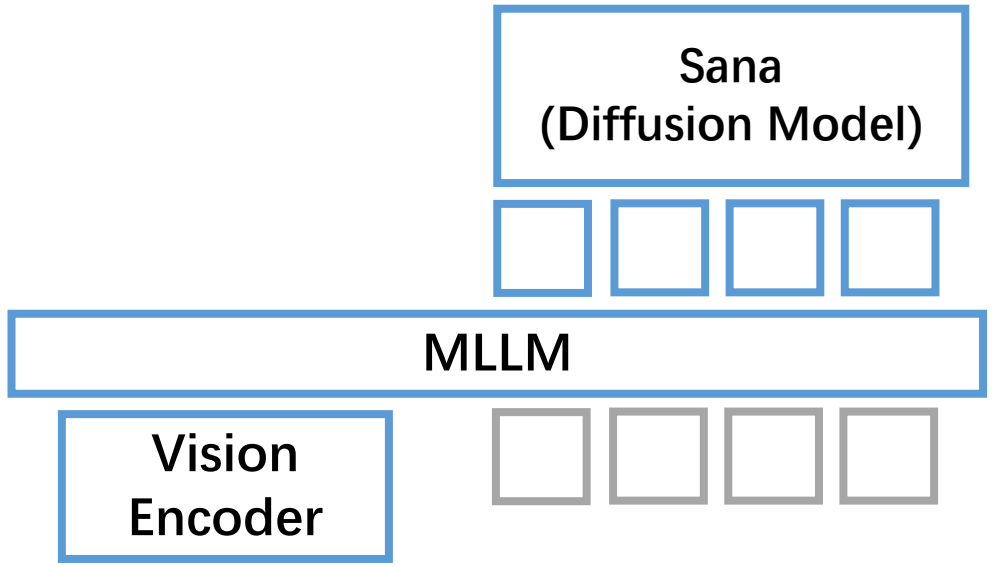
- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.

Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

■Low Editing Fidelity → Probing Fine-Grained Information Loss

Image Reconstruction: Flawless reconstruction achieved via an intermediate MLP layer with Sana training enabled.





Integration Issue: Upon integrating the pre-trained MLLM, the model fails to reconstruct fine-grained details (e.g., textures).

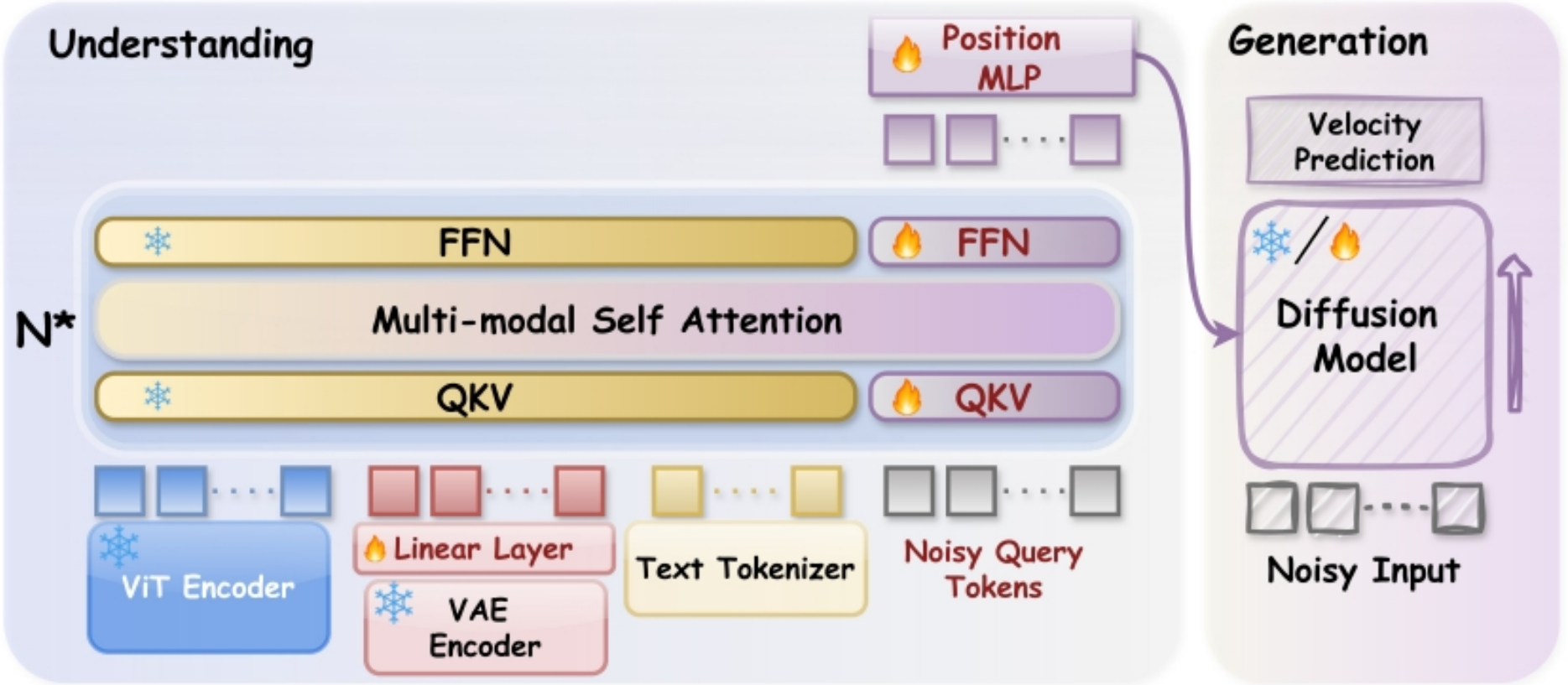
The Bottleneck: Pre-trained MLLMs inherently lose fine-grained details, rather than the Vision Encoder (Qwen 2.5VL ViT) failing to extract them.

Our Solutions: We address this by either (1) unfreezing the Vision Encoder to alter feature representations, or (2) introducing a fine-grained information injection branch.

- Yang et al. "WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens". **CVPR**, 2026.

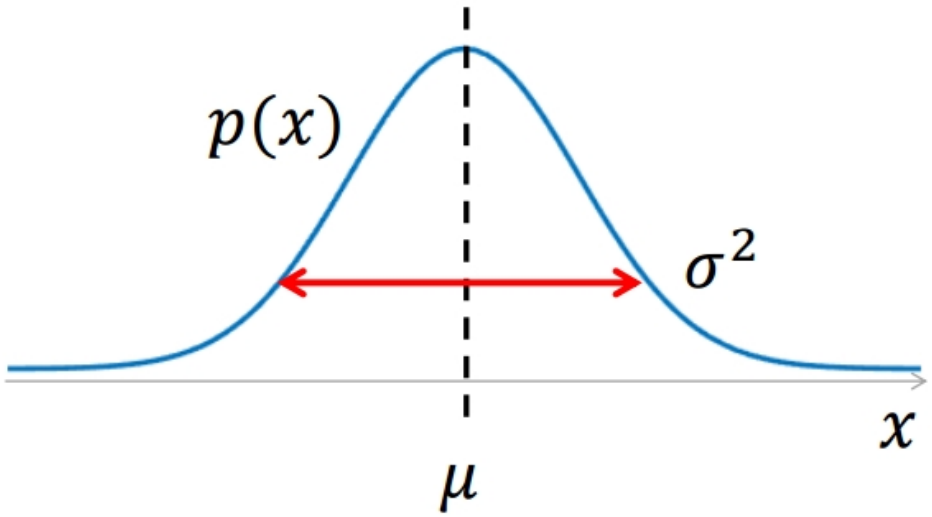
Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

- **Architecture Design:** A decoupled approach where the MLLM handles semantic integration and the Diffusion Model executes visual generation.



- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.

Architecture Design: Noise Query Tokens



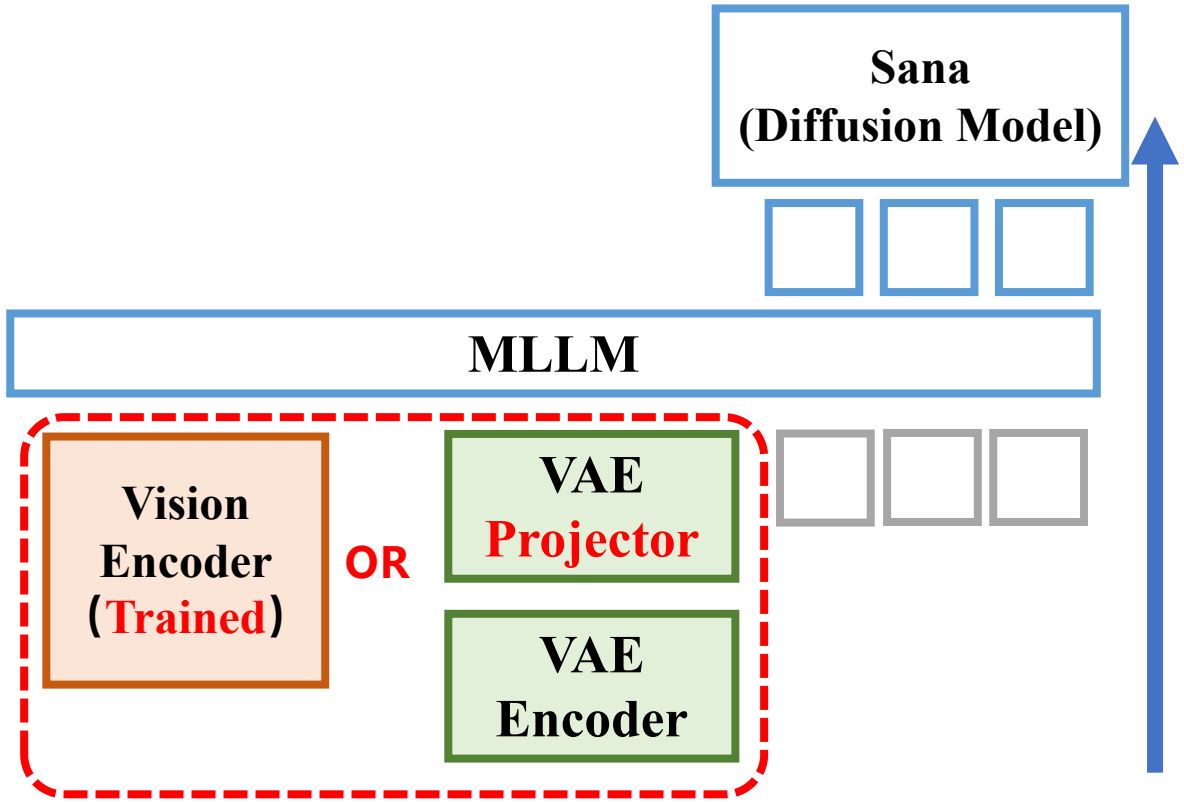
Our noise query tokens, matching the length of the Vision Encoder's features, are sampled from a Gaussian distribution and processed using **M-ROPE** and **bidirectional attention.**

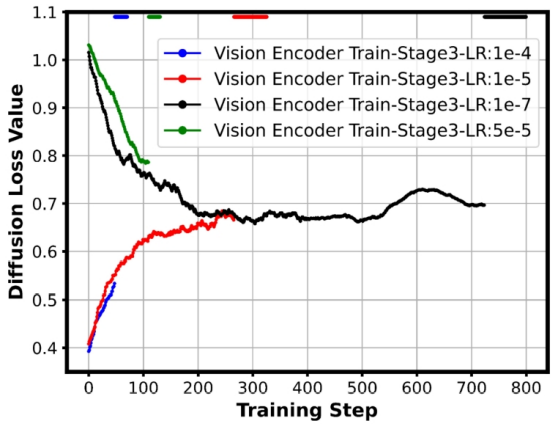
Key Insight: A learnable channel-wise scaling parameter yielded negligible effects ($\mu \approx 1$, $\sigma = 0.0074$) after training on 80M samples. Thus, we conclude that a **vanilla Gaussian distribution** is entirely sufficient.

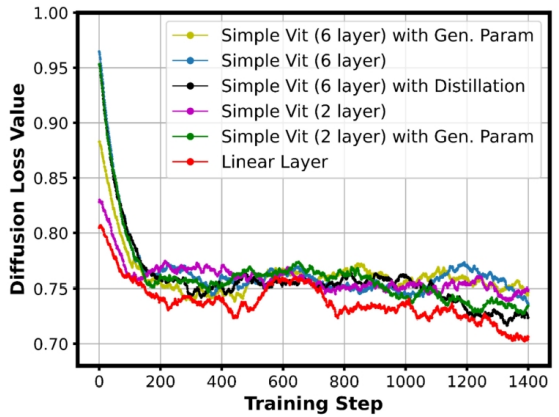
- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.

Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

■ Architecture Design: VAE Branch for Fine-Grained Information Injection







(1) Unfreezing ViT: Yields perfect reconstruction but suffers from instability and collapse during editing tasks (likely due to a brittle pre-trained state).

(2) VAE Injection: The most efficient approach, utilizing just a single linear layer for robust fine-grained information injection.

- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.

Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

Training Strategy: A Four-Stage Progressive Pipeline

Stage	Res.	Batch	Steps	Samples	LR	Warmup	Dataset Source	Task Mixture (Ratio)
1	512^2	2336	34k	~80M	$1.0e^{-4}$	1000	CC12M + LAION-Aesthetics	Rec.: 50, T2I: 47, Uncond.: 3
2	1024^2	584	44k	~25.6M	$1.0e^{-5}$	1500		Rec.: 50, T2I: 47, Uncond.: 3
3	1024^2	584	34k	~20M	$1.0e^{-5}$	1500	HQ Mix† + Uniworld-V1 (**Single**)	S-Edit: 55, T2I: 25, Rec.: 20, Uncond.: 10
4	1024^2	244	18k	~4.3M	$1.0e^{-5}$	1500	HQ Mix† + Uniworld-V1 (**Multi**)	S-Edit: 35, M-Edit: 20, T2I: 25, Rec.: 20, Uncond.: 10

† HQ Mix includes Blip3o, shareGPT-4o, and OpenGPT-4o.

Phase 1: Warm-up: Connector updated; MLLM & Sana frozen.

Phase 2: Joint Adaptation: Sana unfrozen for 1024-res conditional feature adaptation.

Phase 3: High-Fidelity Generation: Single-image editing + T2I generation.

Phase 4: Task Generalization: Multi-image editing tasks.

- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.

Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

Evaluating Editing Capabilities

Type	Method	Size	ImageEdit			GEdit-Bench-EN		
			Hybrid↑	Action↑	Overall↑	G_SC↑	G_PQ↑	G_O↑
Gen. Only	Gemini 2.5 Flash Image [28]	–	3.66	4.59	4.28	7.41	**7.96**	7.10
Unified	EMU3.5 [9]	34B	3.69	4.57	**4.41**	8.11	7.70	7.59
	GPT-4o [21]	–	**3.96**	**4.89**	4.2	7.85	7.62	7.53
	QWen-Image [33]	27B	3.82	4.69	4.27	8.00	7.86	7.56
	Bagel [10]	7B	2.38	4.17	3.2	7.36	6.83	6.52
	OmniGen2 [34]	7B	2.52	4.68	3.44	7.16	6.77	6.41
	UniWorld-V1 [16]	8B	2.96	2.74	3.26	4.93	7.43	4.85
	Query-Kontext [25]	17B	–	–	–	**8.36**	7.37	**7.66**
	WeMMU (Stage 3)	8B	2.82	3.15	3.31	5.86	6.80	5.75
	WeMMU (Stage 4)	8B	2.78	3.17	3.30	5.85	6.79	5.77

Ablation Study

Table 3. Ablation Study on Query Token Design using the ImageEdit Benchmark

Exp. Setting	Hybrid↑	Action↑	Overall↑
Learnable Fixed Query	1.87	2.21	2.53
Learnable Dynamic Query	2.02	2.60	2.88
Noisy Query	2.36	2.75	2.98
Noisy Query + VAE Branch	**2.82**	**3.15**	**3.31**

Image Generation Evaluation

Type	Method	Size	Geneval			DPG-Bench		
			Position↑	Color Attr.↑	Overall↑	Global↑	Entity↑	Overall↑
Gen. Only	FLUX.1-dev [39]	12B	0.20	0.47	0.67	82.1	89.5	84.0
	SD3-Medium [11]	2B	0.33	0.60	0.74	87.90	91.01	84.08
Unified	EMU3.5 [9]	34B	–	–	0.86	–	–	88.26
	QWen-Image [33]	27B	0.76	0.77	0.87	**91.32**	**91.56**	**88.32**
	Bagel* [10]	7B	0.78	0.77	**0.88**	88.94	90.37	85.07
	OmniGen2* [34]	7B	0.71	0.75	0.86	88.81	88.83	83.57
	UniWorld-V1* [16]	8B	0.74	0.71	0.84	83.64	88.39	81.38
	Query-Kontext* [25]	17B	0.85	**0.79**	**0.88**	–	–	–
	MetaQuery-XL* [22]	9B	–	–	0.80	–	–	82.05
	Bifrost-1 [17]	19B	–	–	0.81	–	–	77.67
	WeMMU (Stage 3)	8B	**0.86**	0.77	**0.88**	87.46	89.37	83.69
	WeMMU (Stage 4)	8B	0.85	0.78	**0.88**	87.66	89.07	83.60

- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.

Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

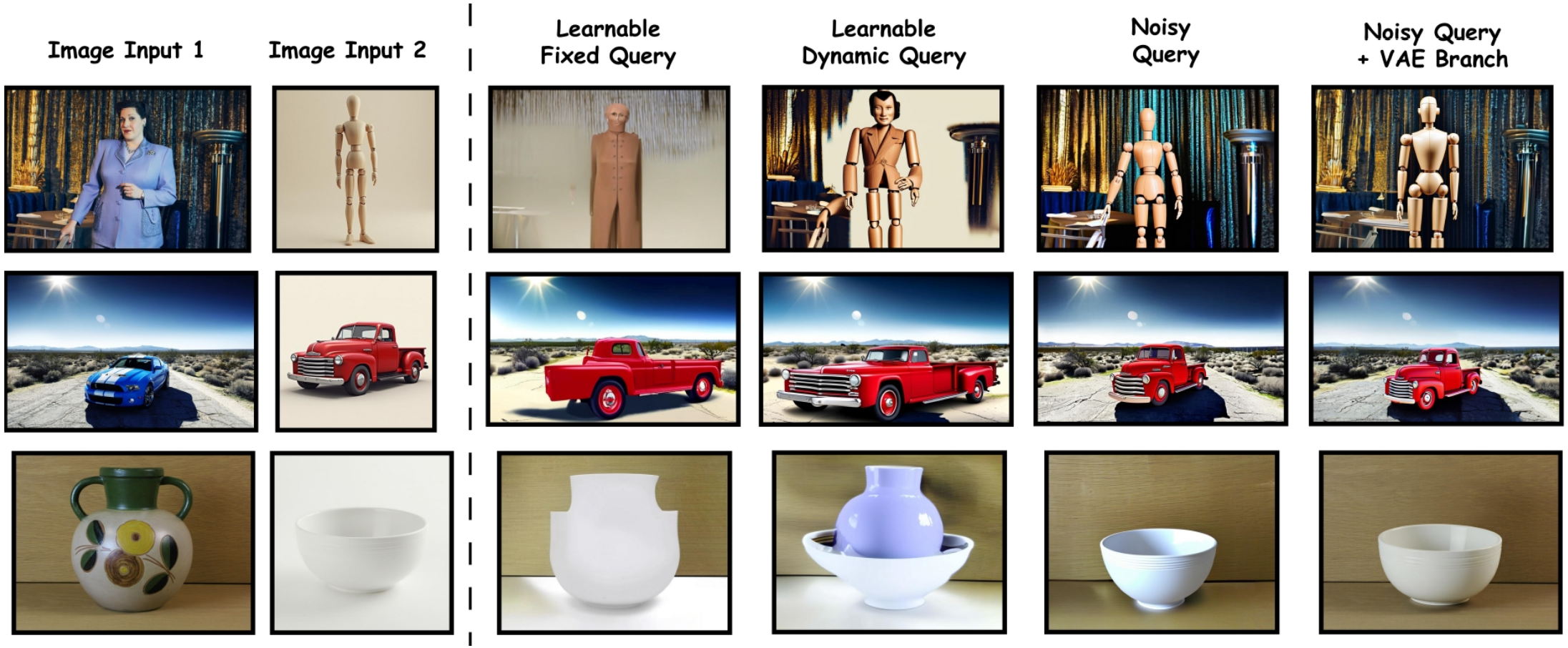
Attention Visualization of Noisy Query Tokens

Attention Gap Analysis: By examining the average attention difference between image and text condition tokens (yielding values of 1.80, 1.01, -0.99, and -0.68 from left to right), we demonstrate that introducing **Noisy Query Tokens** progressively diminishes the attention disparity across modalities.

- Yang et al. "WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens". **CVPR**, 2026.

Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

Investigating Task Generalization: Scaling to Multi-Image Editing



- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.

Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

Visual Examples of Generation and Editing

- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.

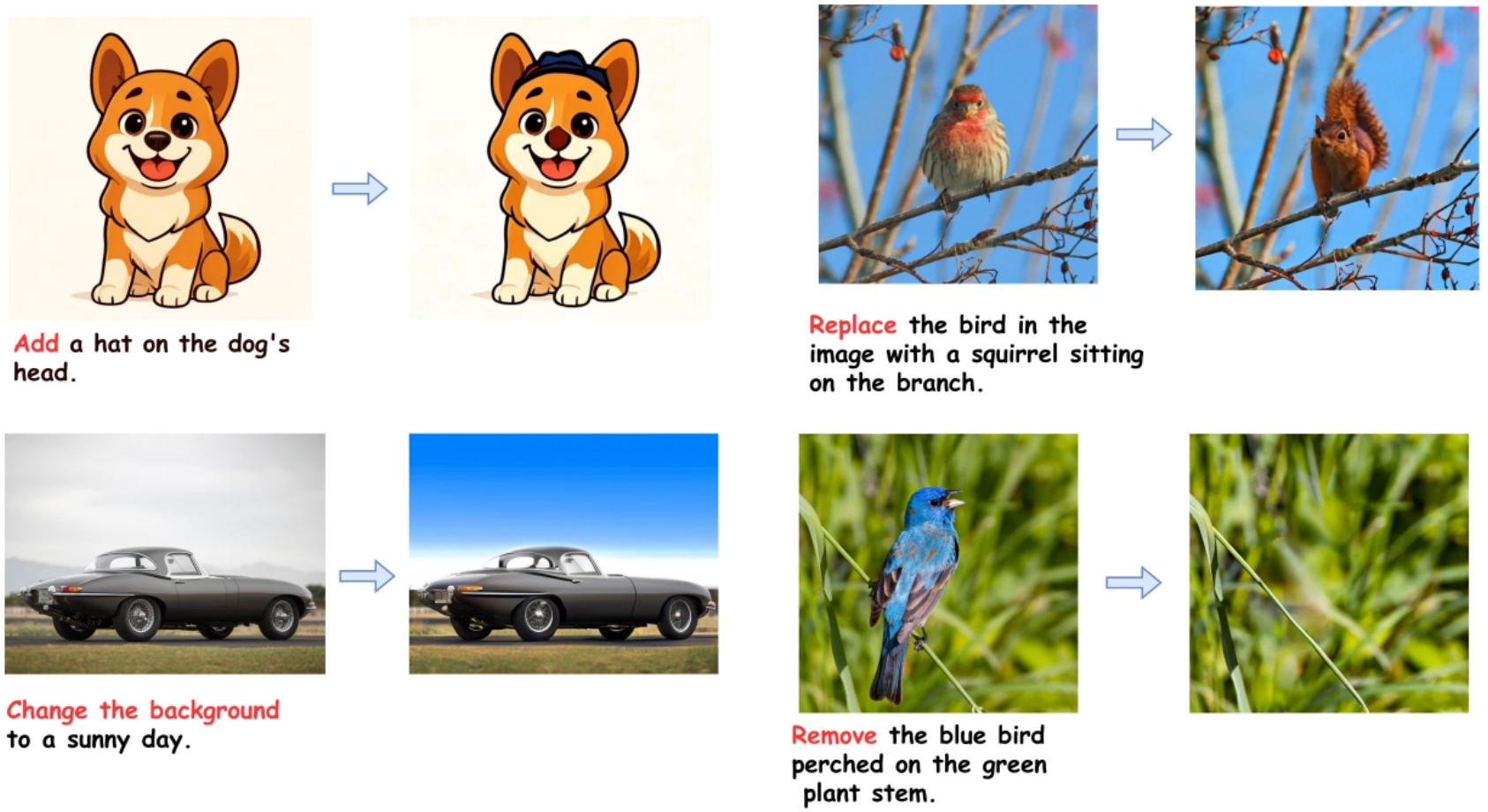
Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

Visual Examples of Generation and Editing

- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.

Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

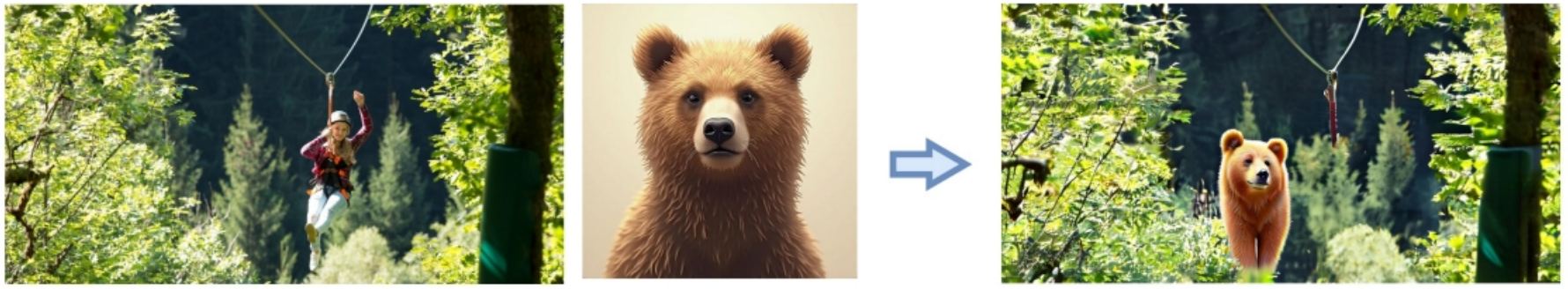
Visual Examples of Generation and Editing

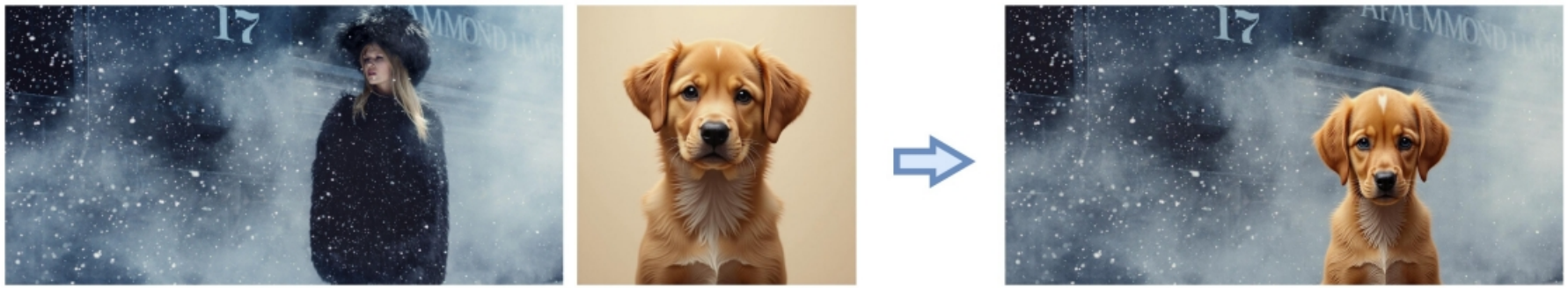


- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.

Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens

Visual Examples of Generation and Editing





- Yang et al. “WeMMU: Enhanced Bridging of Vision-Language Models and Diffusion Models via Noisy Query Tokens”. **CVPR**, 2026.