



XVerse: Consistent Multi-Subject Control of Identity and Semantic Attributes via DiT Modulation

Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen,
Xu Wang, Kang Du, Xinglong Wu

Project Page: <https://bytedance.github.io/XVerse>

GitHub: <https://github.com/bytedance/XVerse>



Contents

01 Background & Motivation

02 Model Architecture

03 Benchmark & Evaluation

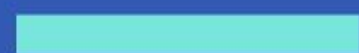
04 Conclusion & Outlook





01

Background & Motivation





From Single to Multi-Subject Personalization

- Rapid progress in T2I models enables realistic and diverse image synthesis from text.
- Early personalization methods focused on **single-subject** control (e.g., DreamBooth, IP-Adapter).
- Growing demand for **multi-subject** and complex scene generation with fine-grained control.





Existing Subject-controlled Generation

- Existing techniques inject reference images via attention mechanisms. This often disrupts the base model, leading to:



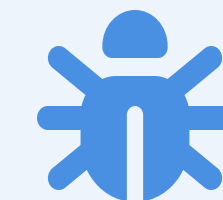
Identity Leakage

Features bleed between subjects.



Attribute Entanglement

Pose & style become coupled.



Structural Artifacts

Direct attention injection warps the image.

- Text-stream modulation offers a separate, semantic-aligned control pathway.



Fine-Grained Multi-Subject Control

- By transforming reference images into offsets for token-specific text-stream modulation, XVerse provides:
 - High-fidelity, editable **multi-subject** image synthesis
 - Powerful control over **individual subject characteristics**
 - Fine-grained manipulation of **semantic attributes**





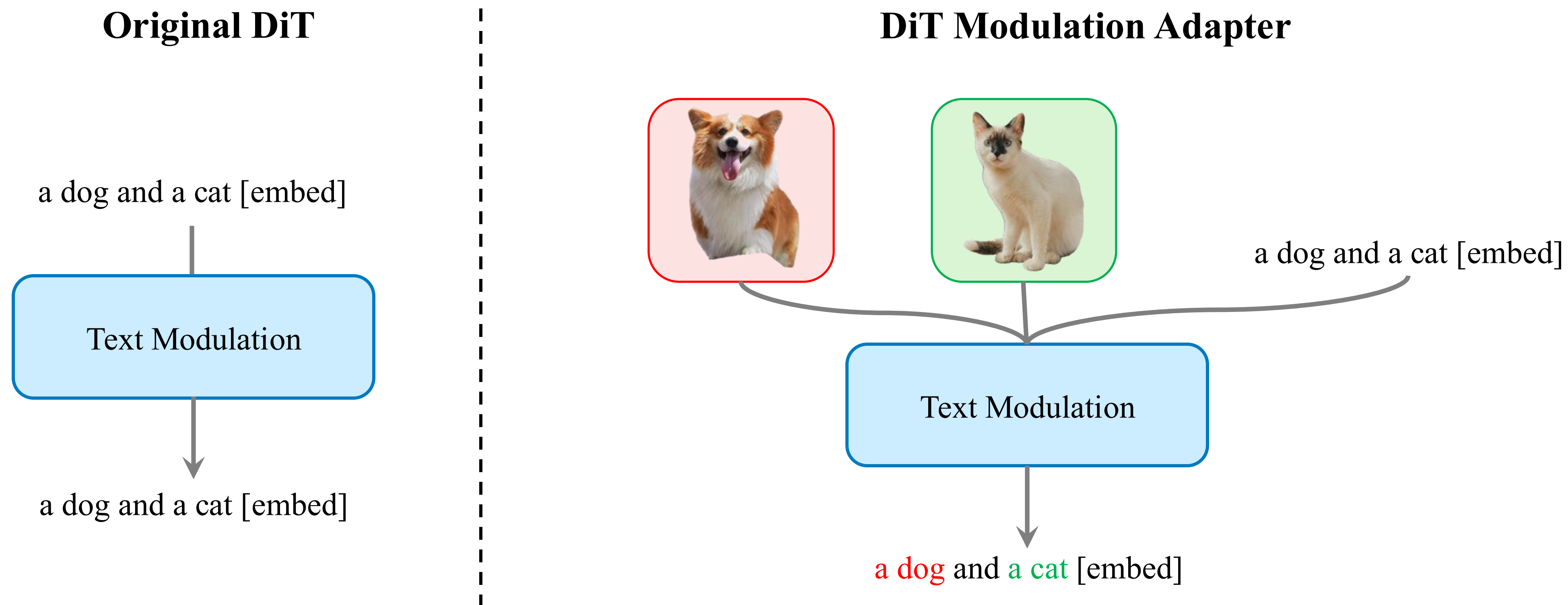
02

Model Architecture



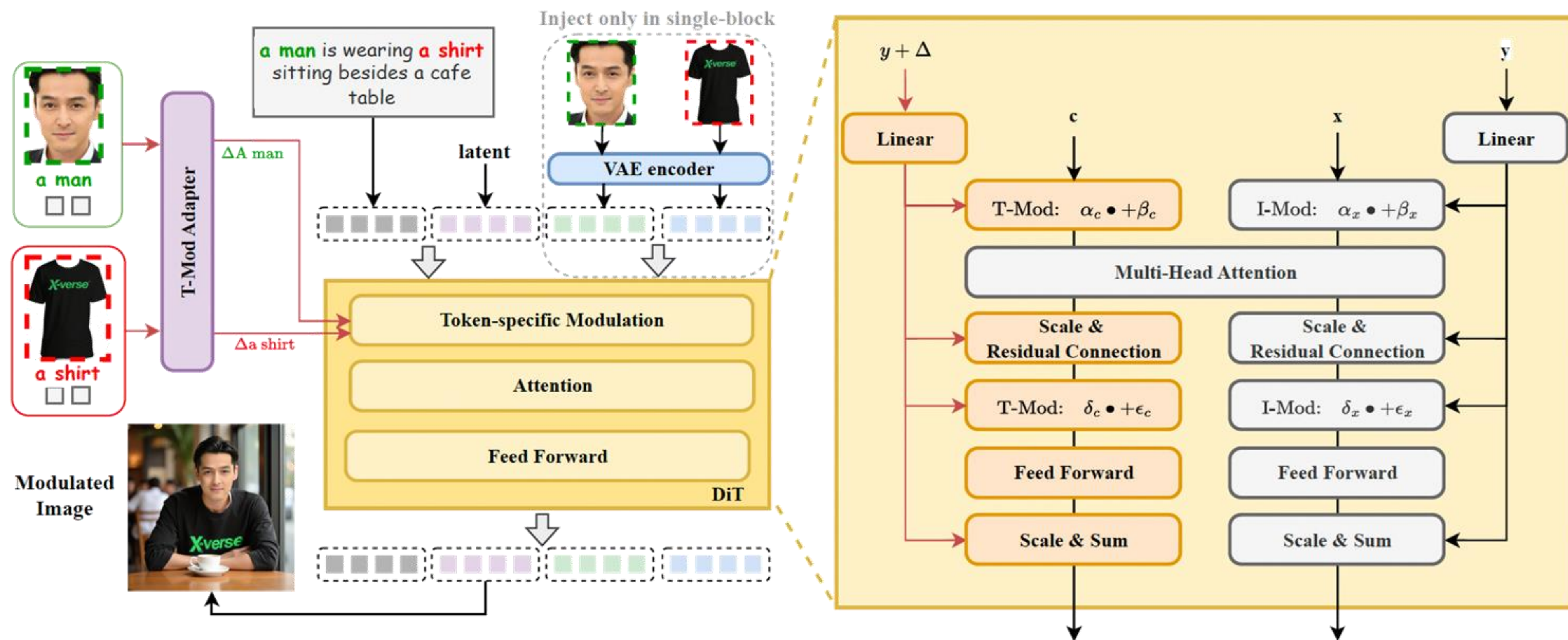


Core Innovation: Text-Stream Modulation Adapter





Core Innovation: Text-Stream Modulation Adapter

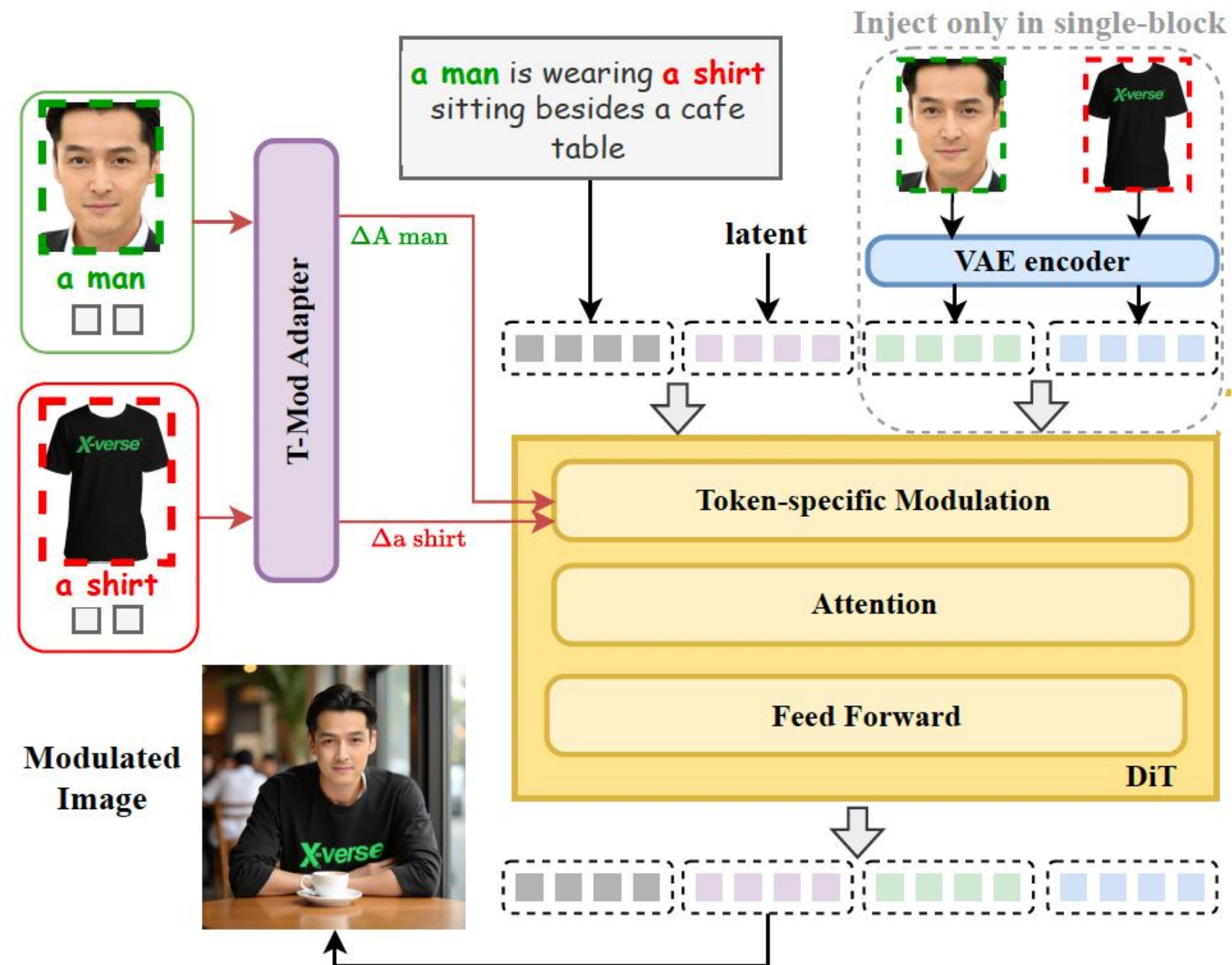


Perceiver Resampler generates modulation offsets Δ across from CLIP features.

Offsets adjust **text-stream modulation** (AdaLN) per token.

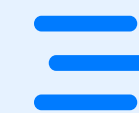


VAE-Enhanced Detail Injection



Auxiliary Cue

VAE features act as a supporting signal, not the main driver, to avoid artifacts.



Single-Stream Injection

Features are injected only into FLUX's single-stream blocks to supplement detail without disrupting the main generation flow.

VAE features capture fine-grained texture and detail.



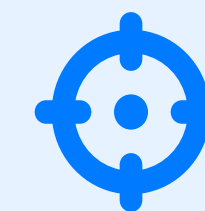
Regularization for Disentanglement



Region Preservation Loss (RPL)

Enforces consistency with the vanilla T2I output in unmodulated regions, preventing subject feature bleed.

$$L_{region} = ||(1 - M_c) \odot (V_{\theta'} - V_{\theta})||^2$$



Text-Image Attention Loss (TIAL)

Aligns cross-attention maps between modulated and T2I branches, preserving semantic locality and editability.

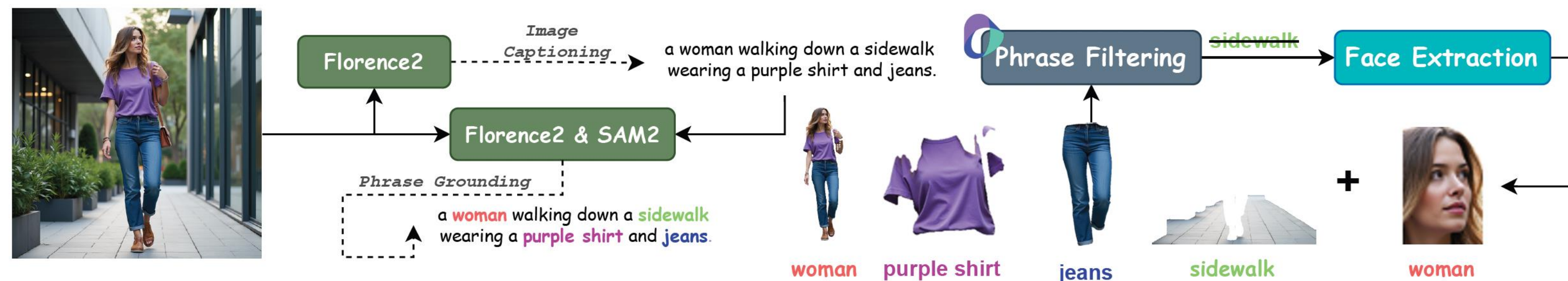
$$L_{attn} = ||Attn_{\theta'} - Attn_{\theta}||^2$$

Together, they suppress **subject fusion** and stabilize training.

Collection of Training data

Single-Image Dataset

- Utilizes **Florence2** for image description and phrase grounding
- Filters the labels by **LLMs**
- Employs **SAM2** for subject segmentation, and a **Face Detector** for face extraction
- Constructs a high-quality training dataset containing multi-subject control



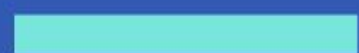
Cross-Image Dataset

- Collect from Subject-200K, and in-house single person multi-view dataset
- Use **DINO-V2** or **ArcFace** for similarity filtering



03

Benchmark & Evaluation

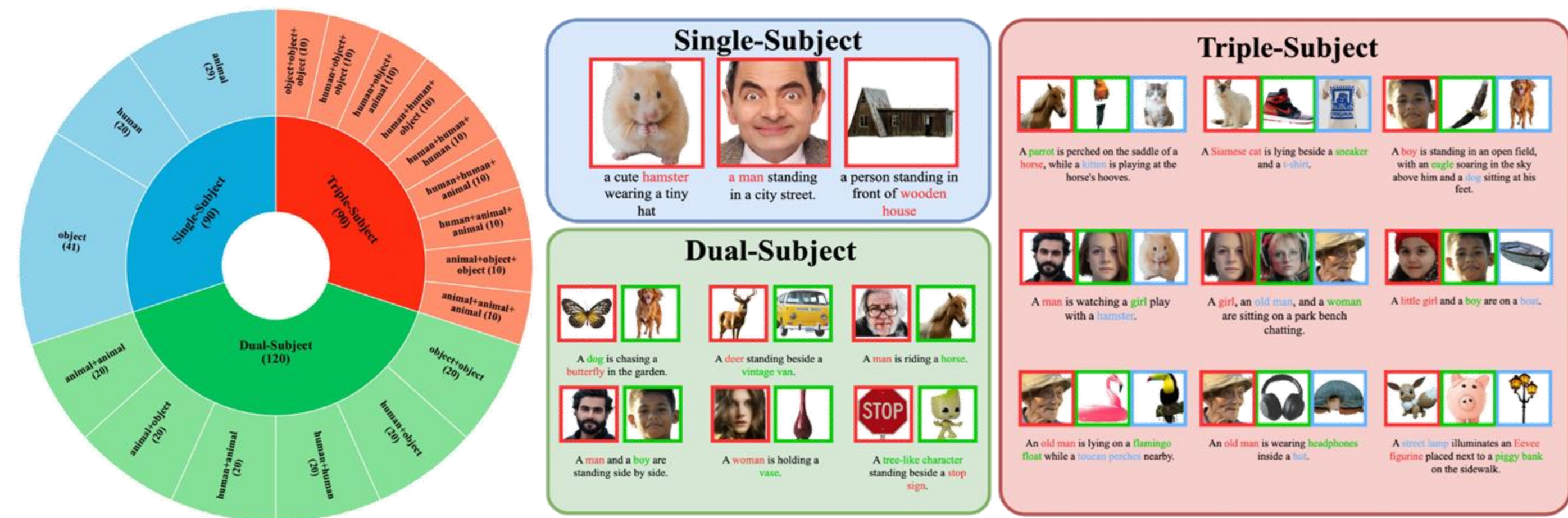




XVerseBench

- To comprehensively evaluate **multi-subject control image generation** capabilities, we propose the XVerseBench benchmark.

- 👤 20 different human identities
- 🚗 74 unique objects
- 🐾 45 different animal species/individuals
- 📝 A total of 300 unique test prompts



Evaluation Metric	Description
DPG Score	Evaluates the model's editing capability
Face ID Similarity	Evaluates the model's ability to maintain human identity
DINOv2 Similarity	Evaluates the model's ability to maintain object features
Aesthetic Score	Evaluates the aesthetic quality of generated images






































Comparison with State-of-the-Art

- XVerse achieves the highest Overall score on XVerseBench.
- In single-subject generation, XVerse attains better subject similarity.
- XVerse demonstrates a more significant performance advantage in challenging multi-subject generation.

Method	Single-Subject					Multi-Subject					Overall
	DPG	ID-Sim	IP-Sim	AES	AVG	DPG	ID-Sim	IP-Sim	AES	AVG	
MS-Diffusion [36]	96.94	6.58	51.06	59.69	53.57	87.27	4.81	40.90	55.87	47.21	50.39
MIP-Adapter [37]	77.48	28.39	66.32	52.09	56.07	84.52	19.49	49.89	51.78	51.42	53.75
OmniGen [38]	85.19	<u>60.17</u>	70.73	51.89	67.00	81.71	42.18	52.11	51.35	56.84	61.92
UNO [13]	91.82	37.22	74.35	55.21	64.65	87.57	26.00	60.62	53.04	56.81	60.73
DreamO [14]	97.51	58.74	67.69	53.80	<u>69.44</u>	<u>89.75</u>	<u>44.21</u>	<u>60.87</u>	51.16	<u>61.50</u>	<u>65.47</u>
XVerse (Ours)	93.50	63.02	<u>71.35</u>	<u>56.63</u>	71.13	91.77	51.03	61.04	<u>53.68</u>	64.38	67.76



Comparison with State-of-the-Art

Inputs	XVerse	DreamO	UNO	OmniGen	MIP	MS-Diffusion
 a person standing in front of a hut						
 A man and a woman standing side by side in a park.						
 A little girl is standing beside a man .						
 A man is walking with a dog on the street.						
 A robin is perched on a branch above while a white tiger slowly approaches a curious kitten in a forest.						



04

Conclusion & Outlook





Conclusion

Summary

- **XVerse pioneers text-stream modulation offsets for gentle, precise control.**
- **Auxiliary VAE cues and dual regularizers ensure high fidelity and disentanglement.**
- **Achieves SOTA performance on the comprehensive XVerseBench.**

Future Work

- **Address the need for large-scale, high-quality cross-image datasets.**
- **Explore image-side modulation for pixel-level regional editing.**
- **Further mitigate potential societal risks like deepfakes and bias amplification.**



THANKS.



Project Page: <https://bytedance.github.io/XVerse>

GitHub: <https://github.com/bytedance/XVerse>