EchoGen: Generating Visual Echoes in Any Scene via Feed-Forward Subject-Driven Auto-Regressive Model

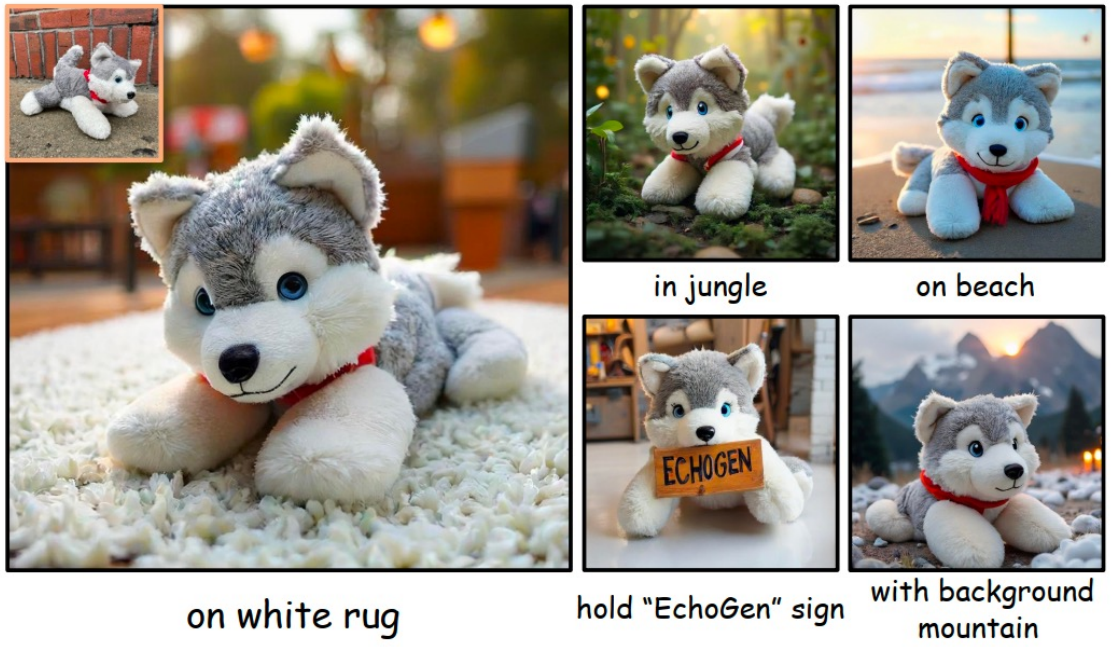
Ruixiao Dong[1,2*], Zhendong Wang[1*], Keli Liu[1], Li Li[1†], Ying Chen[2†],
Kai Li[2], Daowen Li[2], Houqiang Li[1]
[1]University of Science and Technology of China
[2]Alibaba Group

中国科学技术大学
University of Science and Technology of China

Alibaba

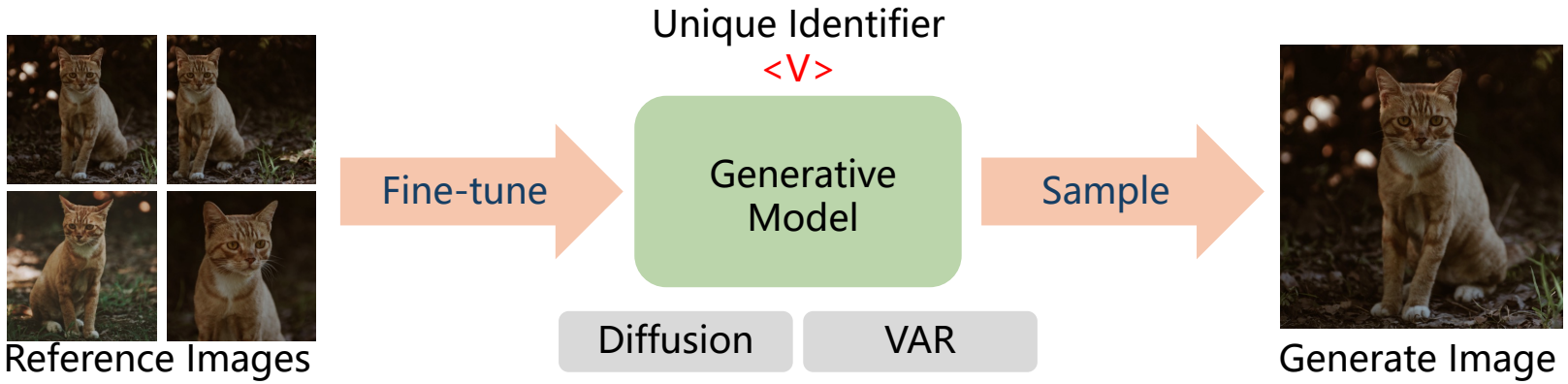
Background

- Subject-driven Image Generation
  - Input: one or more reference images of a subject;
  - Goal: generate the same subject in novel scenes following text prompts;
  - Challenge: preserve subject identity while following diverse instructions.

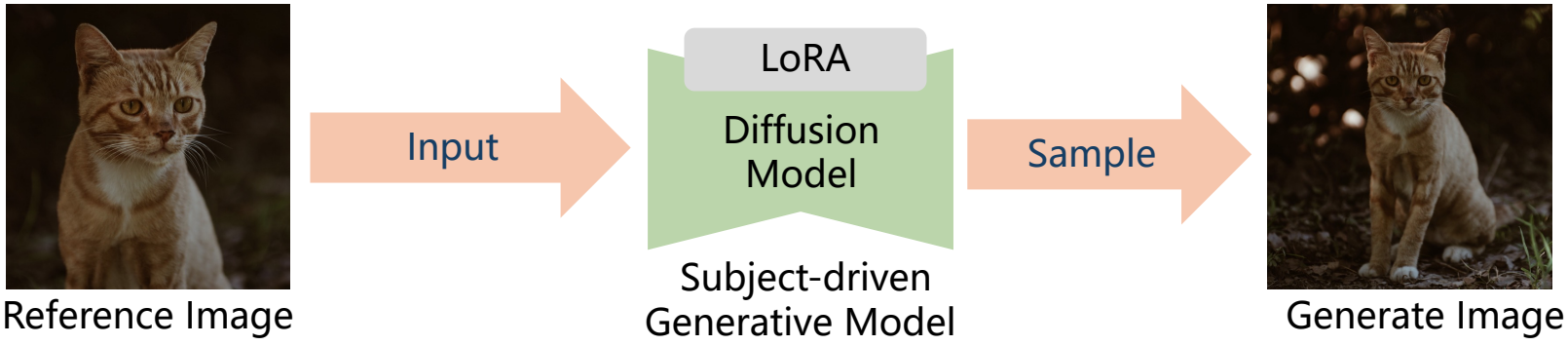


Background

- Subject-driven Image Generation Methods
  - Test-Time Fine-tuning Methods



  - Feed-Forward Diffusion Methods



Problem Definition

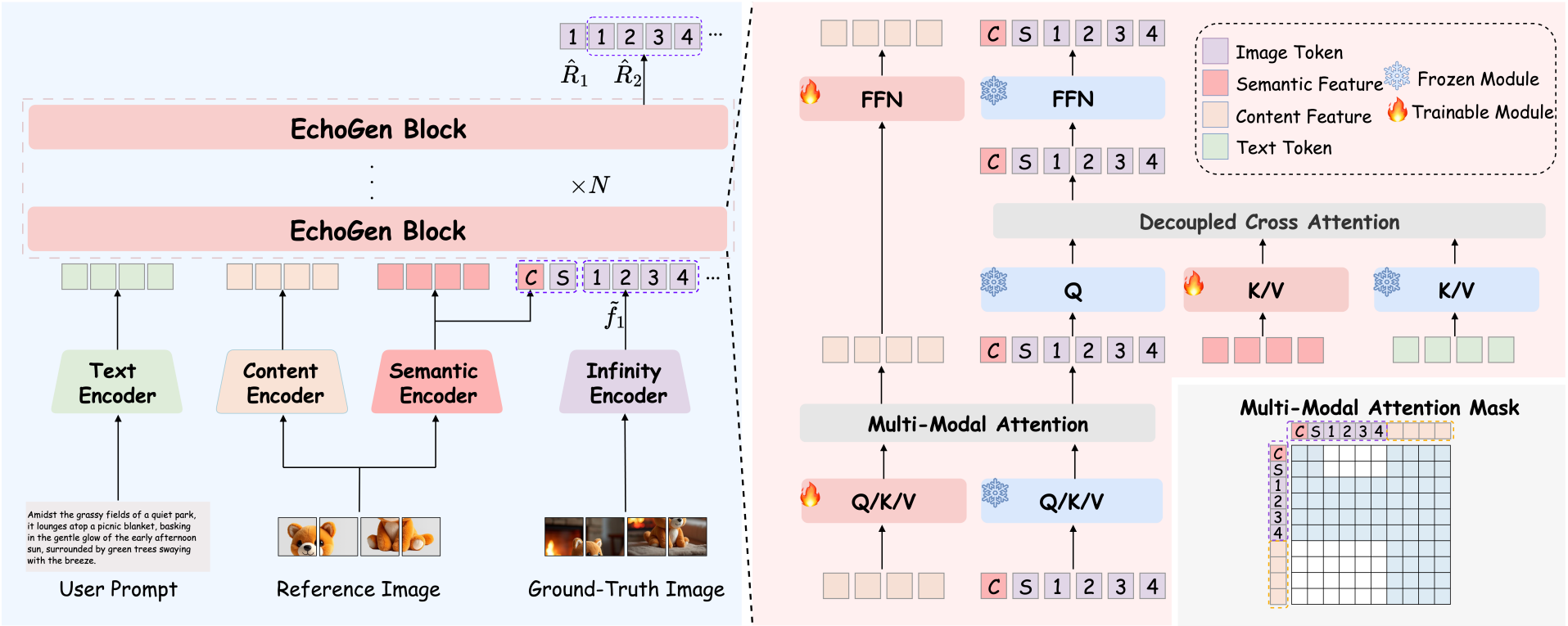
- Problems
  - Test-Time Fine-Tuning Methods: **very slow due to per-subject optimization;**
  - Feed-Forward Diffusion Methods: **still slow due to iterative sampling**.
- Our Solution—**EchoGen**
  - **Feed-forward subject-driven generation built on Visual Auto-Regressive (VAR).**
  - Naïve prompts alone are insufficient; **dual-path subject injection** preserves subject identity.
  - Subject Segmentation and Subject-Text CFG further improve quality and controllability.

	Inference Speed	**Quality**
Test-Time Fine-Tuning Methods	Very Slow	Good
Feed-forward Diffusion Methods	Slow	Good
VAR with Only Naïve Prompts	Fast	Bad
Ours	**Fast**	**Good**

EchoGen

ICLR

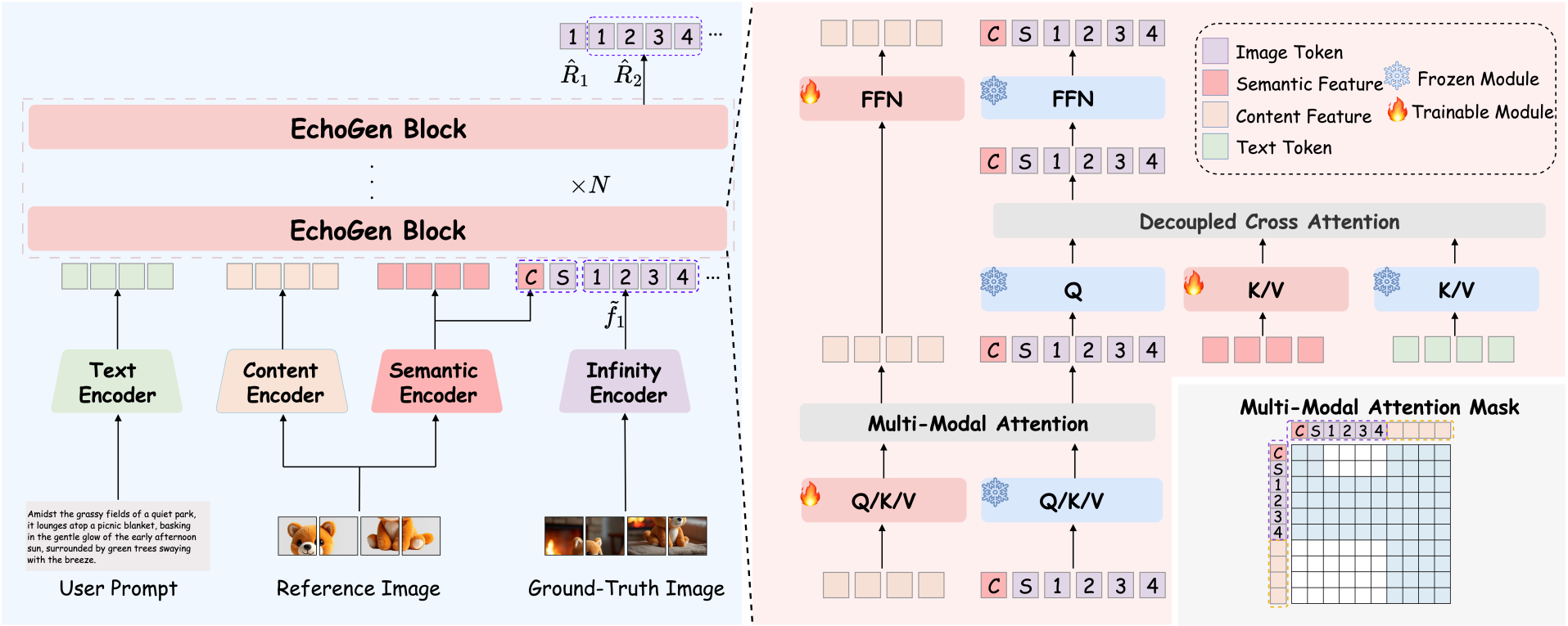
- Dual-Path Subject Injection: Semantic Path
  - Semantic features for structural coherence, extracted with the DINOv2 model.
    - Inject fine-grained semantic features via decoupled cross-attention;
    - Inject global semantic features through prefixing and AdaLN.



EchoGen

Dual-Path Subject Injection: Content Path

- Content features for local details are extracted with the FLUX.1-dev VAE.
  - Multi-modal attention injects these features into generation process.

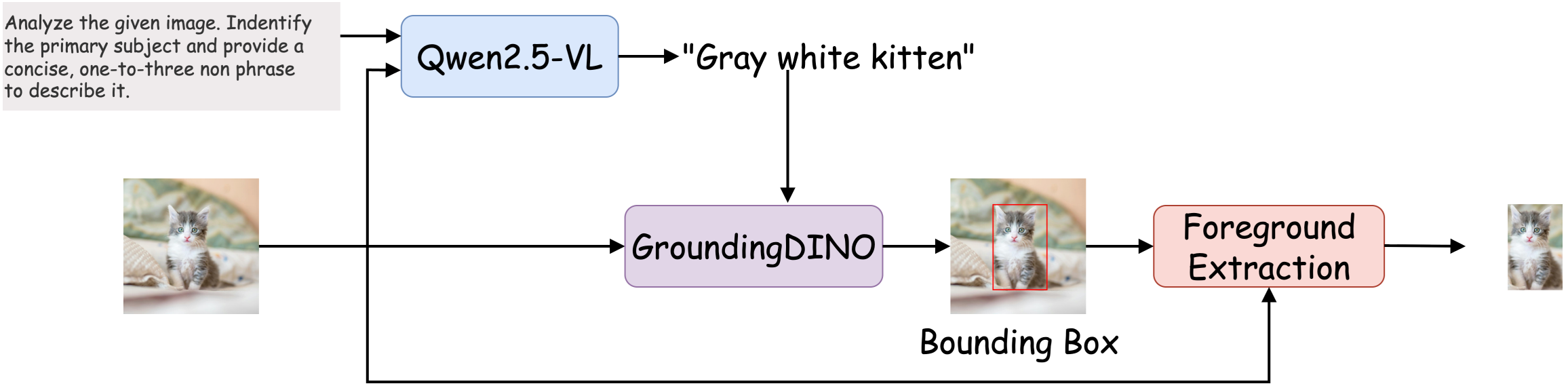


EchoGen

ICLR

Subject Segmentation

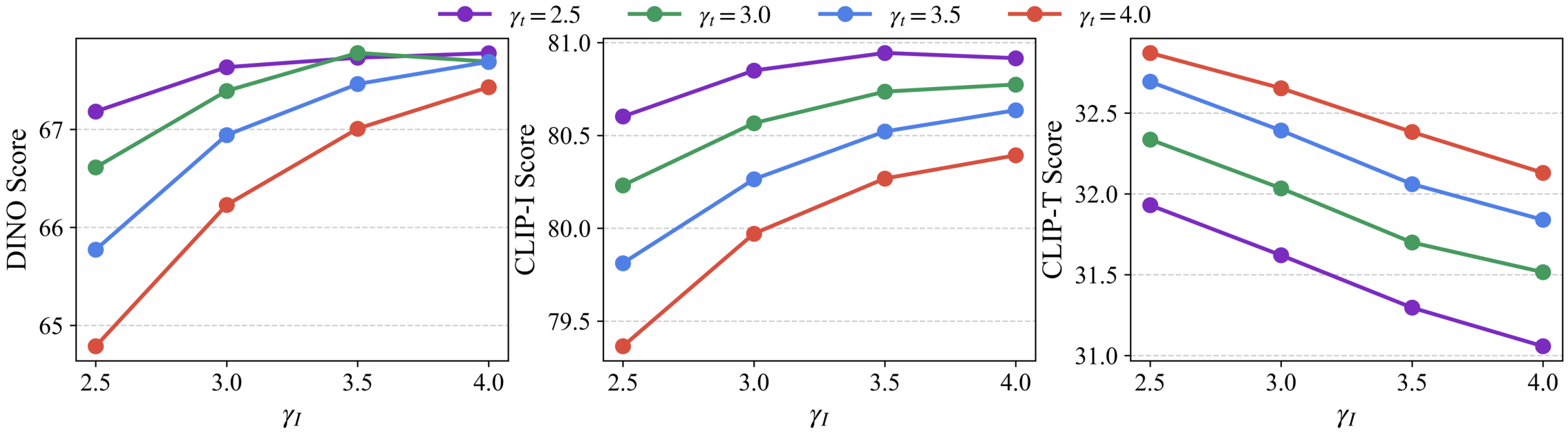
- Pre-process reference images with subject segmentation to remove distracting backgrounds;
- Use Qwen2.5-VL and GroundingDINO for subject segmentation.



EchoGen

- Sampling with Subject-Text Classifier-free Guidance
  - Enhance control over subject features and text prompts.
  - Control the trade-off between subject fidelity and text alignment.

$$\hat{l} = l(\Phi_t, \Phi_s, \Phi_c) + \gamma_t \times (l(c_t, \Phi_s, \Phi_c) - l(\Phi_t, \Phi_s, \Phi_c)) + \gamma_I \times (l(c_t, c_s, c_c) - l(c_t, \Phi_s, \Phi_c))$$



Results

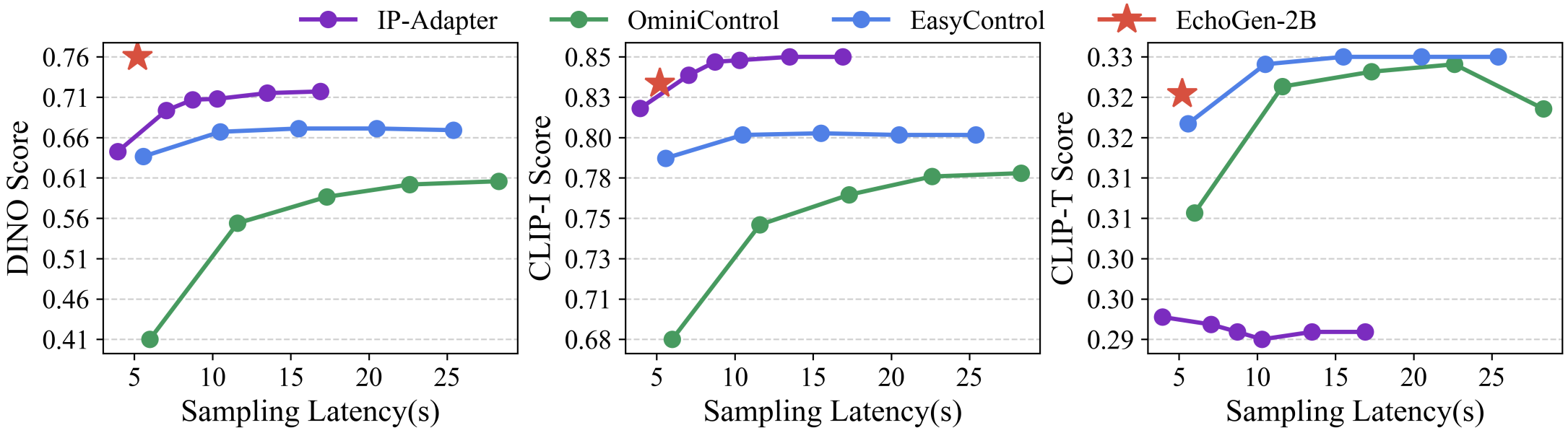
Quantitative Comparison

- EchoGen achieves **competitive or superior quality with substantially lower latency.**

Method	Base Model	DINO↑	CLIP-I↑	CLIP-T↑	Latency ↓
Test-time Fine-tuning					
Textual-Inversion	SD-v1.5	0.569	0.780	0.255	50min
DreamBooth	SD-v1.5	0.668	0.803	0.305	15min
BLIP-Diffusion	SD-v1.5	0.670	0.805	0.302	-
AR-Booth	Infinity-2B	**0.750**	**0.808**	0.269	2.8h
Unified Generation					
OmniGen	OmniGen	**0.693**	0.801	0.315	93.4s
Feed-Forward					
ELITE	SD-v1.4	0.621	0.771	0.293	**11.0s**
Re-Imagen	Imagen	0.600	0.740	0.270	-
BLIP-Diffusion	SD-v1.5	0.594	0.779	0.300	-
λ-Eclipse	Kan-v2.2	0.613	0.783	0.307	-
MS-Diffusion	SDXL	0.671	0.792	0.321	39.6s
IP-Adapter	SDXL	0.613	**0.810**	0.292	16.9s
IP-Adapter	FLUX.1-dev	0.561	0.725	**0.351**	-
OminiControl	FLUX.1-dev	0.684	0.799	0.312	27.5s
EasyControl	FLUX.1-dev	0.652	0.789	**0.325**	25.4s
EchoGen-0.1B	Infinity-0.1B	0.675	0.806	0.321	**0.5s**
EchoGen-2B	Infinity-2B	**0.755**	**0.835**	**0.325**	**5.2s**

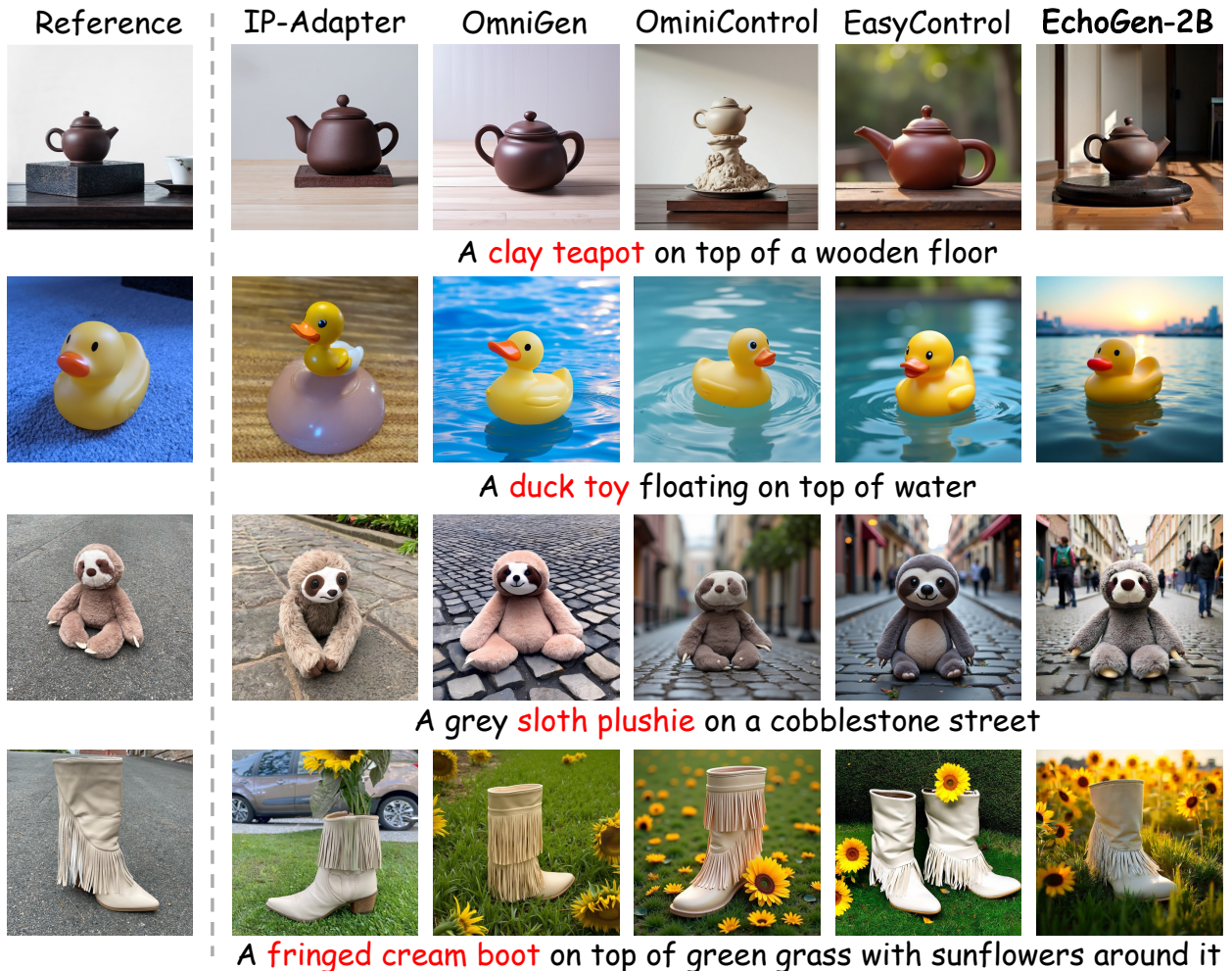
Results

- Performance-Latency Trade-offs Comparison
  - EchoGen-2B consistently **offers a strong trade-off between quality and latency**.



Results

- Qualitative Comparison
  - EchoGen exhibits advantages in both subject fidelity and instruction alignment.



Results

- Human Evaluation
  - EchoGen-2B **ranks first in subject fidelity and photorealism**, and remains competitive in text alignment.

Method	Subject Fidelity↑	Text Alignment↑	Photorealism↑
OmniGen	0.15	0.13	0.09
IP-adapter	0.21	0.05	0.14
OminiControl	0.12	0.21	0.15
EasyControl	0.15	**0.31**	0.28
EchoGen-2B	**0.37**	0.30	**0.34**

Results

- Ablations: Quantitative
  - Quantitatively validate the effectiveness of dual-path subject injection.

Encoder	DINO↑	CLIP-I↑	CLIP-T↑
SigLIP-2	0.438	0.720	0.320
FLUX.1-dev	0.433	0.706	0.320
DINOv2	**0.632**	**0.788**	**0.328**

(a) Significance of fine-grained semantic injection

Experiment	DINO↑	CLIP-I↑	CLIP-T↑
Baseline	0.670	0.798	**0.322**
+Cross-Attn	0.667	0.803	0.318
+MM-Attn	**0.672**	**0.806**	0.321

(b) Impact of injecting subject details.

Results

- Ablations: Qualitative
  - Qualitatively validate the effectiveness of dual-path subject injection.
    - Semantic path improves subject identity consistency;
    - Content path further adds texture details.

Conclusion

Our Contributions

- Efficient feed-forward subject-driven generation method built on VAR:
  - Dual-path subject injection preserves subject structure and local details;
  - Subject segmentation isolates the target subject from complex backgrounds;
  - Subject-text CFG enhances control over subject features and text prompts.

Results

- The model achieves comparable or superior quality than baselines with substantially faster inference.

Thanks!

Paper

Code

Paper: https://arxiv.org/abs/2509.26127

Project Page: https://github.com/drx-code/EchoGen

中国科学技术大学
University of Science and Technology of China

Alibaba