Fine-Tuning Visual Autoregressive Models for Subject-Driven Generation

Jiwoo Chung, Sangeek Hyun, Hyunjun Kim, Eunseo Koh, MinKyu Lee, and Jae-Pil Heo

Sungkyunkwan University

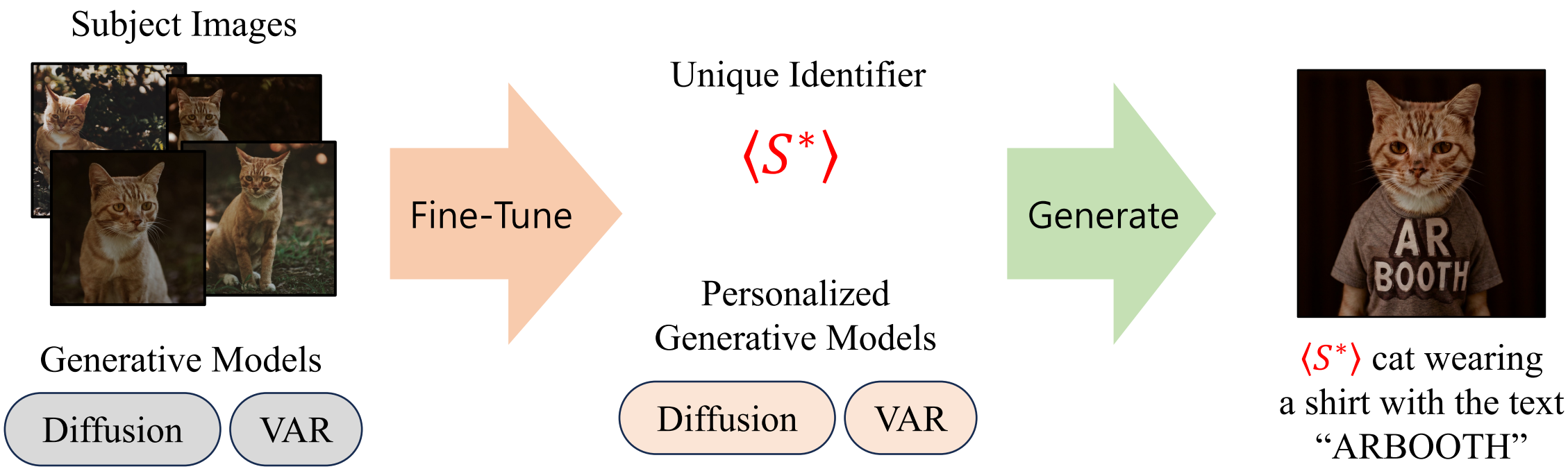
Project Page

Backgrounds (Subject-driven generation)

Input: a few images of a subject (*e.g., a dog*)

Goal: generate new images aligned with prompts

Key challenge: preserve identity while adapting to diverse contexts



Problem Definition

Problem

- Diffusion personalization → **high quality but slow**
- Naïve VAR fine-tuning → **fast, but identity underfitting & language drift**

Our Answer

- Efficient VAR framework with **Selective Layer Tuning, Scale-wise Weighted Tuning, and Prior Distillation**

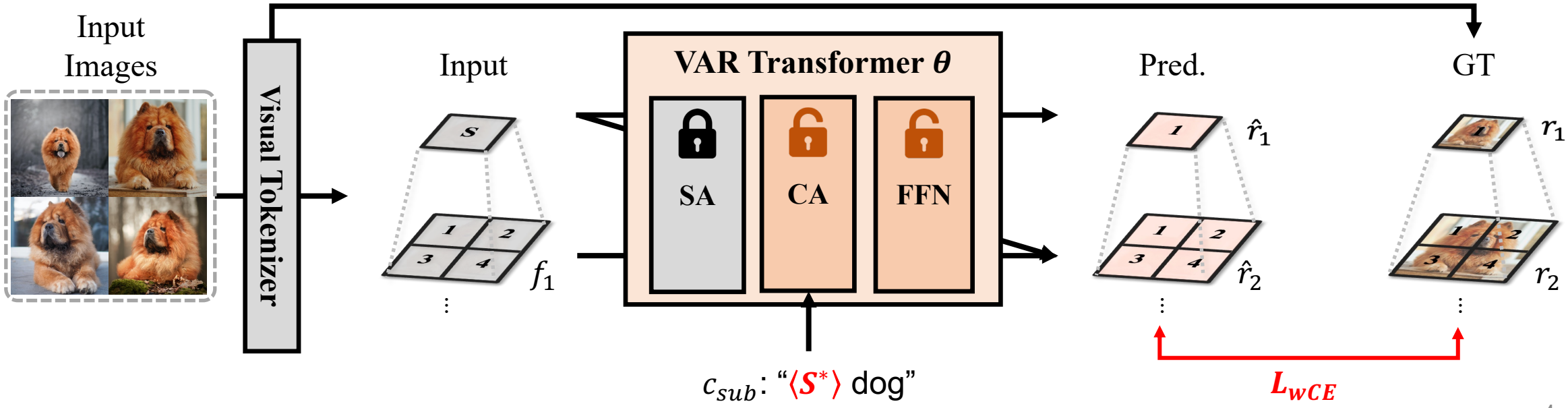
→ **fast (~0.5 s) and faithful generation**

Model Type	Inference Speed	Quality
Diffusion-based	Slow 	Nice 
Naïve VAR Fine-Tuning	Fast 	Bad 
Ours	**Fast** 	**Good** 

Our Framework (ARBooth)

Subject-driven Fine-Tuning

- Encode subject images into multi-scale token maps
- Fine-tune **only key layers (CA & FFN)** with subject token
- Use **Weighted CE Loss** to emphasize coarse scales



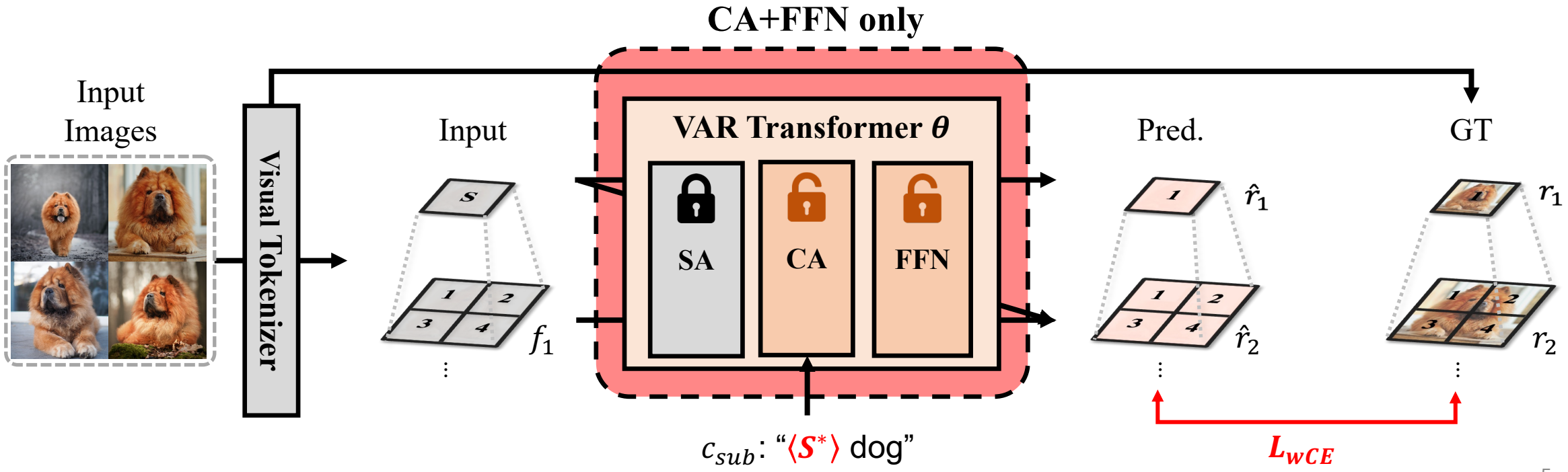
Selective Layer Tuning (1)

Idea

- Fine-tune only CA & FFN layers → efficient & effective personalization

Why?

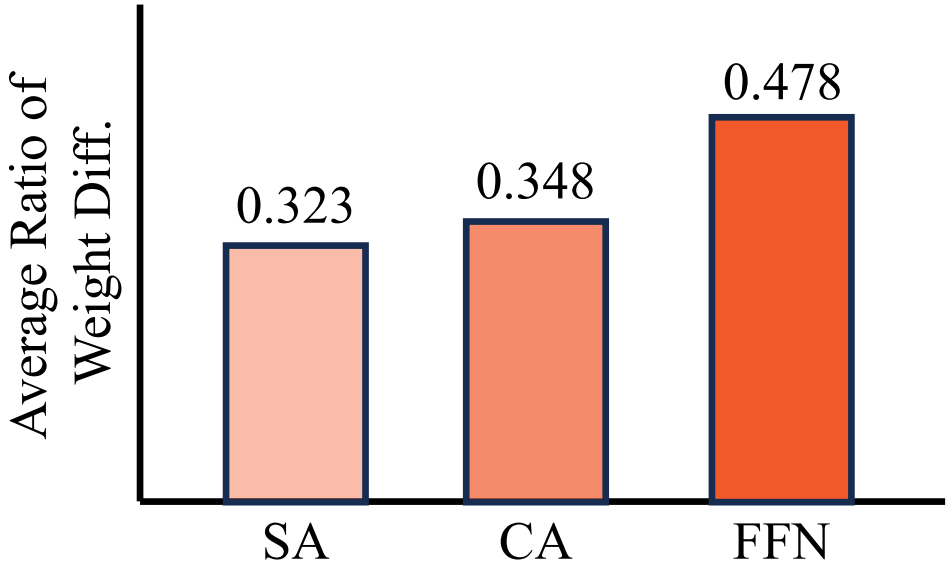
- Effectively captures subject identity with fewer parameters

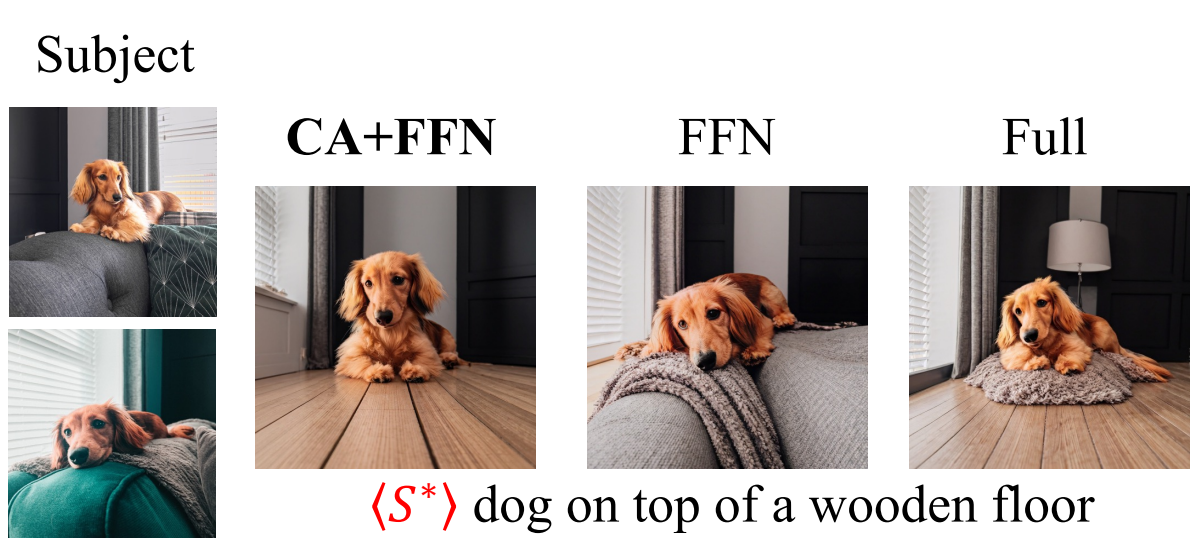


Selective Layer Tuning (2)

Observation

- When all layers are tuned, **CA & FFN change the most**
- Empirical results: tuning only CA+FFN → strong identity and prompt alignment





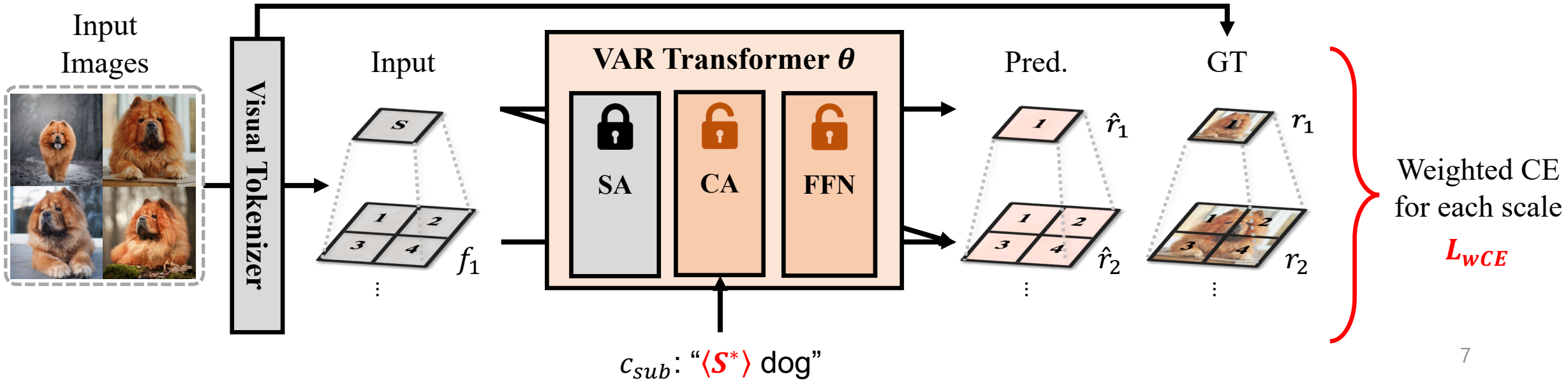
Scale-wise Weighted Tuning (1)

Idea

- Early (coarse) scales mainly determine subject identity
- Later (fine) scales refine only minor details

Our Approach

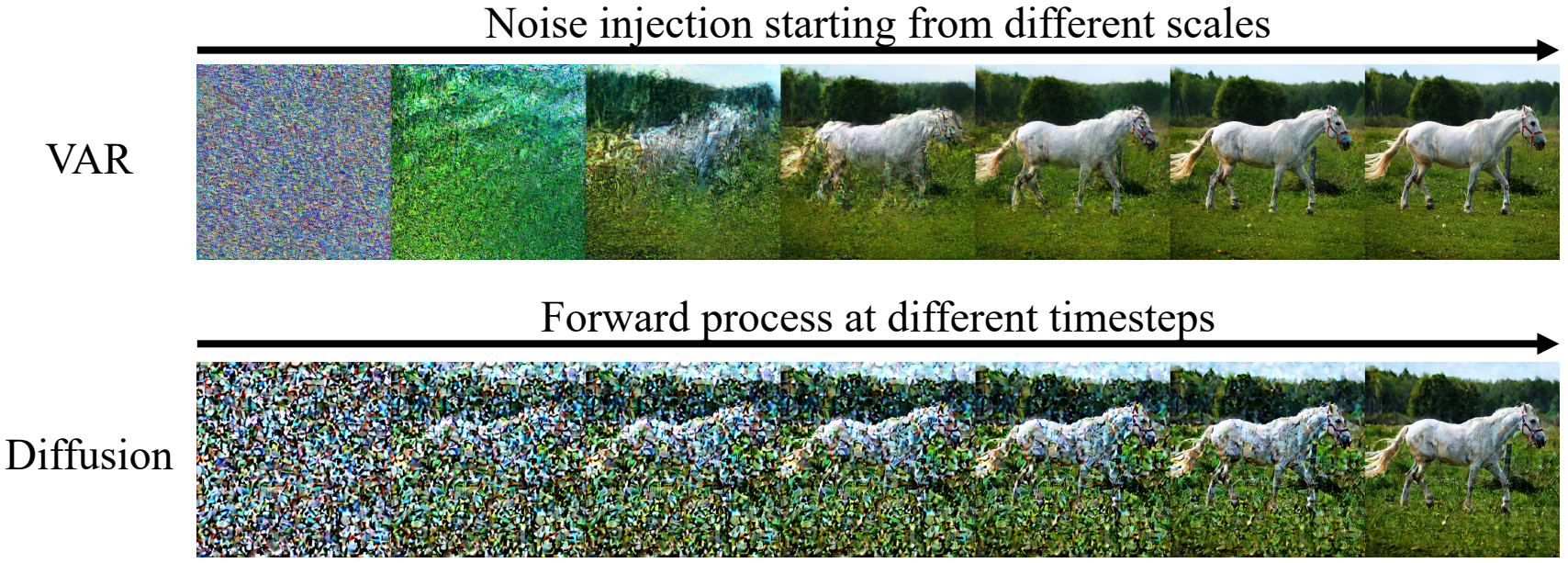
- Apply **a Weighted Cross-Entropy (L_{wCE})** loss
- Assign larger weights to coarse scales → better preservation of subject identity

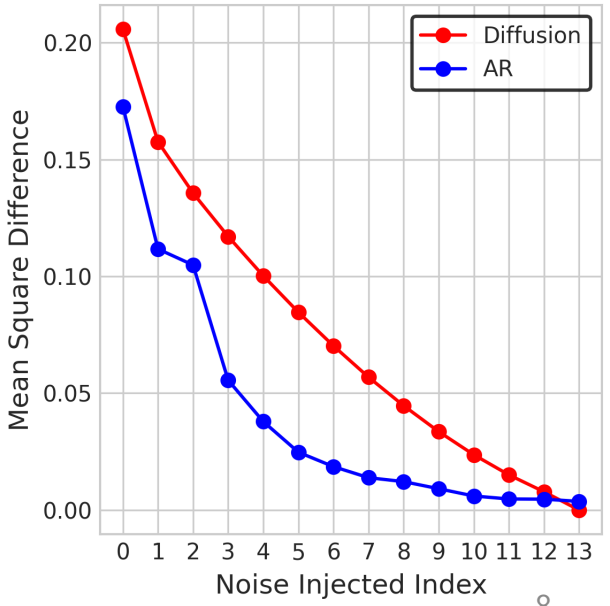


Scale-wise Weighted Tuning (2)

Observation

- Noise injection analysis: replacing coarse scales drastically changes the generated content
- Fine-scale replacement barely affects subject identity (similar to diffusion models)
- Emphasizing coarse scales through SWT → improves fidelity and stability





Scale-wise Weighted Tuning (3)

Observation

- **Without SWT** → identity drift
- **With SWT** → subject preserved

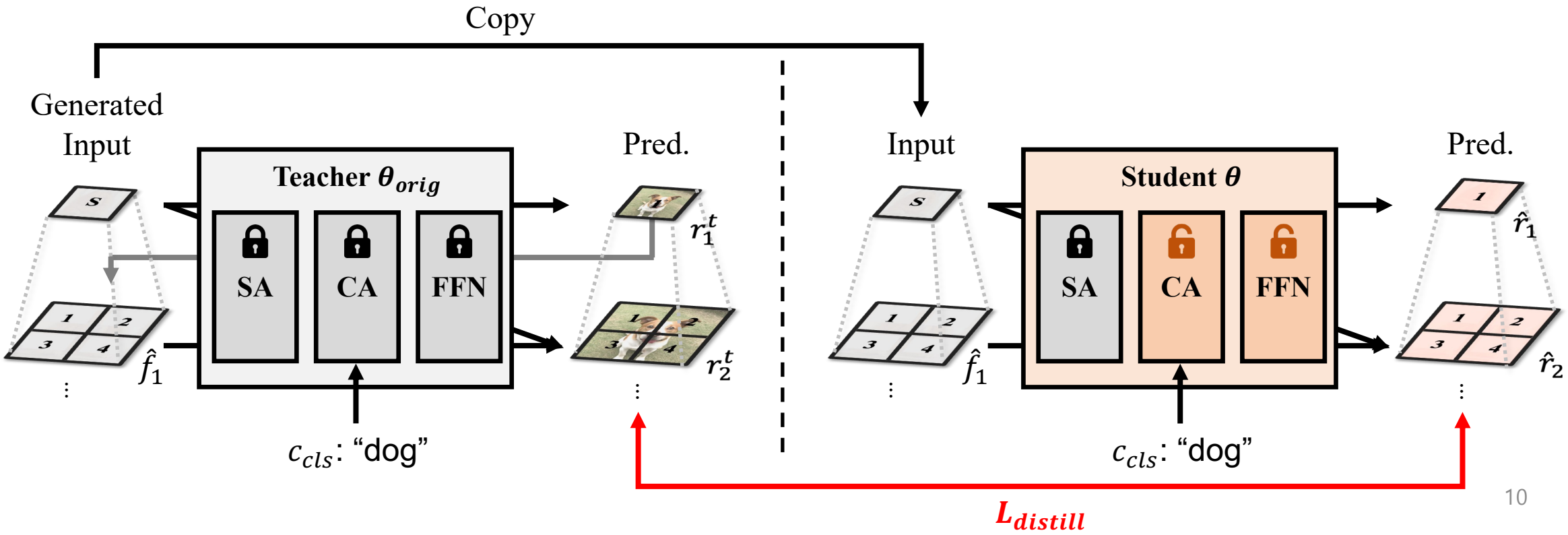
Prior Distillation (1)

Idea

- Naïve VAR fine-tuning → language drift & reduced diversity

Our Approach

- Distill class priors from pretrained model (Teacher → Student)
- KL-divergence loss → preserve knowledge & maintain diversity without extra data



Prior Distillation (2)

Observation

- **Without Prior Distillation** → model overfits, collapsing to the fine-tuned subject
- **With Prior Distillation** → retains pretrained knowledge, generating diverse and faithful subject instances

Subject

with prior distillation

$\langle S^* \rangle$ cat in the snow cat in the snow

without prior distillation

$\langle S^* \rangle$ cat in the snow cat in the snow

Qualitative Comparison (1)

Input images

DreamMatcher ELITE CD DreamBooth TI **Ours**

a $\langle S^* \rangle$ teddybear floating on top of water

a purple $\langle S^* \rangle$ teddybear

Qualitative Comparison (2)

Input images

DreamMatcher	ELITE	CD	DreamBooth	TI	**Ours**
					

a $\langle S^* \rangle$ cat wearing a santa hat






a cat $\langle S^* \rangle$ in a police outfit

Quantitative Comparison

- Dataset: ViCo [1] → 16 subjects × 31 prompts
- Evaluation: 8 images per pair → total 3,968 generations

Model	Subject Fidelity (I_{dino} ↑)	Text Fidelity (T_{clip} ↑)	Time ↓
TI	0.529	0.220	18 s
DreamBooth	0.640	0.815	18 s
CD	0.659	0.815	18 s
ELITE	0.584	0.783	11 s
DreamMatcher	0.682	0.823	32 s
Ours	**0.705**	**0.824**	**0.5 s**

[1] Ham, Cusuh, et al. "Personalized residuals for concept-driven text-to-image generation." CVPR 2024.

Contributions

Our Contributions

- We propose the *first VAR-based subject-driven generation framework (ARBooth)*
- Key components:
  - **Selective Layer Tuning** → fine-tune only CA & FFN layers for efficient personalization
  - **Scale-wise Weighted Tuning** → emphasize coarse scales to preserve subject identity
  - **Prior Distillation** → distill class priors from the pretrained model to prevent language drift

Results

- Achieves fast inference (~0.5 s) with high subject fidelity and text alignment

Thank you!

Subject

$\langle S^* \rangle$ dog seen from the top

$\langle S^* \rangle$ dog seen from the back

$\langle S^* \rangle$ dog wearing a top hat

$\langle S^* \rangle$ dog in an astronaut outfit

$\langle S^* \rangle$ dog diving into a pool

$\langle S^* \rangle$ dog inside a cave entrance

$\langle S^* \rangle$ dog sitting by a campfire under the stars

$\langle S^* \rangle$ dog surfing on a giant wave at sunset

$\langle S^* \rangle$ dog exploring an underwater coral reef

$\langle S^* \rangle$ dog riding a motorcycle along a coastal highway

$\langle S^* \rangle$ dog scaling a challenging rock wall at a canyon

$\langle S^* \rangle$ dog riding gracefully on a giant koi fish in a pond