Rethinking Training Dynamics in Scale-wise Autoregressive Generation

Supplementary Material

In this supplementary, Section A presents extended training curves and a thorough throughput–FID comparison across representative generative model families, offering a deeper perspective on the optimization behavior of SAR. In Section B, we analyze the generation process step by step, comparing SAR with FlexVAR under matched initial conditions. Section C showcases qualitative results generated by our method. Section D provides a detailed study of student-forcing behavior and illustrates how SAR mitigates drift and stabilizes scale-wise autoregressive training.

A. Training Curves and Model Comparison

A.1. Training Dynamics

Figure 1 provides a joint visualization of the training behavior of SAR and its performance relative to state-of-the-art generative models. We compare three variants trained on ImageNet 256×256:

- **FlexVAR** [1]: a FlexVAR-d16 model trained from scratch for 180 epochs.
- **SAR (from scratch)**: SAR trained from scratch for 180 epochs.
- **SAR (init FlexVAR)**: SAR initialize from a well-trained FlexVAR model at 170 epochs, then further train for 10 epochs.

We observe several important trends:

1. **Faster convergence.** SAR trained from scratch descends sharply in the early stage and consistently maintains a lower FID across training. This indicates that SAR provides a stable student-forcing training scheme and could effectively improve teacher-forcing training.
2. **Lower final error.** By Epoch 180, SAR achieves a noticeably lower FID than FlexVAR. The smoother convergence curve suggests improved stability in token prediction during rollouts.
3. **Strong initialization benefits.** Initializing SAR from FlexVAR drastically accelerates early-stage optimization. Within only a few epochs, SAR quickly surpasses the best performance of a fully-trained FlexVAR model, converging to the lowest FID among all variants.

Overall, the training analysis confirms that SAR not only enhances sample quality but also makes student-forcing training in scale-wise AR effective and stable.

A.2. Model Comparison

Figure 2 visualizes the throughput–FID trade-off across popular generative model families.

Key observations include:

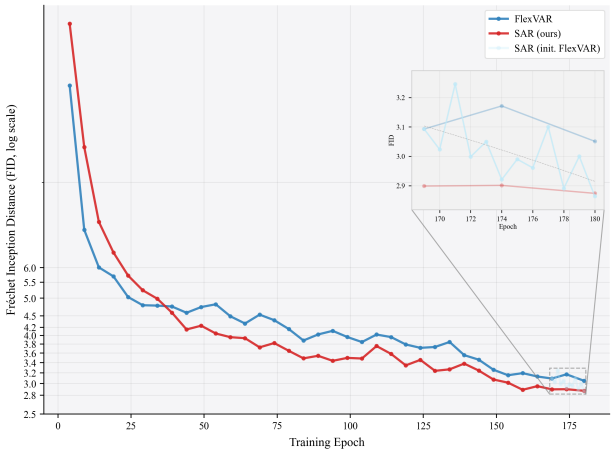

Figure 1. **Training dynamics.** Training curves for FlexVAR, SAR trained from scratch, and SAR initialized from a pretrained FlexVAR checkpoint.

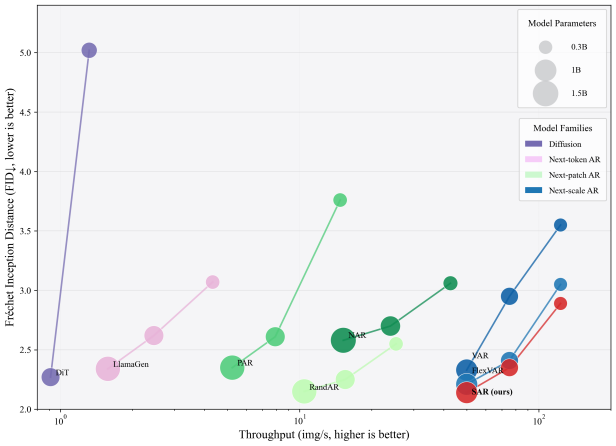

Figure 2. **Throughput–FID trade-off.** Comparison of throughput, parameter count, and FID across representative generative model families, including Diffusion, Next-token AR, and Next-scale AR.

- **Diffusion models** (purple) generally achieve good FID but suffer from low throughput due to long sampling trajectories.
- **Next-token AR models** (greens) improve throughput by using autoregression in latent space, but token-level causality limits their global coherence and negatively impacts FID.
- **Next-scale AR models** (blues), including VAR and Flex-

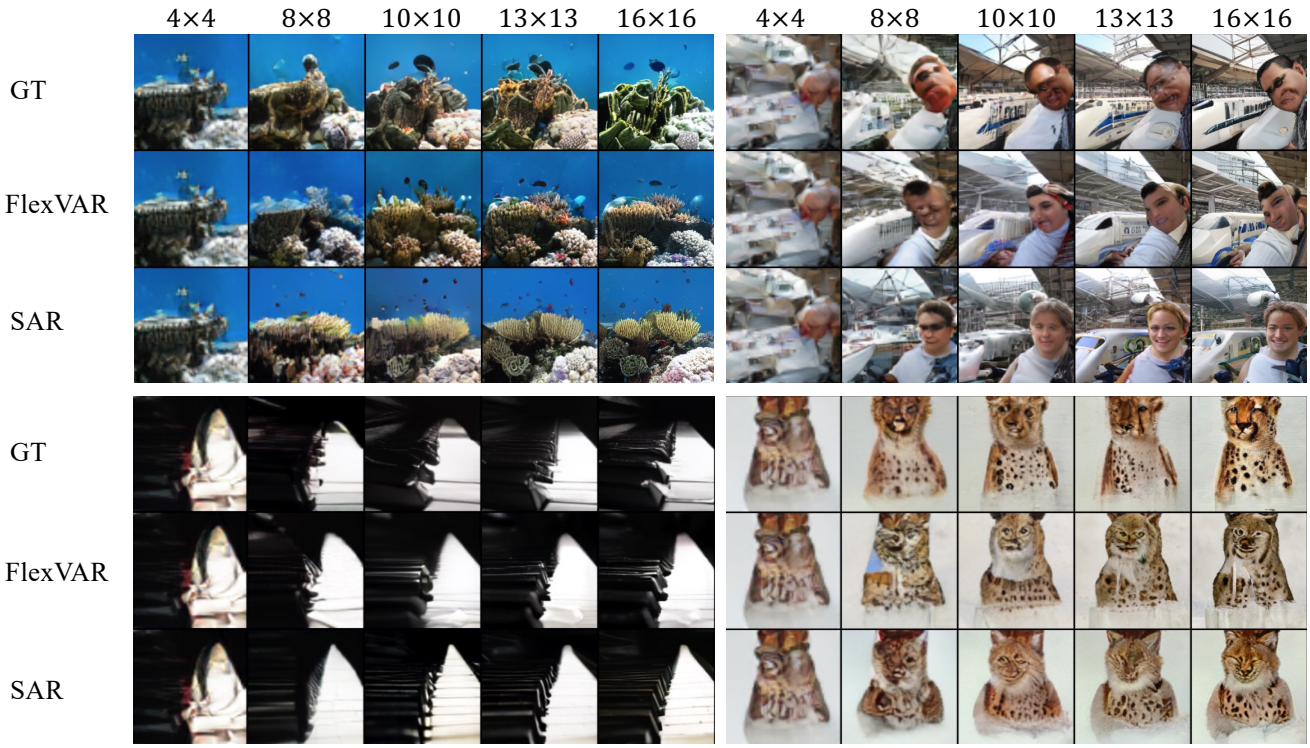

Figure 3. Comparison of the generation process of SAR and FlexVAR. Both models start from the same 4×4 latent and follow identical scale schedules. SAR demonstrates smoother transitions and stronger error correction across scales.

VAR, provide an efficient coarse-to-fine sampler that greatly boosts throughput.

- **SAR** (red) attains the *best overall trade-off*: the **highest throughput among autoregressive models** and further improve next-scale prediction AR model with the **lowest FID across all AR baselines**.

The introduction of SAR retains the structural advantages of VAR while improving generation quality.

B. Visualization of the Generation Process

To better understand the qualitative differences between SAR and FlexVAR, we visualize the full generation trajectory of both models under identical initial conditions. We begin by selecting images from the ImageNet validation set, encoding them using the VAE to obtain the scale-wise latent representations. At inference time, both models are initialized with the same 4×4 latent, and perform coarse-to-fine autoregressive generation following the scale schedule:

$$4 \to 8 \to 10 \to 13 \to 16.$$

Figure 3 presents the intermediate outputs at each scale for both SAR and FlexVAR.

Across all examples, two consistent patterns emerge:

- **Smooth and stable refinement.** SAR exhibits smoother transitions between scales, with fewer abrupt changes in structure or texture.
- **Error correction during refinement.** When early-scale predictions contain distortions or misaligned structures, SAR reliably corrects these errors at subsequent scales. In contrast, FlexVAR often amplifies early mistakes, leading to degraded high-resolution details.

These visualizations highlight the student-forcing–aligned training in SAR produces more stable hierarchical latents, enabling more coherent and robust refinement during generation.

C. Qualitative Results

We show the qualitative results of SAR in Figure 4.

D. Visualization of Student-Forcing Inputs During Training

To better understand the behavior of student-forcing (SF) within our SAR training framework, we visualize the intermediate latents produced during two consecutive SF rollout steps. This provides an intuitive, fine-grained view of how self-generated latents evolve, how errors propagate across

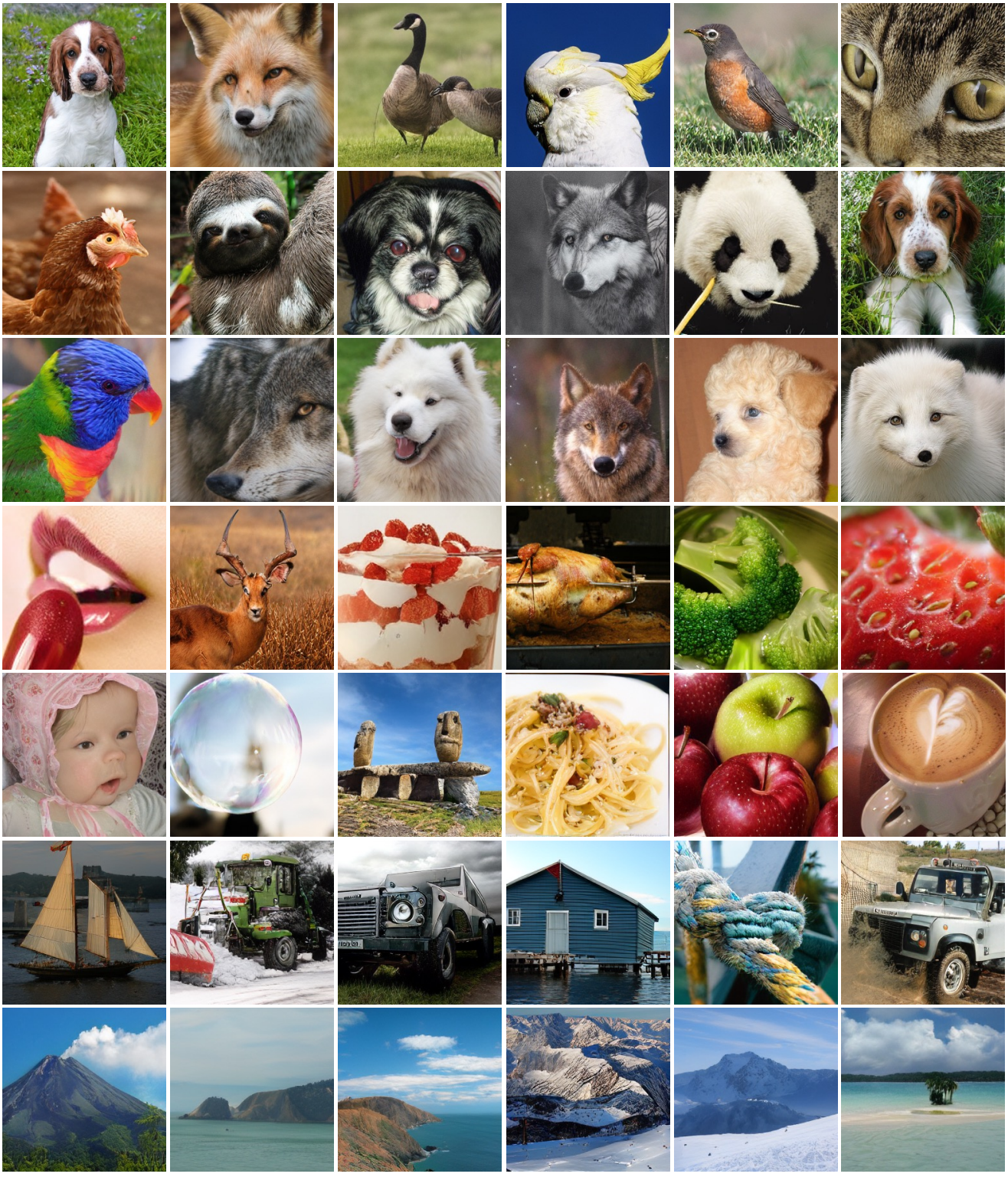

Figure 4. Qualitive results of SAR.

scales, and how they differ from the teacher-forcing (TF) pathway.

Our visualization uses the same notation and formulation introduced in the main paper. Recall that naïve student-forcing predicts the next-scale latent from its own previously generated latents:

$$\tilde{f}_i^{(S)} = g_\theta(\tilde{f}_{1:i-1}^{(S)}), \qquad \mathcal{L}_{\mathrm{SF}} = \sum_{i=1}^{N} \ell(\tilde{f}_i^{(S)}, f_i), \quad (1)$$

which exposes the model to inference-like conditions but suffers from drift, error amplification, and destabilized supervision (Sec. 4.1).

In contrast, SAR replaces naïve SF with Stagger-Scale Rollout (SSR), which produces both teacher-forced predictions

$$\hat{f}_i^{(T)} = g_\theta(f_{1:i-1}), \quad (2)$$

and a one-step, structure-preserving student-forced trajectory:

$$\tilde{f}_i^{(S)} = \mathrm{Upsample}(\hat{f}_i^{(T)}), \quad (3)$$

$$\hat{f}_i^{(S)} = g_\theta(\tilde{f}_{1:i-1}^{(S)}). \quad (4)$$

D.1. Visualization Setup

Figure 5 shows the intermediate latents obtained when performing two consecutive student forcing rollout steps, visualized across scale transitions: $10 \to 13, 13 \to 16$.

Each row in the figure corresponds to a specific component in the SAR rollout pipeline:

- **Row 1: Ground-truth latents (f_{10}, f_{13}, f_{16}).** These represent the teacher-forcing targets at each scale.
- **Row 2: TF model inputs ($\mathrm{up}(f_8)$, $\mathrm{up}(f_{10})$, $\mathrm{up}(f_{13})$).** Ground-truth latents are upsampled to match the next-scale resolution.
- **Row 3: SF inputs ($\tilde{f}_8^{(S)}$, $\tilde{f}_{10}^{(S)}$, $\tilde{f}_{13}^{(S)}$).** These are obtained by shifting TF predictions to the next scale at scale 8:

$$\tilde{f}_i^{(S)} = \mathrm{Upsample}(\hat{f}_i^{(T)}).$$

  And two consecutive SF steps at scales 10 and 13.
- **Row 4: SF predictions ($\hat{f}_{10}^{(S)}$, $\hat{f}_{13}^{(S)}$, $\hat{f}_{16}^{(S)}$).** These are produced by feeding the shifted predictions into the model:

$$\hat{f}_i^{(S)} = g_\theta(\tilde{f}_{1:i-1}^{(S)}).$$

- **Row 5: Difference maps (Δ_i).** To quantify the student-forcing deviation at each scale, we compute:

$$\Delta_i = \left| f_i - \tilde{f}_i^{(S)} \right|.$$

Rows 3–5 make the effect of student-forcing explicit: how student-generated latents enter the next-scale prediction, how the model responds, and where errors emerge.

D.2. Observations

Three consistent observations emerge from the visualization:

1. **Error amplification across scales.** Deviations introduced at coarse scales (e.g., $10 \to 13$) propagate and typically enlarge at finer scales ($13 \to 16$), demonstrating why naïve SF destabilizes training.
2. **Structured corrections under SSR.** Because SSR starts from teacher-forced predictions, the SF inputs remain structurally meaningful. As a result, the model corrects local distortions rather than diverging, which stabilizes hierarchical learning.
3. **Alignment between trajectories.** The CSFL drives $\hat{f}_i^{(S)}$ to match the stable TF predictions $\hat{f}_i^{(T)}$, yielding visibly smaller difference maps compared to naïve SF. This explains why SAR can leverage student-forcing without drifting off-manifold.

Together, these visualizations clarify why naïve SF fails and why SAR succeeds: SSR ensures that self-generated inputs remain structured, while CSFL ensures that student-forced latents follow the correct predictive trajectory. This resolves the supervision inconsistency and the train–test gap in scale-wise autoregression.

References

[1] Siyu Jiao, Gengwei Zhang, Yinlong Qian, Jiancheng Huang, Yao Zhao, Humphrey Shi, Lin Ma, Yunchao Wei, and Zequn Jie. Flexvar: Flexible visual autoregressive modeling without residual prediction. *arXiv preprint arXiv:2502.20313*, 2025. 1

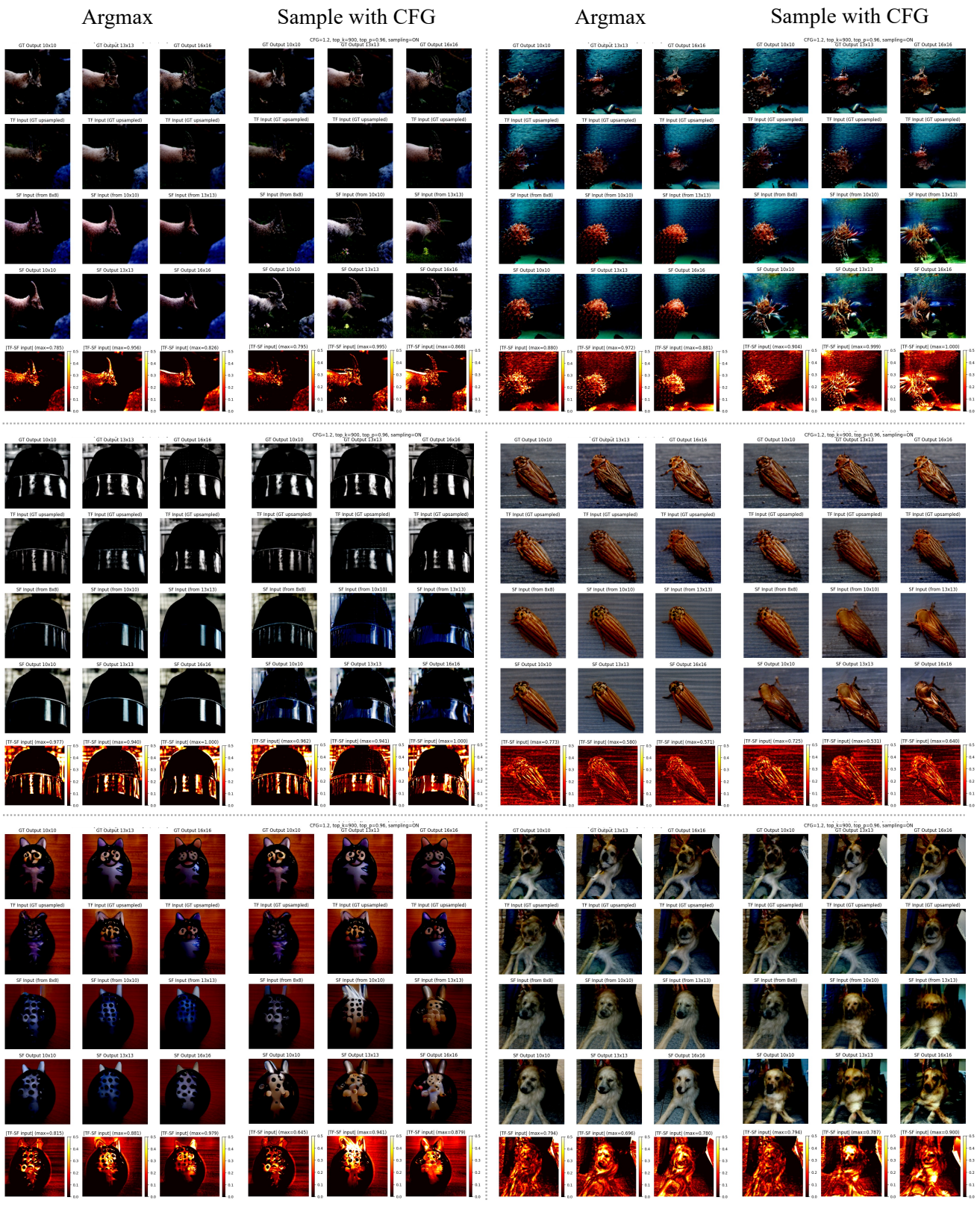

Figure 5. **Visualization of student-forcing inputs during training.** We show ground-truth latents, teacher-forcing (TF) inputs, student-forcing (SF) inputs, SF predictions, and the corresponding difference maps across two consecutive rollout steps ($10 \rightarrow 13$ and $13 \rightarrow 16$).