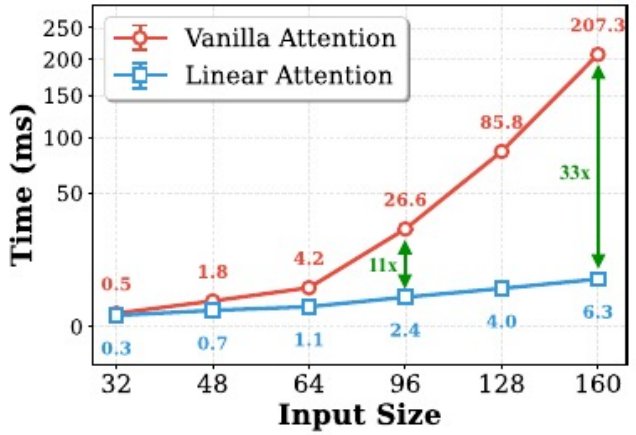
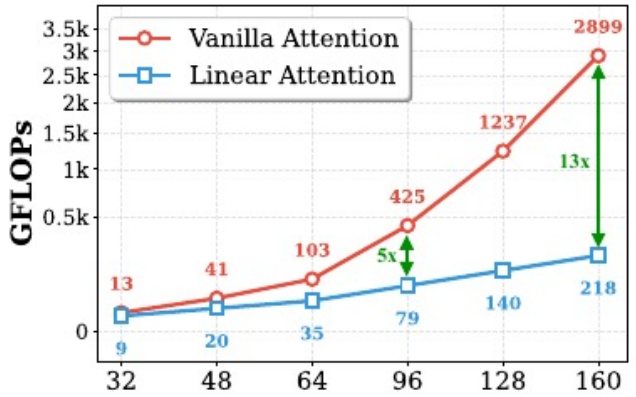
LinearSR: Unlocking Linear Attention for Stable and Efficient Image Super-Resolution

First robust recipe for making linear attention stable, efficient, and high-fidelity in diffusion SR.





Why linear attention was still missing in SR

Quadratic attention is the main scaling bottleneck in photorealistic SR.

O(N^2) ->O(N)

1) Guidance

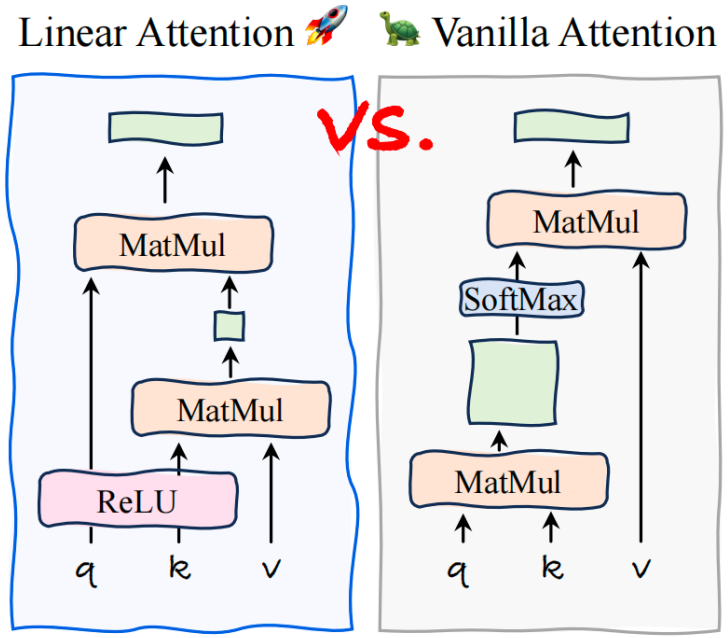
Verbose external descriptions do not match the intrinsic semantics already contained in the LR image.

2) Stability

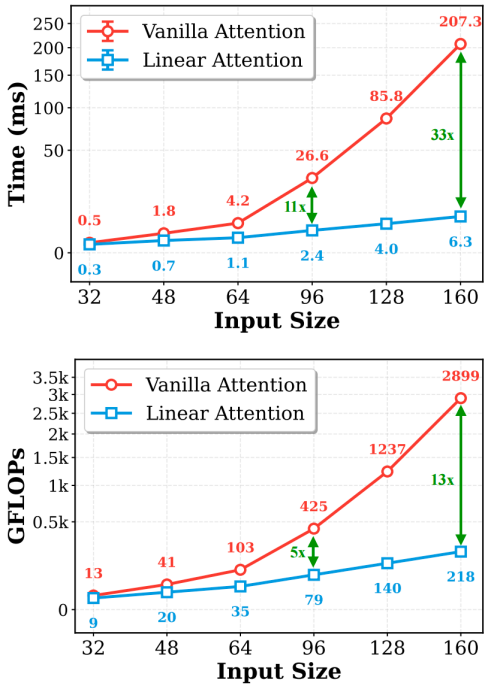
Naively fine-tuning a converged linear-attention SR model can diverge to NaN and collapse training.

3) Perception vs. distortion

Improving perceptual realism often hurts fidelity, so different denoising stages should not share one generic strategy.

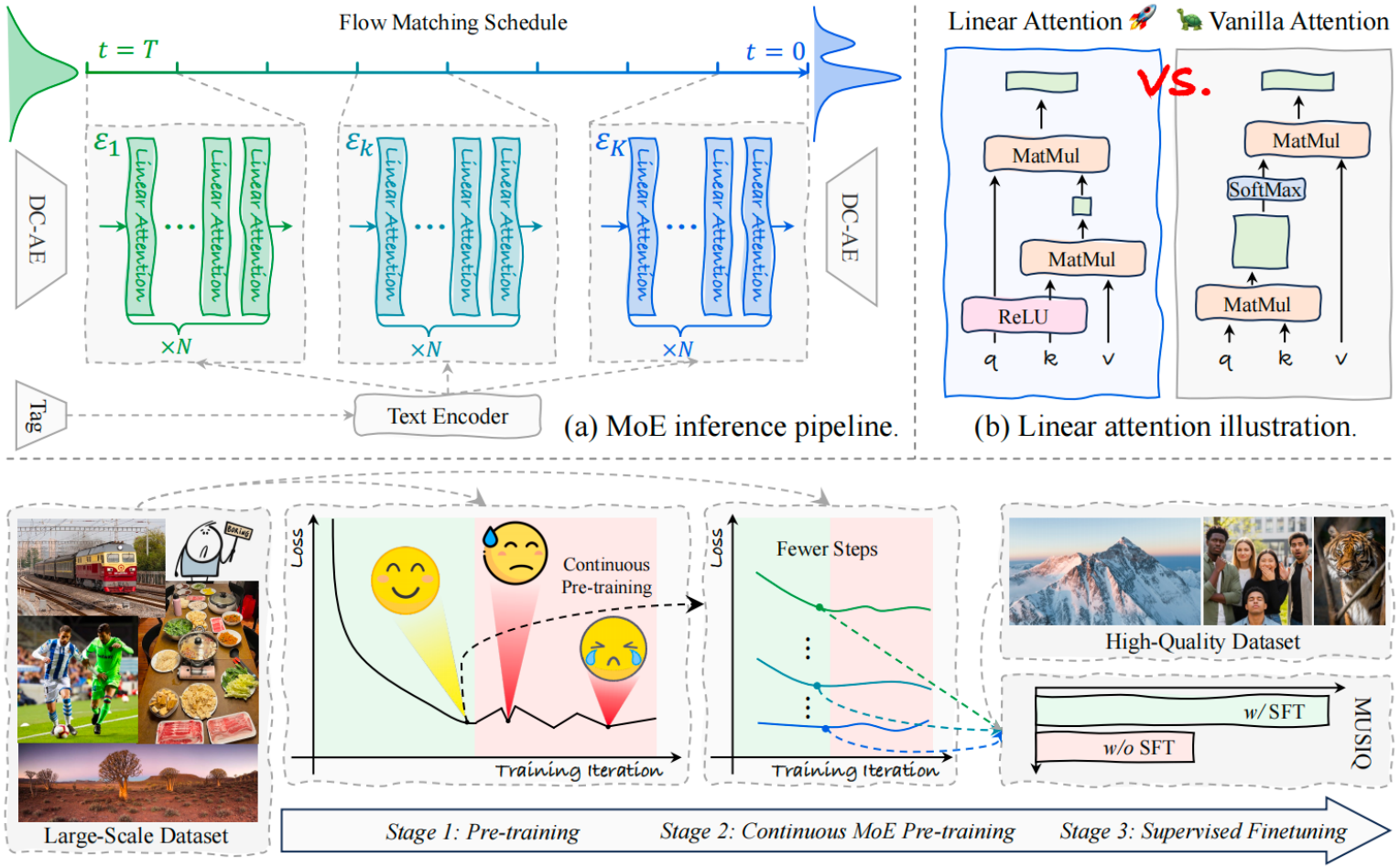


(b) Linear attention illustration.



Method

LinearSR framework



TAG guidance

Use concise object tags instead of long captions. The paper calls this “precision over volume.”

ESGF

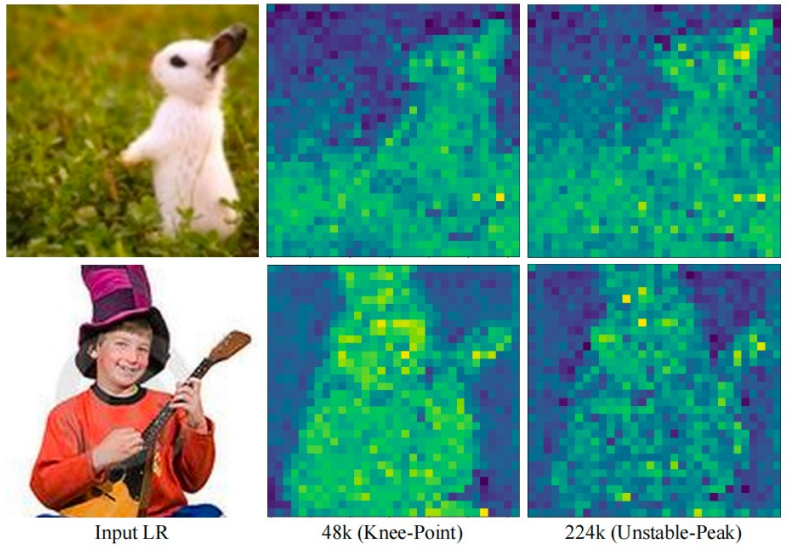
Select the knee-point checkpoint before fine-tuning. This is the stability anchor of the framework.

SNR-based 4-expert MoE

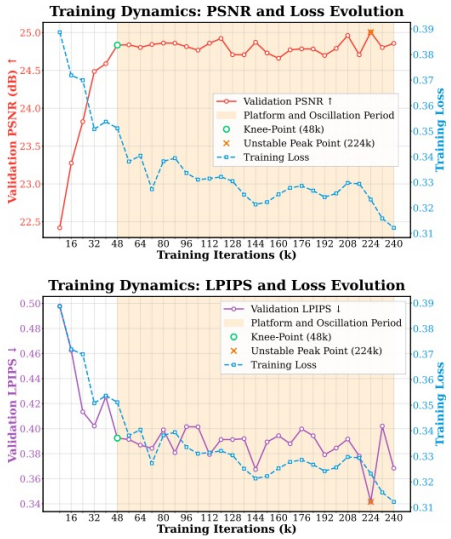
Different experts specialize in different denoising regimes: structure, refinement, texture, and polishing.

Core insight

Why ESGF and SNR-MoE matter



(a) Linear Attention Features: Knee-Point vs. Unstable Peak



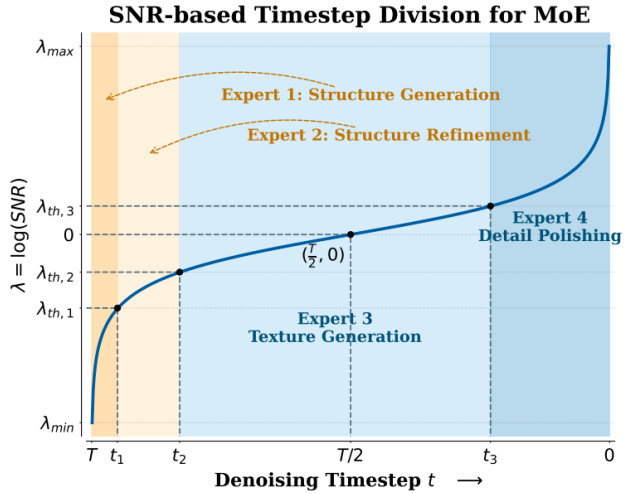
(b) Training Dynamics

After the knee point, loss keeps decreasing but validation quality becomes unreliable.

So loss alone is a deceptive model-selection signal for linear-attention SR fine-tuning.

ESGF in one line

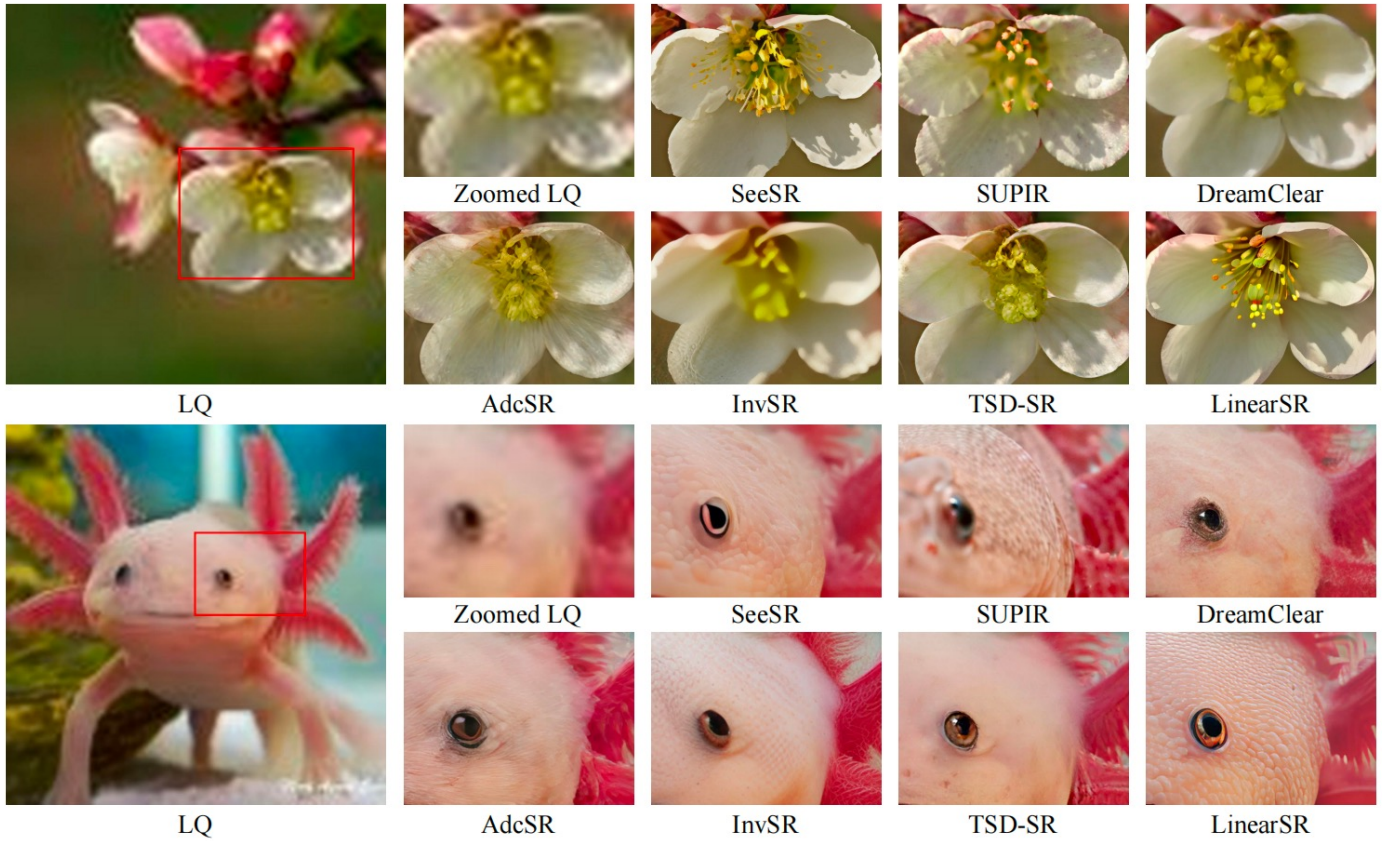
Start stage-2 fine-tuning from the 48k knee-point checkpoint, not from the 224k unstable peak. In the paper, the naive route collapses after about 2k steps.



SNR-based gating gives one active expert per timestep, so specialization does not add inference overhead.

Experiments

Main results



0.036 s

Fastest 1-NFE forward time for 1024×1024 output in the paper's comparison.

0.830 s

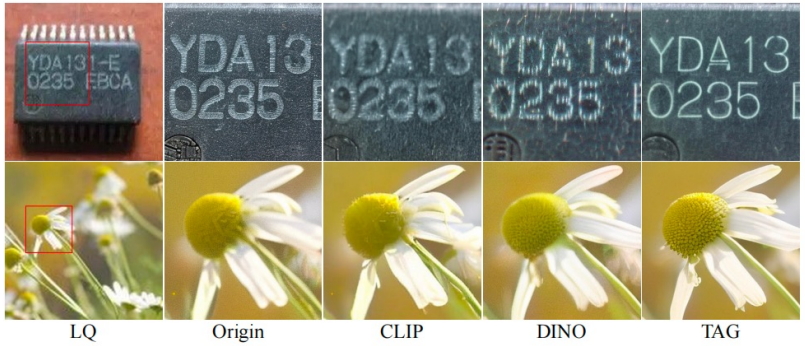
Overall multi-step inference time: competitive end-to-end speed, even without distillation.

RealLQ250 clean sweep

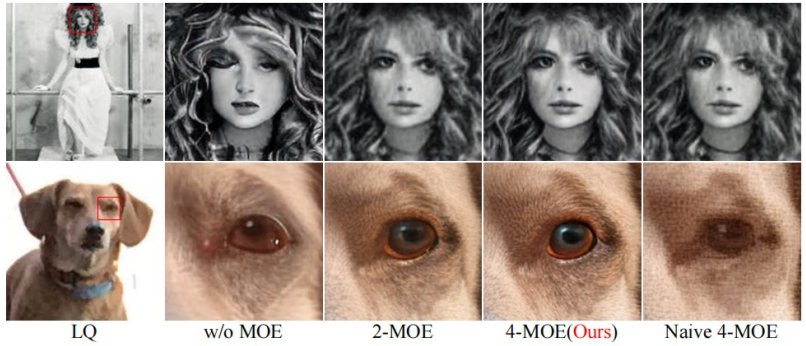
MANIQA	0.515
MUSIQ	71.914
CLIPIQA	0.720

The paper is strongest on no-reference perceptual metrics. Full-reference fidelity remains competitive rather than dominant.

Ablations and takeaways



(a) Ablation on guidance methods



(b) Ablation on MoE configurations

Guidance

TAG beats verbose captions, and also outperforms CLIP and DINO features.

Training

Without ESGF, naive fine-tuning collapses. ESGF is a prerequisite, not a small tuning trick.

Experts

The SNR-based 4-expert MoE beats simpler or uniformly partitioned expert layouts.

Takeaway: LinearSR turns linear attention from a theoretical acceleration idea into a practical SR backbone with stable training, strong perceptual quality, and genuine architectural efficiency.

Thank you!