Qualcomm

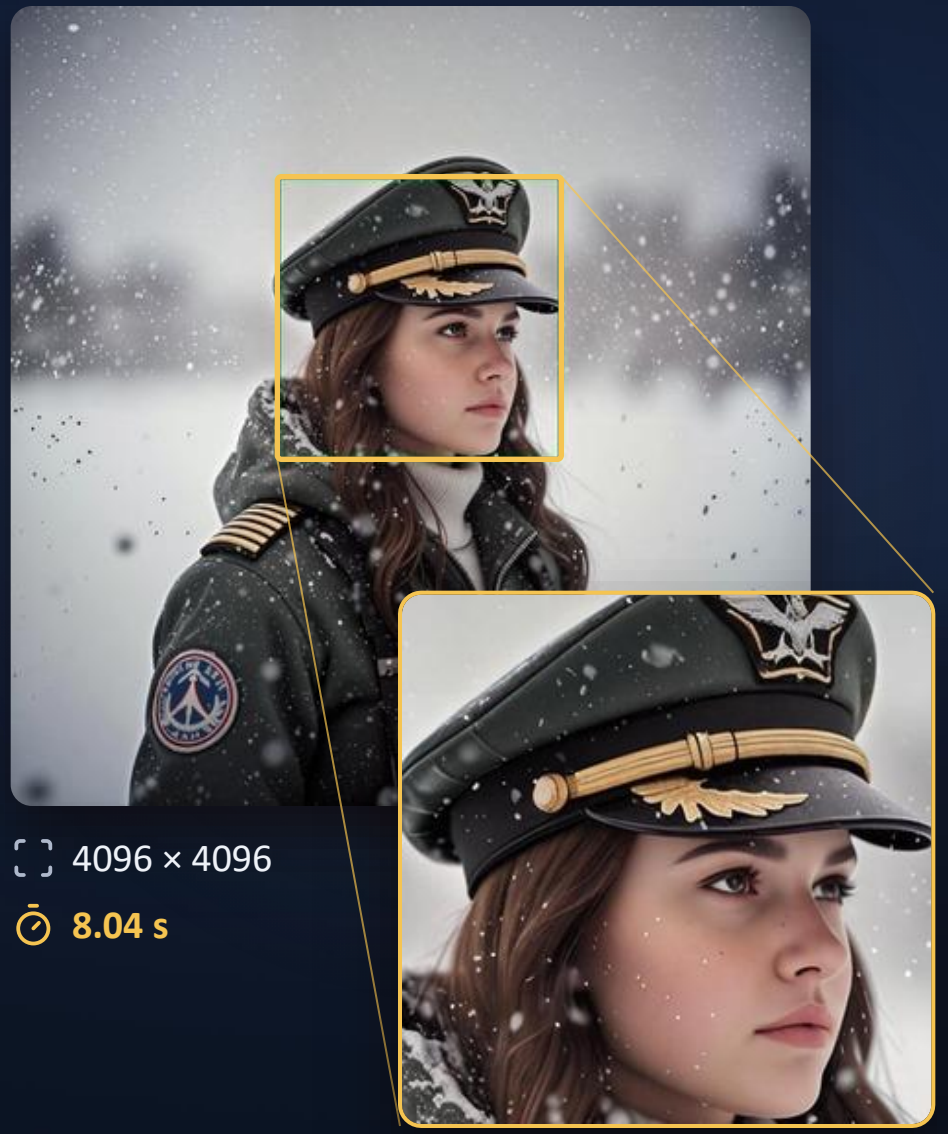
NeurIPS 2026

UltraFlash

Accelerating megapixel visual synthesis

Phuc Lai*, Anh Nguyen*, Phong Nguyen, Anh Tran

Qualcomm AI Research



Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

*Equal contribution

Three routes to megapixel images



Native 4K models

SANA, HunyuanImage

high-res data and training

Costly computation that scales quartically w.r.t. the side length

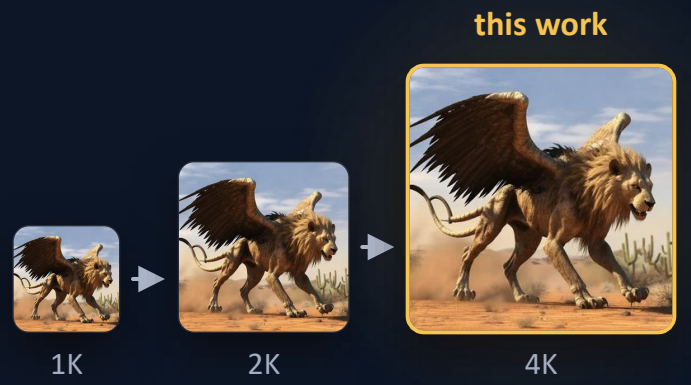


Direct extrapolation

HiFlow, DyPE for FLUX

no retraining

repeated objects, distorted layout



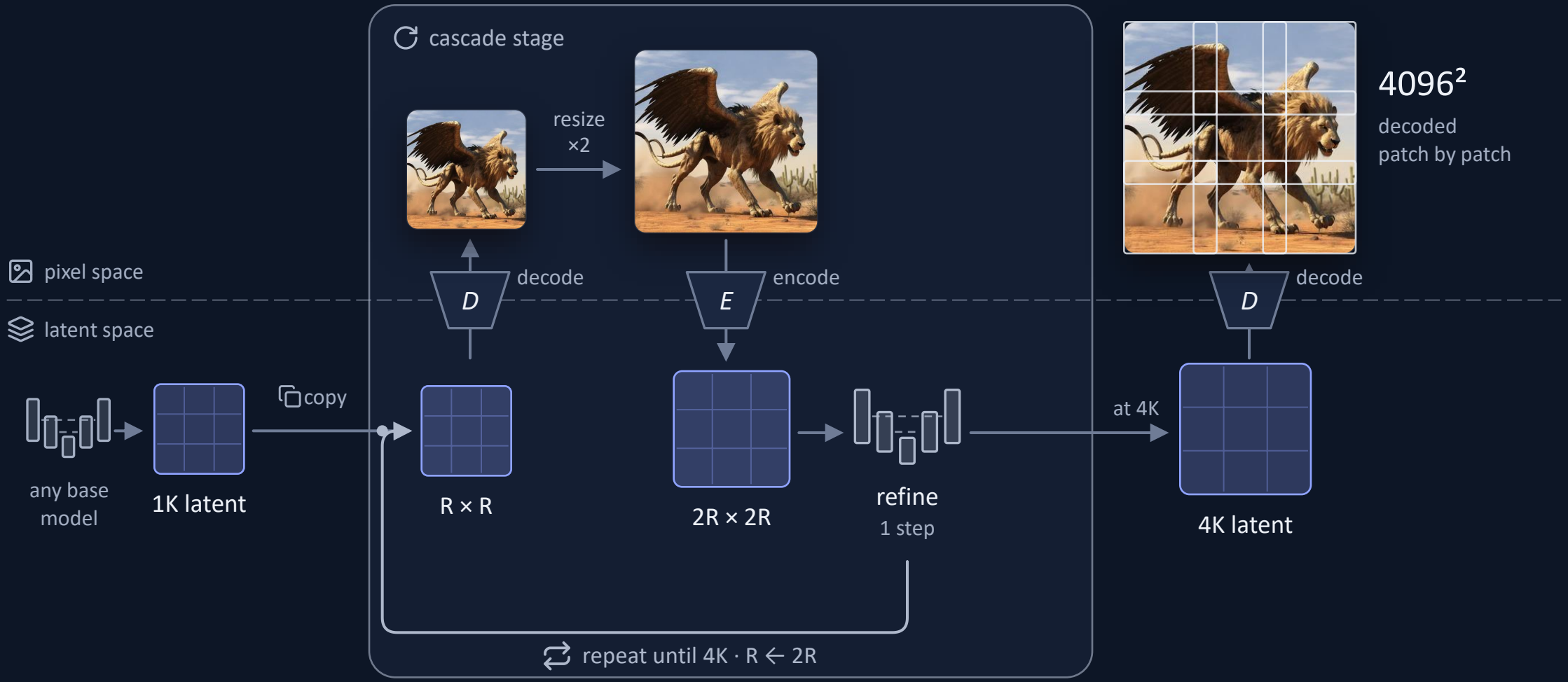
Cascade refinement

DemoFusion, HiWave, PixelRush

reuses pretrained models

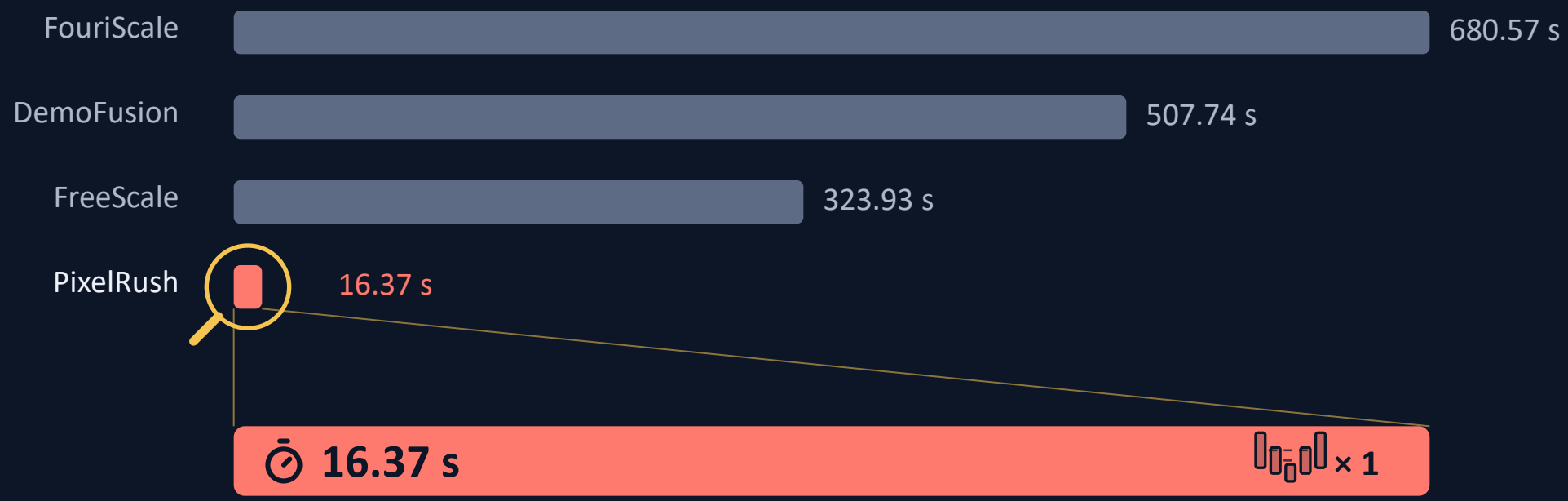
UltraFlash speeds up this route

Cascades reach 4K by refining a 1K image in stages



4K generation speed

seconds per 4096^2 image, one A100



PixelRush gives the best speed–quality trade-off among cascades.

Refinement takes one step, yet 4K takes 16.37 s.

So where does the time go?

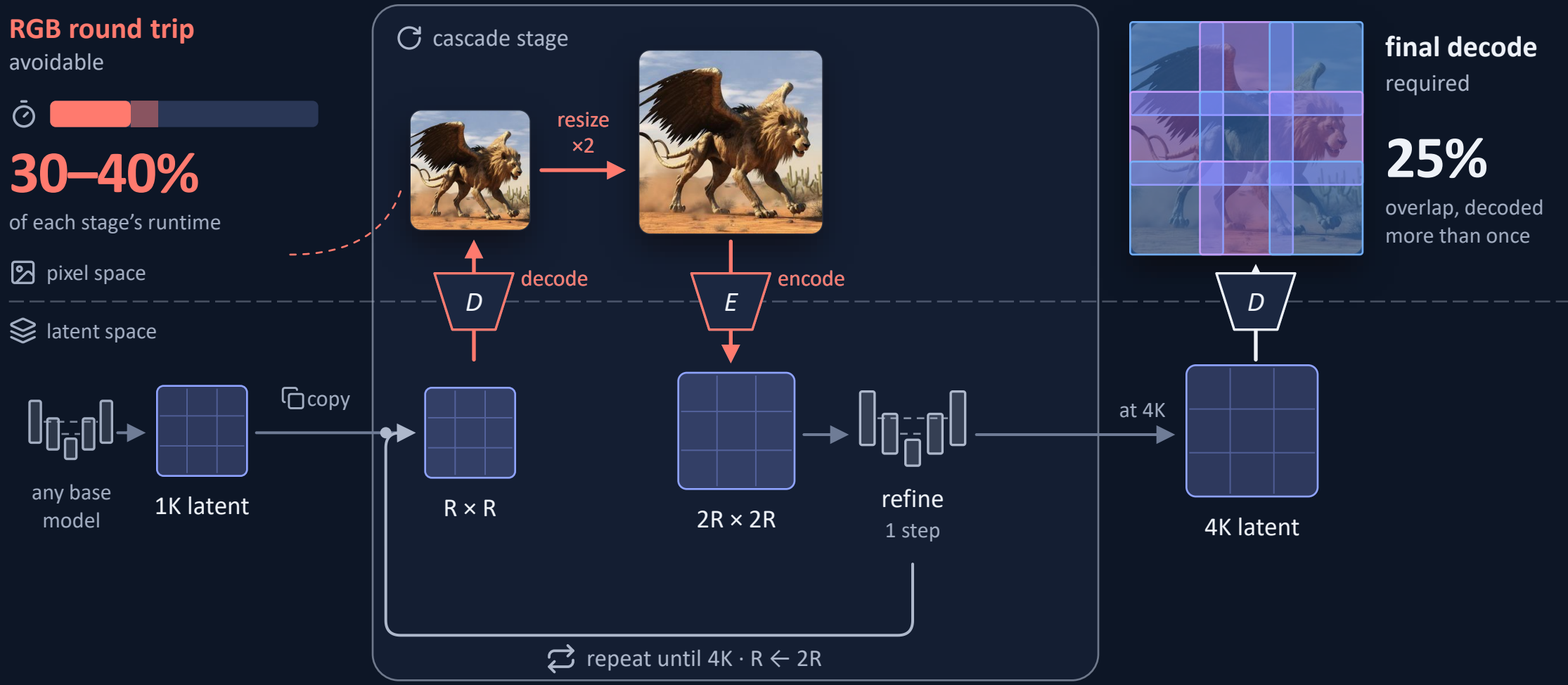
Two transitions: one avoidable, one required

RGB round trip
avoidable

30–40%
of each stage's runtime

pixel space

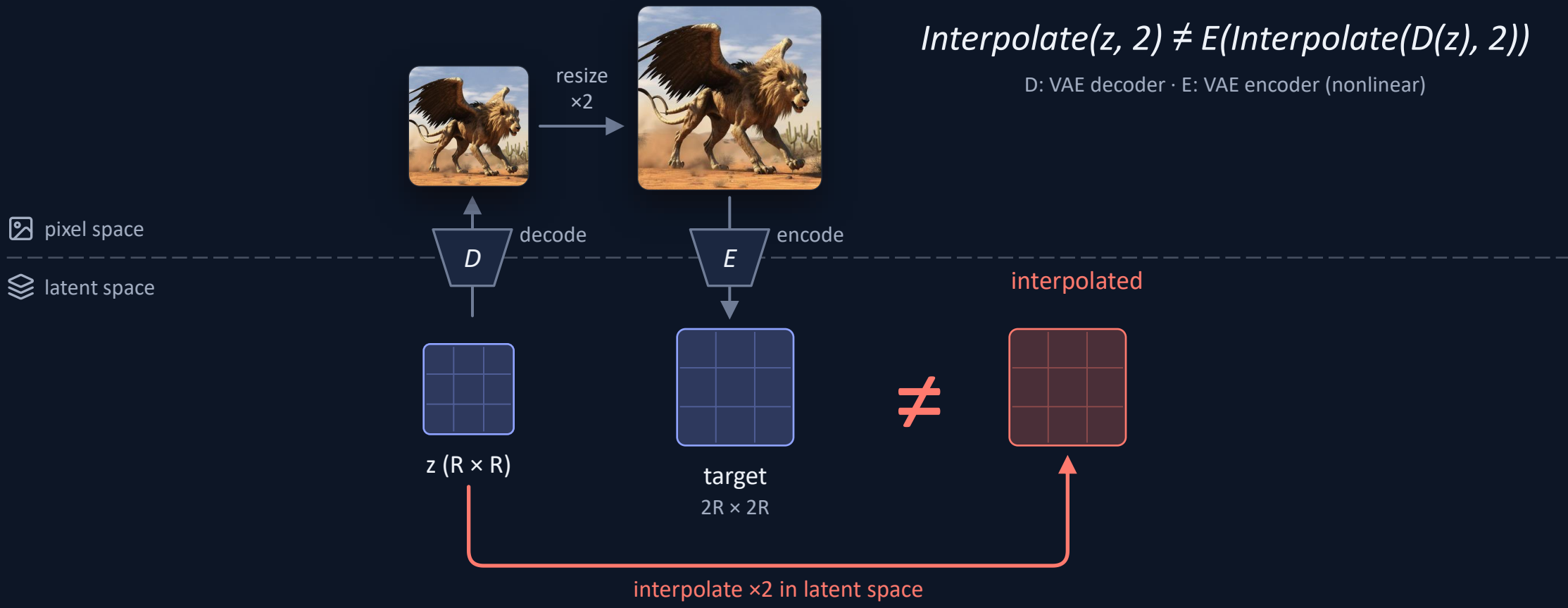
latent space



final decode
required

25%
overlap, decoded
more than once

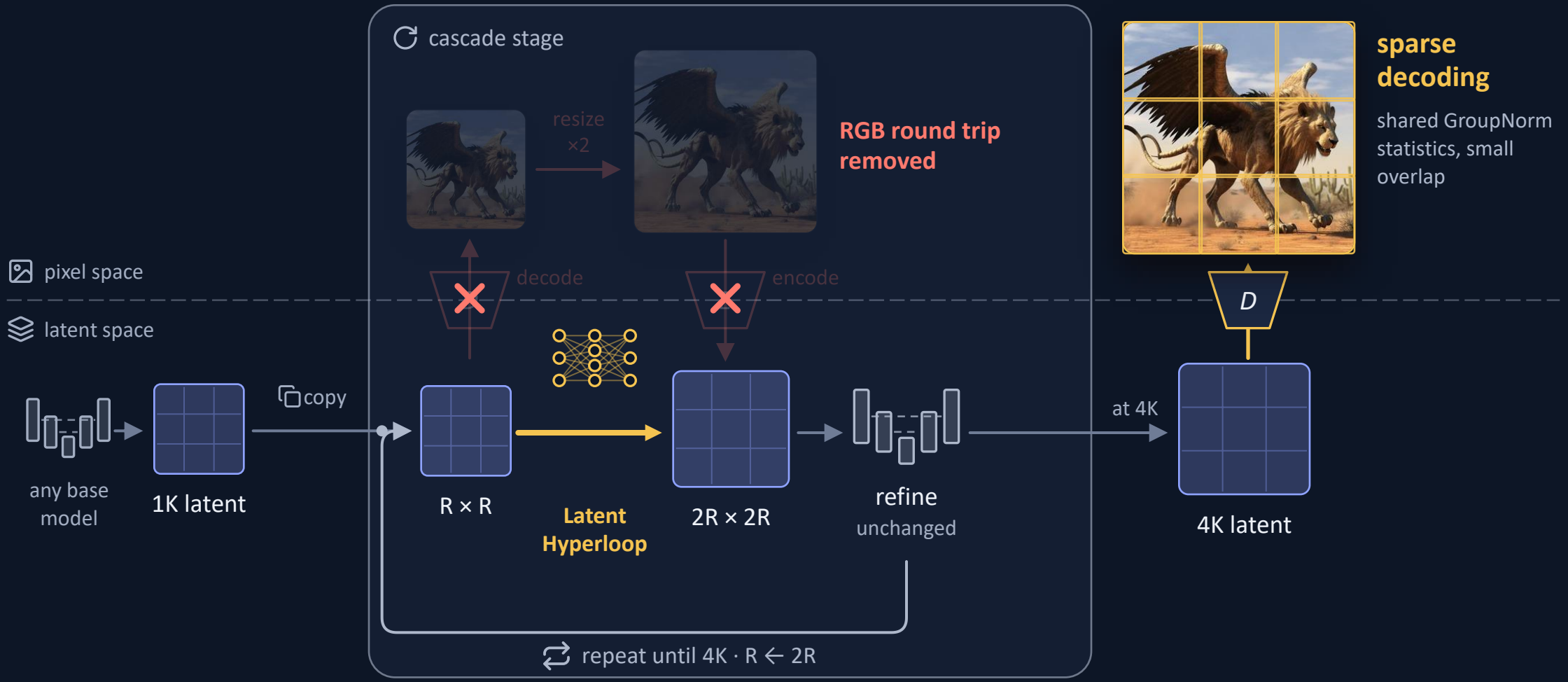
Latent interpolation cannot replace the round trip



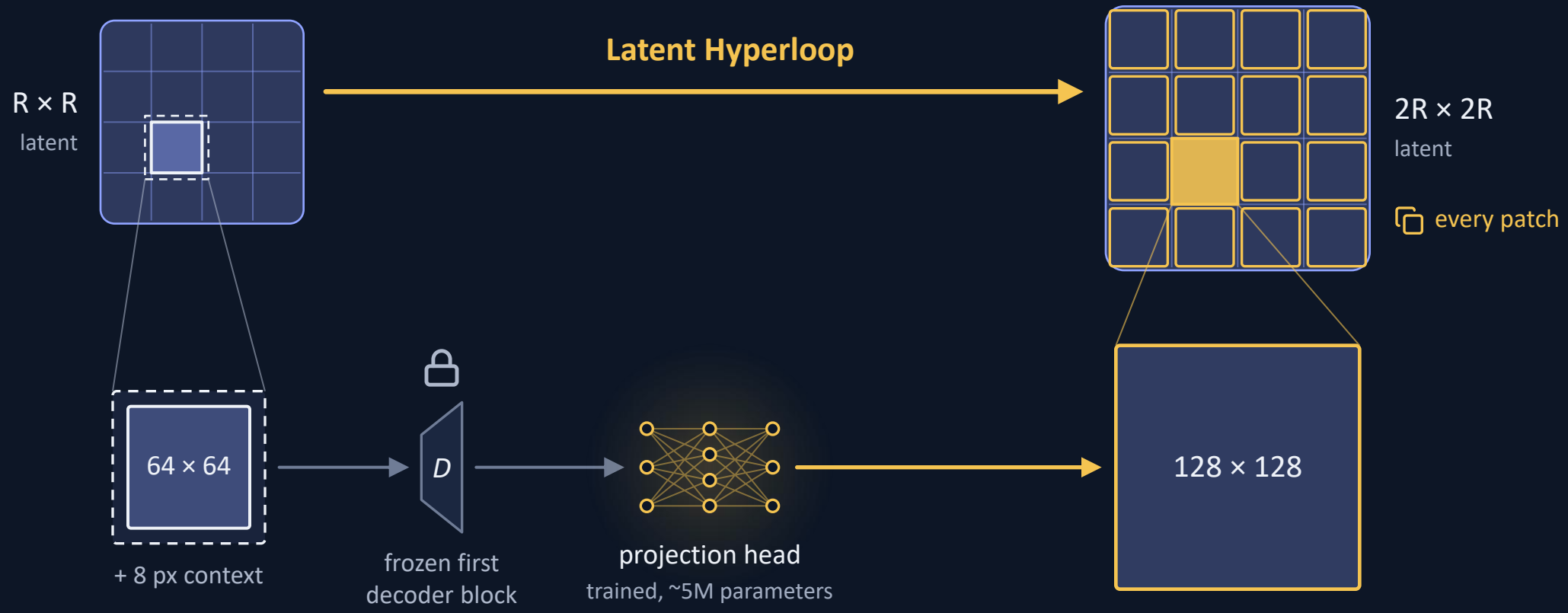
Interpolate(z, 2) ≠ E(Interpolate(D(z), 2))

D: VAE decoder · E: VAE encoder (nonlinear)

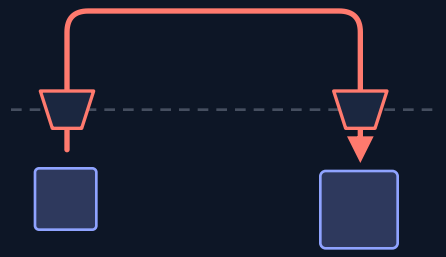
Replace the round trip, then decode sparsely



Latent Hyperloop: a patch-wise latent shortcut



Hyperloop output closely matches the round trip



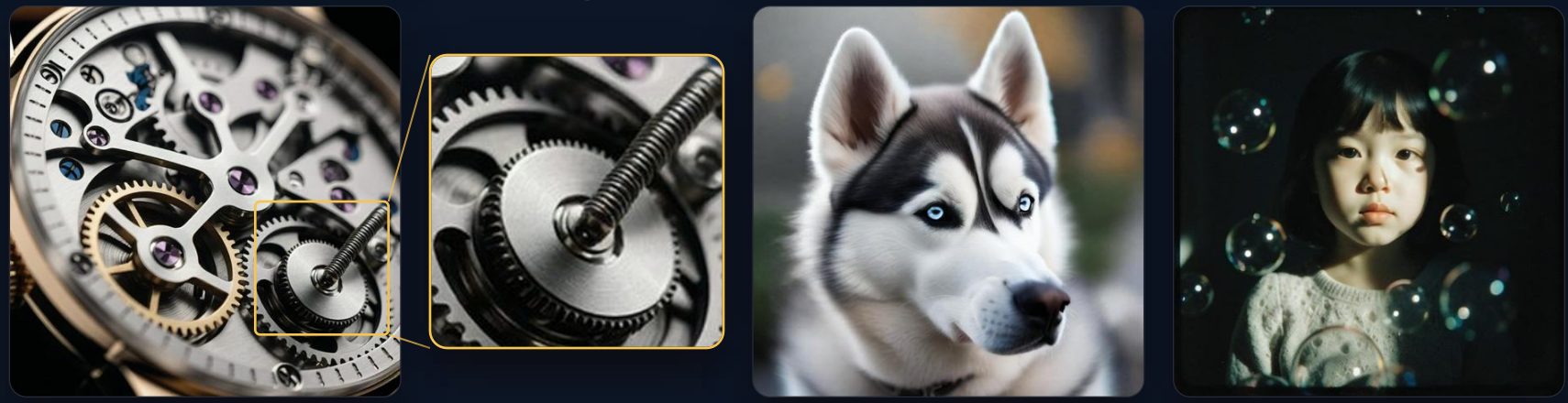
RGB round trip ×2



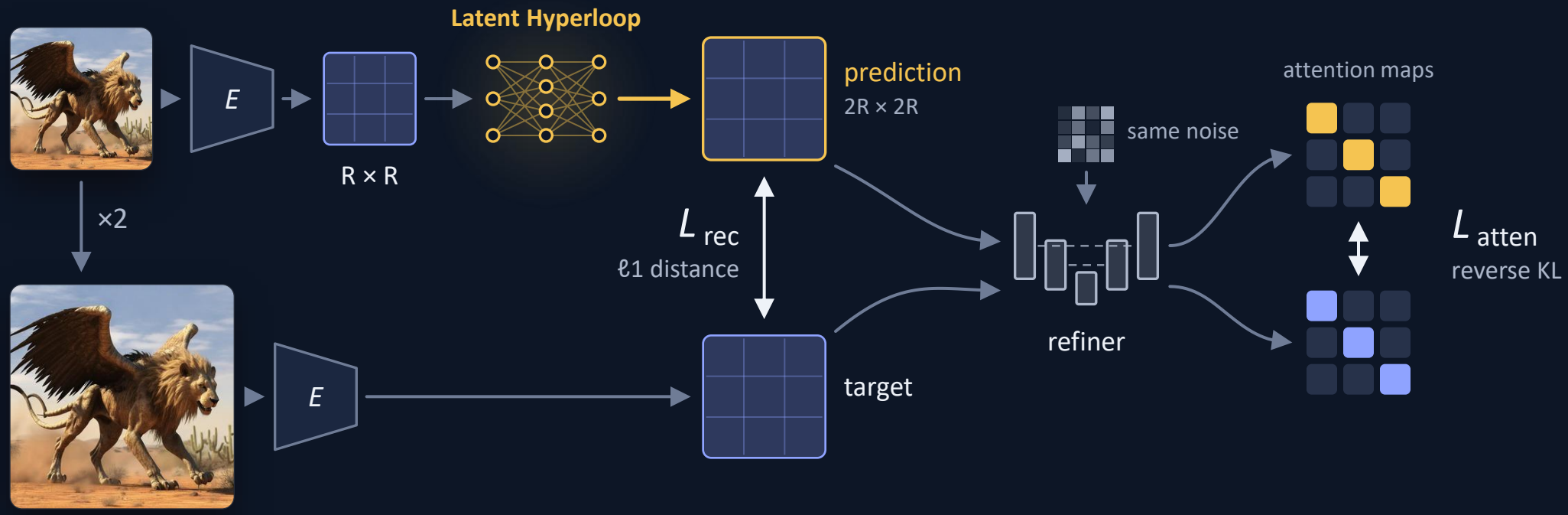
≈



Latent Hyperloop ×2

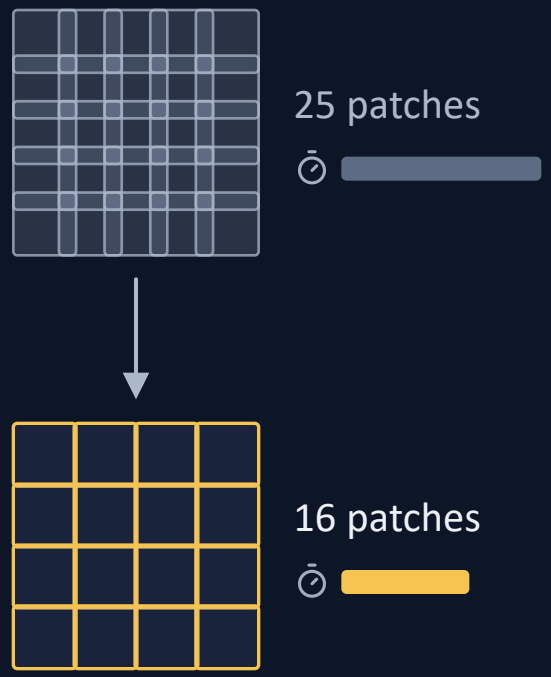


Train on clean pairs with two losses



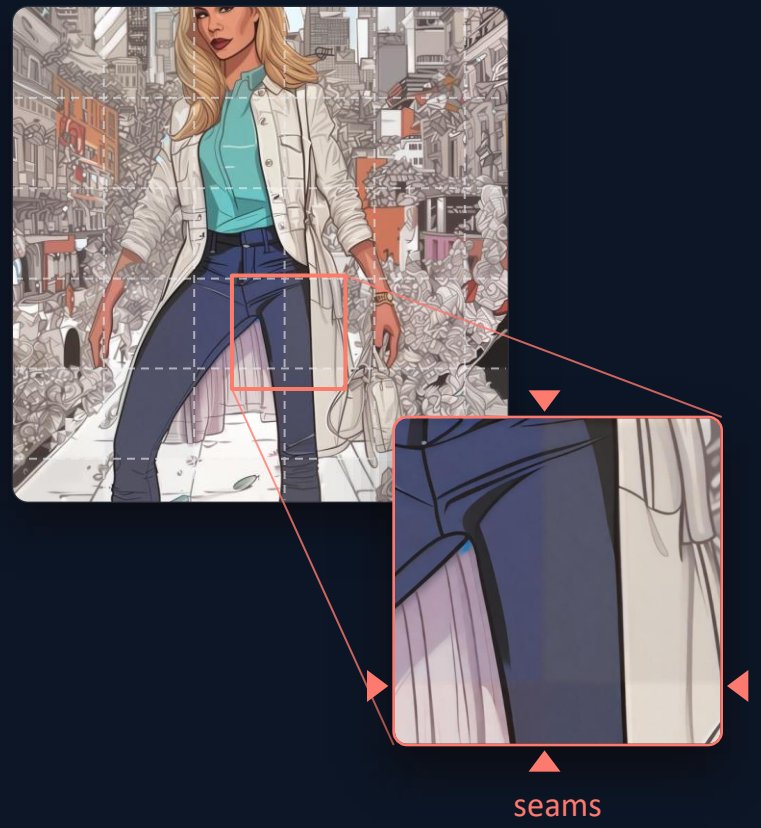
Reduce final decoding cost by reducing overlapping ratio?

1 Idea: smaller overlap

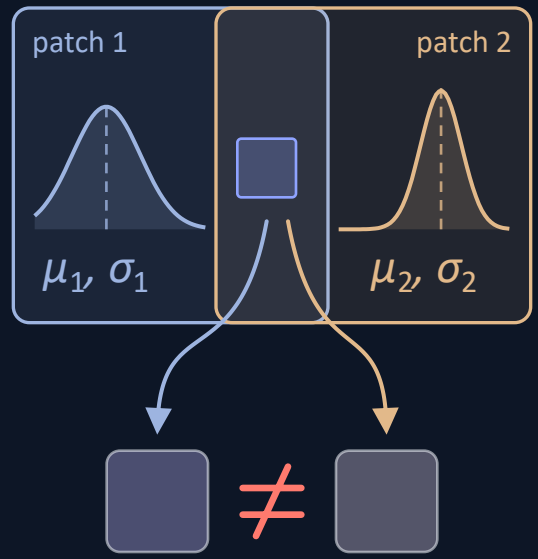


2 Issue: seams at patch borders

overlap factor 0.03125

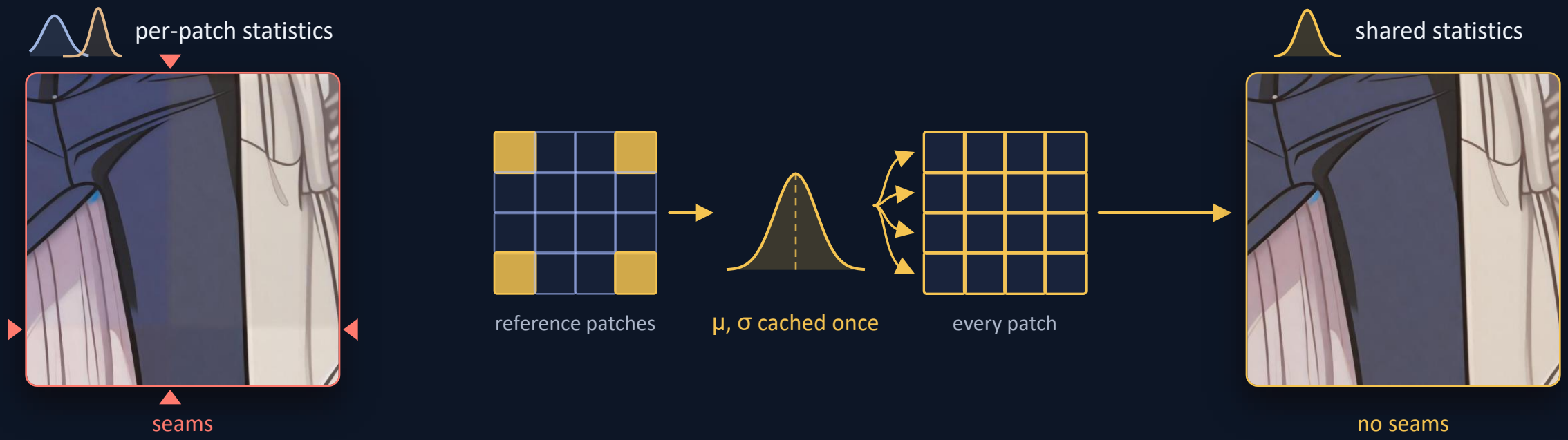


3 Cause: inconsistent GroupNorm statistics



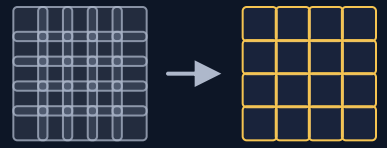
At each location in the overlapping region: same input latent, different decoded pixels from each patch due to mismatched GroupNorm

Share GroupNorm statistics across patches





up to **8×** less overlap



up to **30–40%** fewer patches

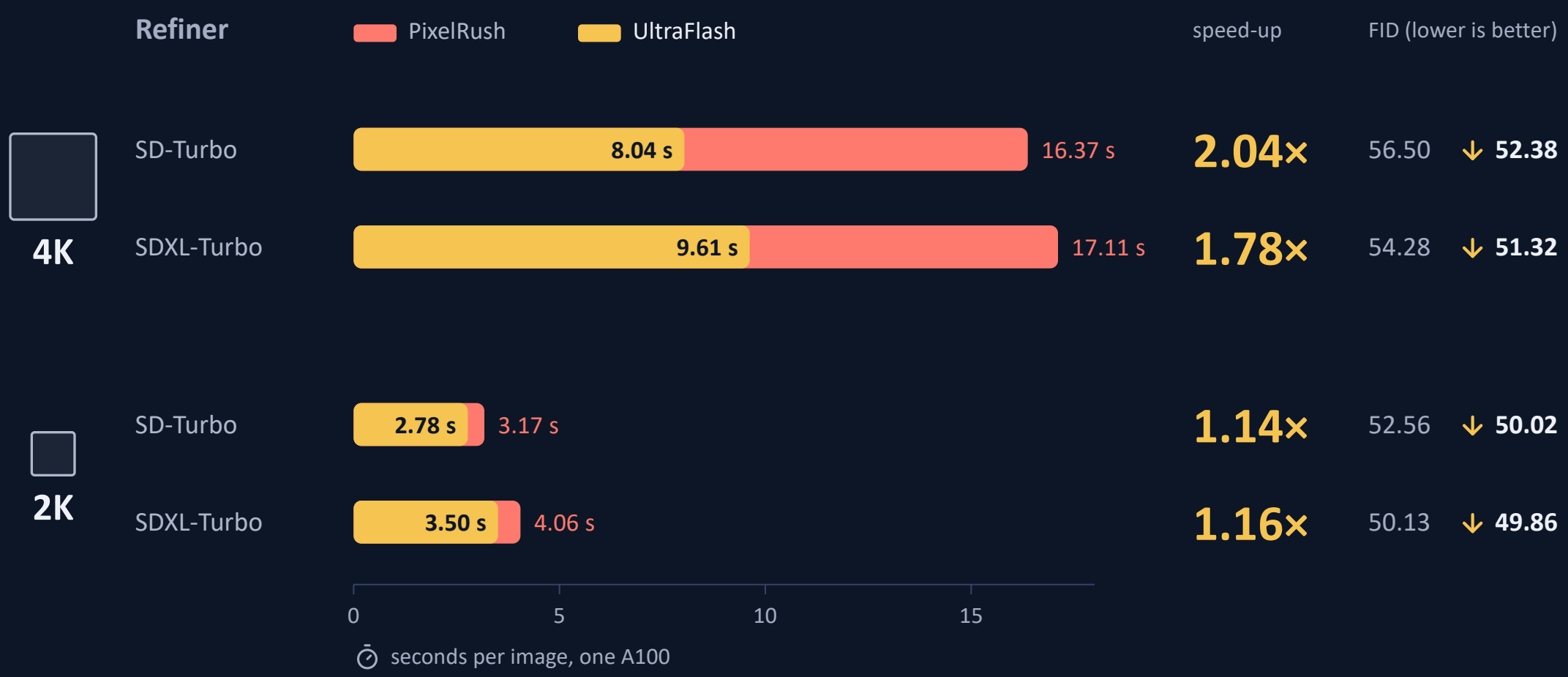


no training
statistics only

UltraFlash halves 4K latency and lowers FID

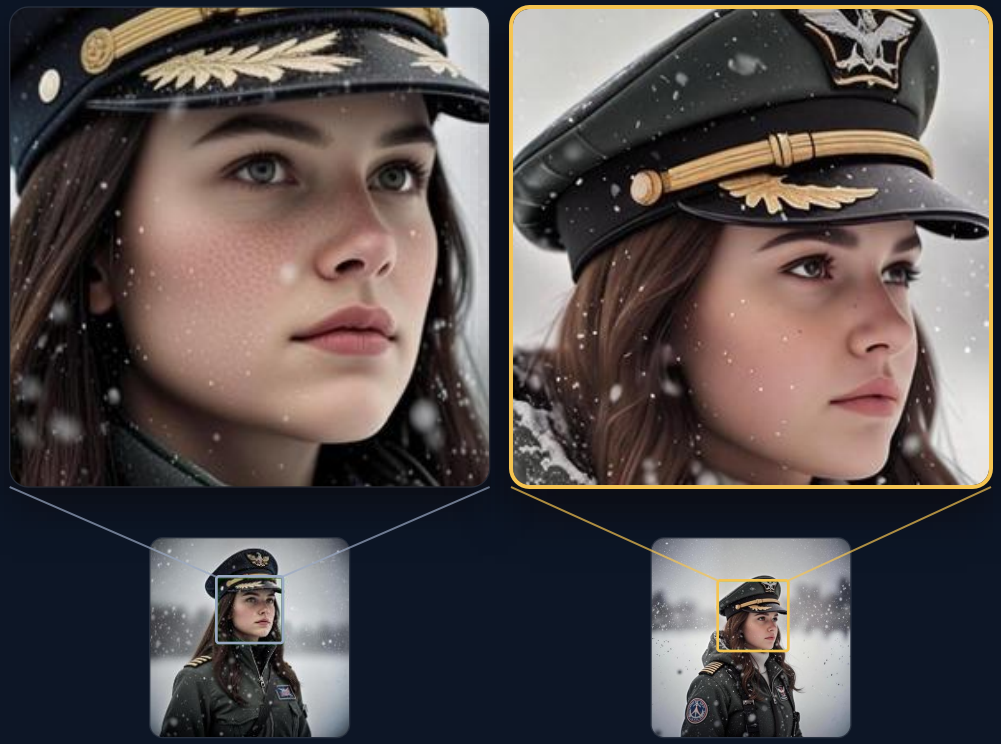


UltraFlash is faster in all four settings

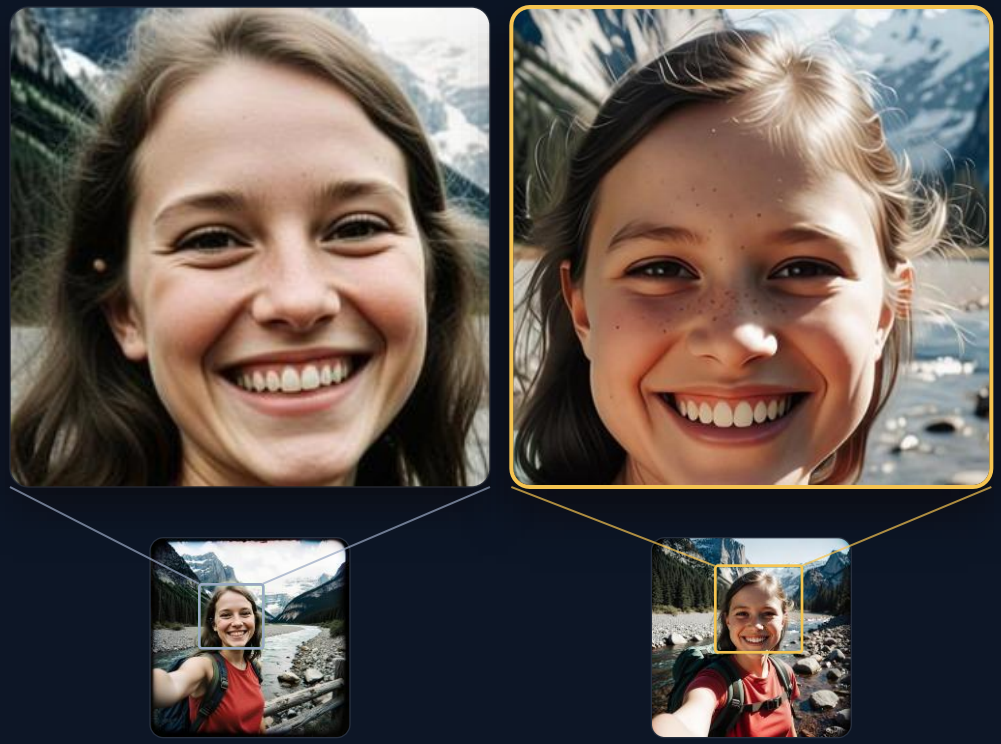


UltraFlash keeps fine detail in half the time

PixelRush 16.37 s UltraFlash 8.04 s



PixelRush 16.37 s UltraFlash 8.04 s

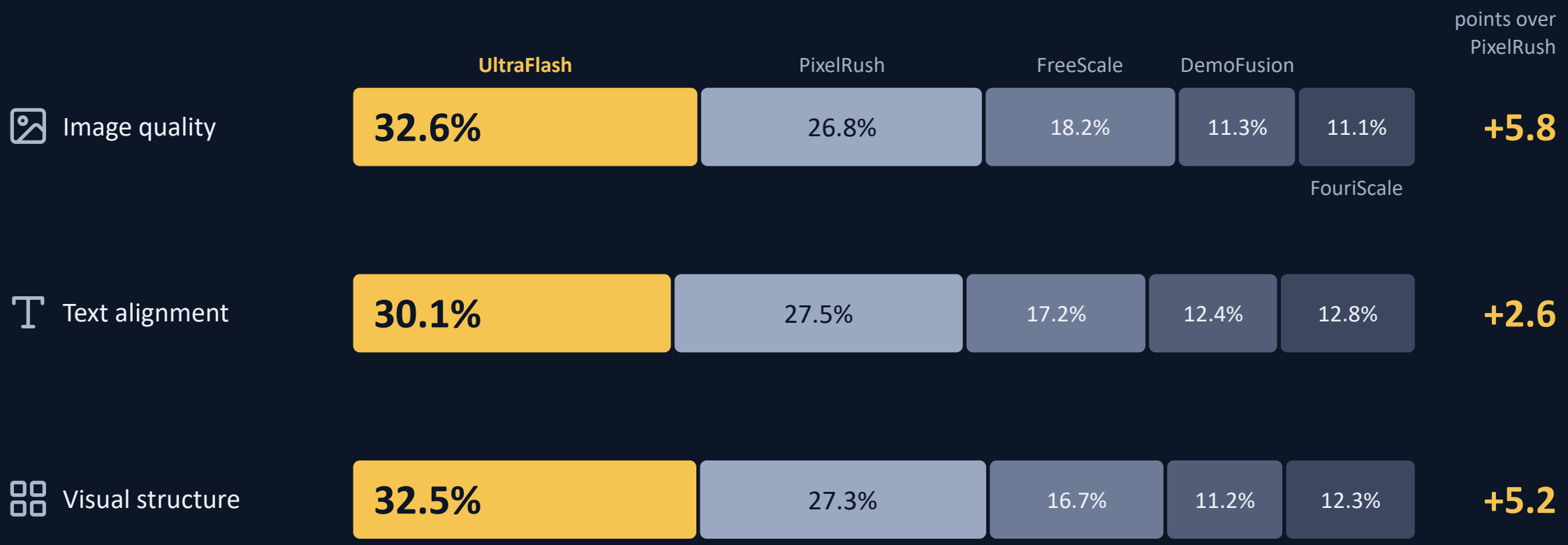


Lower FID and latency than other 4K routes



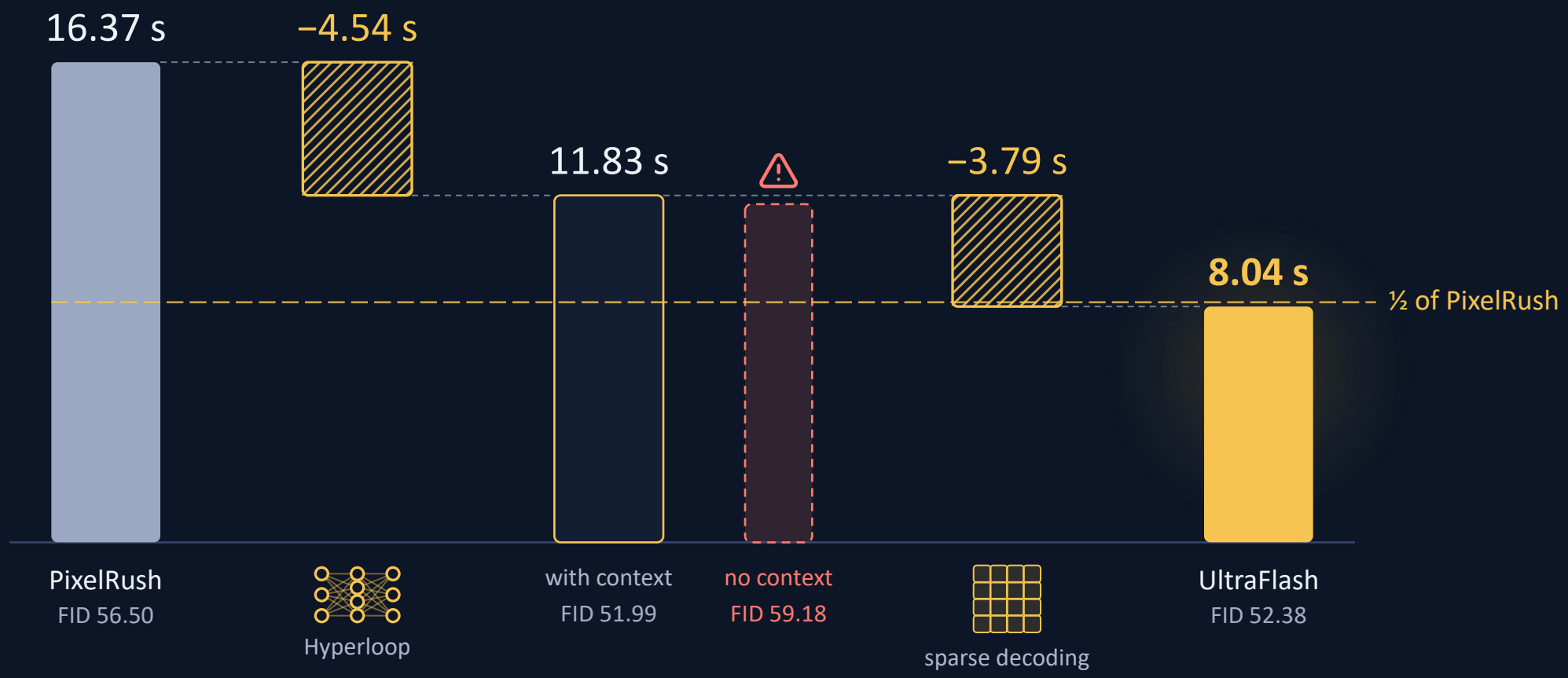
UltraFlash is preferred on all three criteria

137 participants · 25 prompts · 3,425 comparisons

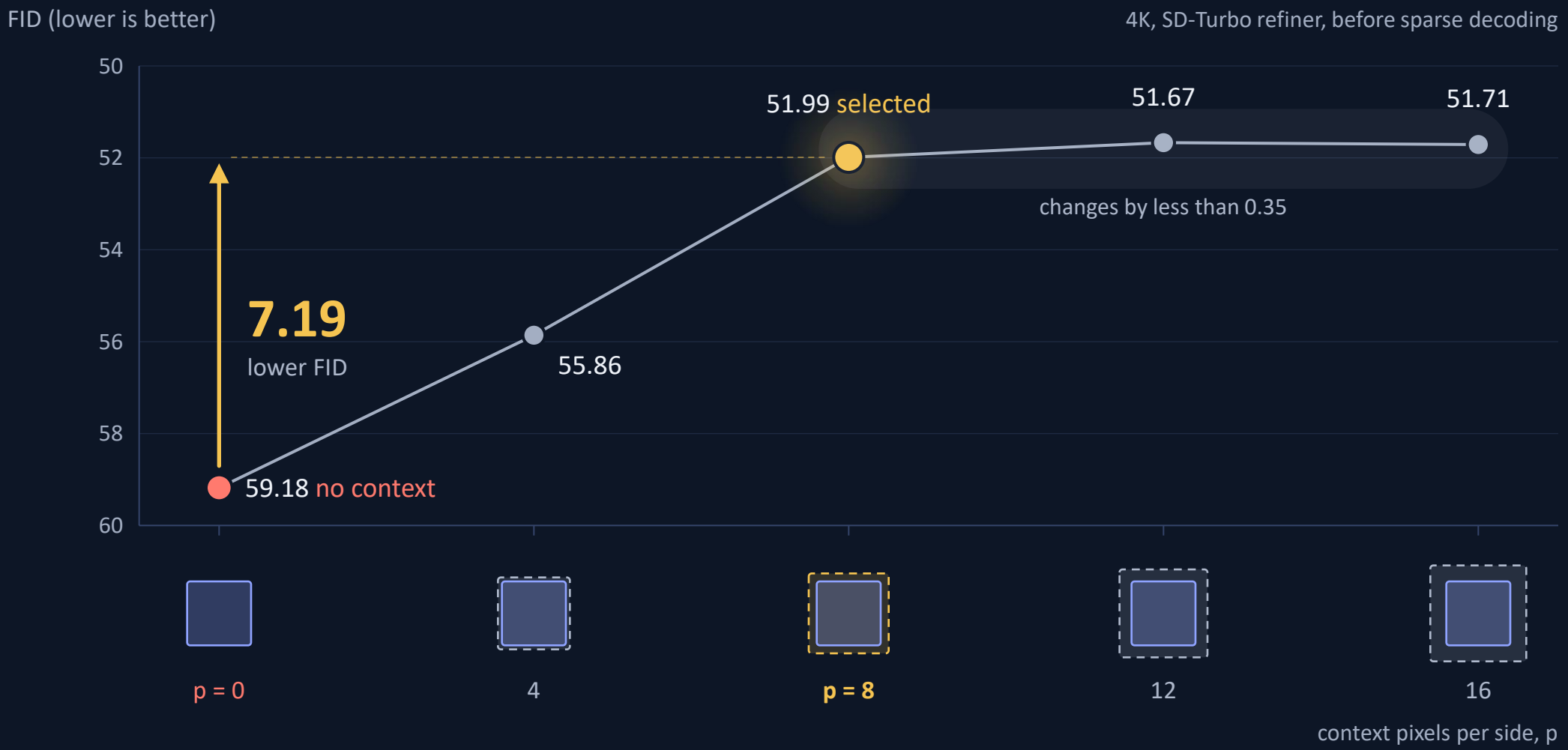


Both components together halve 4K latency

seconds per 4096² image, SD-Turbo refiner



FID improves with context up to p = 8



UltraFlash works across base and refiner models



Latency from 2K to 16K



Uncurated UltraFlash outputs at 4096 × 4096, FLUX base







Uncurated UltraFlash outputs at 4096 × 4096, SDXL base





