RaPD: Resolution-Agnostic Pixel Diffusion via Semantics-Enriched Implicit Representations

Yanhao Ge[1] Shanyan Guan[2] Weihao Wang[1,2] Ying Tai[3] Mingyu You[1,4]

[1] College of Electronic and Information Engineering, Tongji University [2] vivo Mobile Communication Co., Ltd. [3] Nanjing University
[4] State Key Laboratory of Autonomous Intelligent Unmanned Systems; Frontiers Science Center for Intelligent Autonomous Systems, Ministry of Education

Project Page | Code Repository

Can One Generated Representation Support Different Output Resolutions?

(a) VAE latent



coarse, 8× down-sampled

(b) Pixel tokens



fine, but still discrete

(c) Continuous NIF (Ours)



smooth field, query at any q

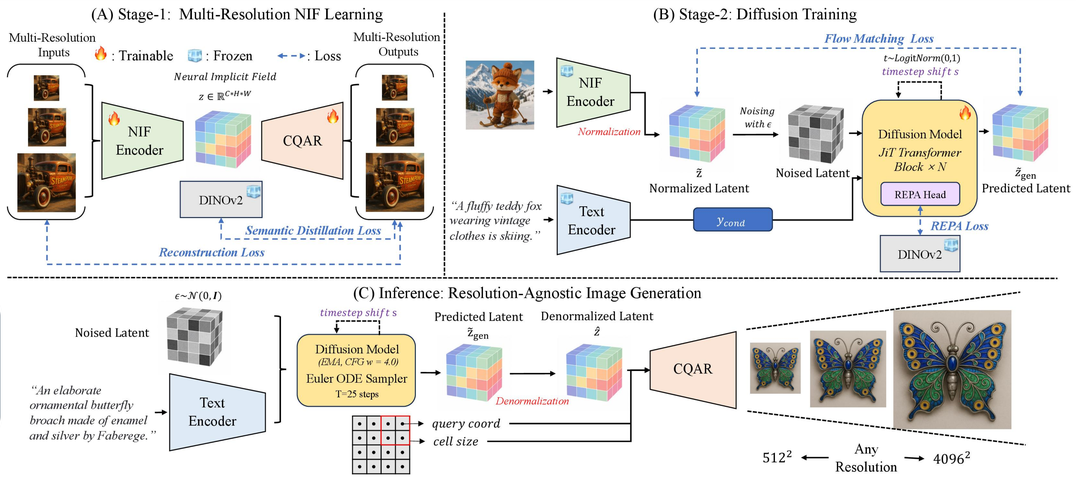
The visual world is continuous, but VAE-latent and pixel-token diffusion both operate on fixed spatial grids.

Higher resolution ⇒ more tokens, more compute

Coordinate-based decoders support flexible output sizes. But reconstruction-trained latents are not necessarily ready for generation.

Question:
Can we learn a generation-ready representation that renders at continuous coordinates?

Method: RaPD makes NIF generation-ready with SRG and CQAR



Stage-1: NIF AE + Semantic Distillation ⇒ generation-aware latent.
Stage-2: Flow matching + REPA with s = 32 timestep shift on dense NIF latent.
Inference: One denoised latent ⇒ any H′ ×W ′ via the CQAR query grid.

SRG: Semantic Representation Guidance

- Semantic distillation. Align the latent with frozen DINOv2 patch features:

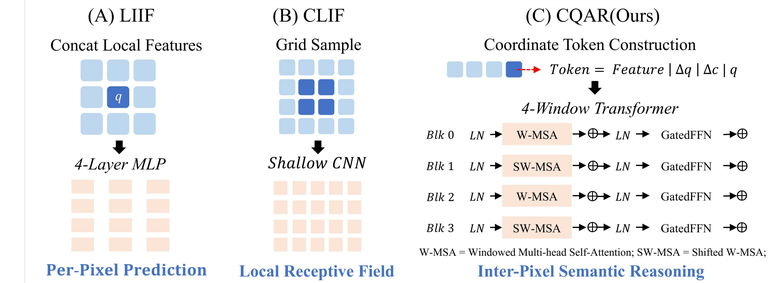
$$Latent \Rightarrow \quad \mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{mcos}} + \mathcal{L}_{\text{mdms}} \quad \Leftarrow DINO\ Teacher$$

- Latent-density-aware timesteo shift. Emphasize noisier states during training:

$$t_0 = \text{sigmoid}(\xi), \quad \xi \sim \mathcal{N}(0,1), \qquad t = \frac{t_0}{t_0 + (1 - t_0) \cdot s}.$$

Make reconstruction-oriented latents generation-ready

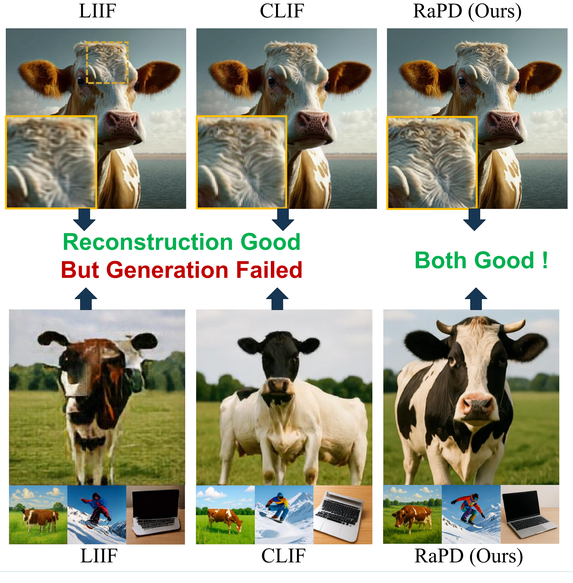
CQAR: Coordinate-Queried Attention Renderer



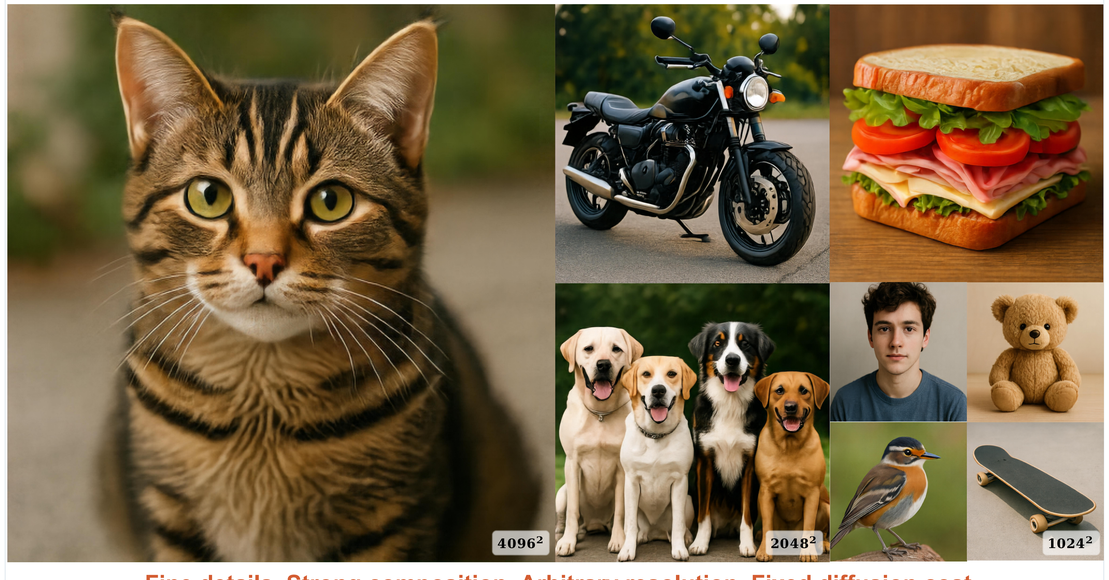
Inter-pixel reasoning for coherent, arbitrary-resolution rendering

The Reconstruction–Generation Gap

LIIF&CLIF enable continuous reconstruction, but are not generation-ready.

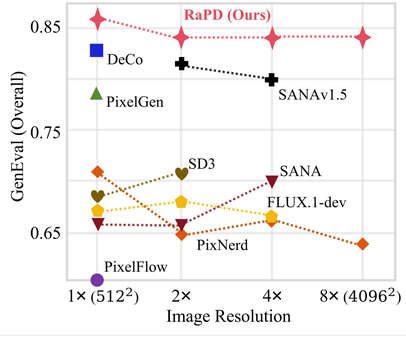


Arbitrary Resolution Generation Results



Fine details. Strong composition. Arbitrary resolution. Fixed diffusion cost.

Generation Quality and Performance



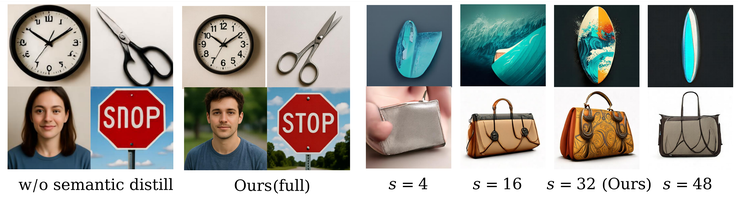
- **0.86 GenEval / 82.9 DPG**

Method	GenEval	DPG
PixNerd	0.71	81.5
PixelGen	0.79	80.4
DeCo	0.83	81.2
RaPD	**0.86**	**82.9**

- **FIxed diffusion cost**
- **4K in 2.13 seconds**

Method	512^2	4096^2
PixelGen	0.19 s	61.46 s
DeCo	0.21 s	62.72 s
RaPD	**0.23 s**	**2.13 s**

Effect of Semantic Representation Guidance



Both semantic distillation and timestep shift helps improve semantic