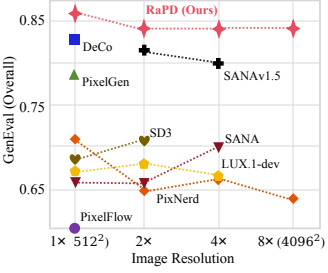
RaPD: Resolution-Agnostic Pixel Diffusion via Semantics-Enriched Implicit Representations

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

[1]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: https://wangwhtj.github.io/rapd-page/

Can one generated representation support different output resolutions?

▷ *We start by questioning a hidden assumption shared by today's image generative models.*

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

Higher resolution ⇒ more tokens, more compute — linear-to-quadratic scaling on the diffusion side.

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

Our question

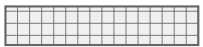
Can we learn a generation-ready representation that renders at continuous coordinates?

(a) VAE latent



coarse, 8× down-sampled

(b) Pixel tokens



fine, but still discrete

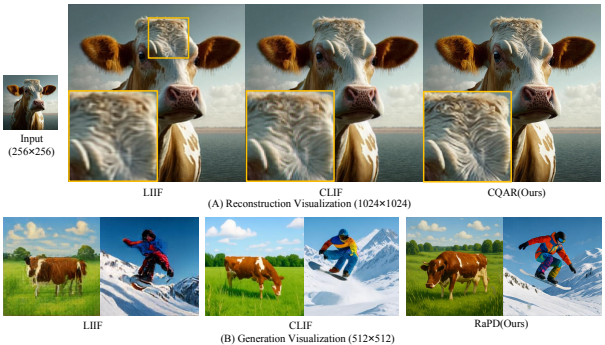
(c) Continuous NIF (Ours)



smooth field, query at any q

Reconstruction-trained image fields are not automatically generation-ready

▷ *Applying a diffusion model to LIIF / CLIF latents exposes a reconstruction–generation gap.*



Reconstruction–generation gap of existing NIF representations: strong super-resolution reconstruction (top) but weak text-to-image generation (bottom).

LIIF / CLIF: trained for **reconstruction** (pixel loss + lightweight decoder).

Apply DiT-style diffusion on top ⇒ only **coarse semantics, poor visual fidelity**.

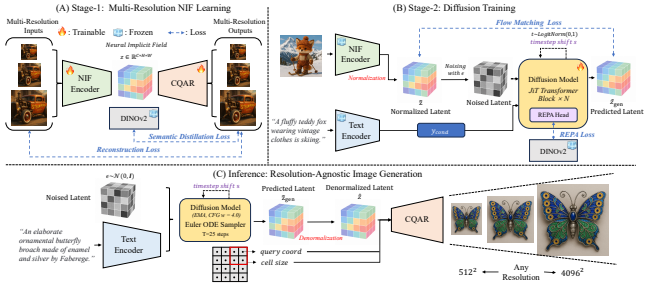
Reconstruction–generation gap

Latent encodes *low-level texture*, not semantic structure.

Decoder is a *local operator* — no inter-pixel reasoning.

RaPD: shape the NIF (*SRG*) and redesign the renderer (*CQAR*), in two stages

▷ *We close the gap with two coupled innovations: SRG makes the latent generation-ready, CQAR makes the renderer semantic-aware.*



Overview of RaPD. (A) Stage-1: semantically guided multi-resolution NIF learning. (B) Stage-2: text-conditioned flow matching in normalized NIF latents. (C) Inference: render one denoised latent at arbitrary resolutions via CQAR.

(A) Stage-1. NIF AE + **Semantic Distillation** ⇒ generation-aware latent.

(B) Stage-2. Flow matching + REPA with $s = 32$ timestep shift on dense NIF latent.

(C) Inference. **One denoised latent** → any $H' \times W'$ via the CQAR query grid.

SRG turns reconstruction-oriented NIF latents into generation-aware ones

▷ *Innovation 1: shape the latent itself. The latent must inherit semantic structure from a vision foundation model.*

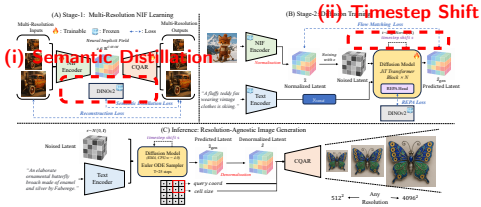
Two components:

Semantic Distillation: pool $\boldsymbol{z} \to 1\times1$ conv $\to$ align with frozen *DINOv2* patch features (marginal-cosine + distance-matrix losses).

Latent-Density-Aware Timestep Shift:

$$t = \frac{t_0}{t_0+(1-t_0)\,s}, \qquad s=32$$

(vs SD3's $s\in[1,3]$); pushes mass toward noisier states where dense NIF needs most denoising.



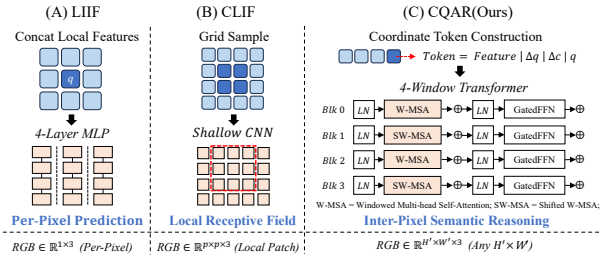
Two SRG components highlighted in red on the RaPD framework.

Why not just REPA / VA-VAE?

Our alignment supervises a **dense NIF latent** directly, unlike prior work on downsampled VAE latents (REPA-E, VA-VAE) or denoiser features (REPA). The projection branch transfers semantic structure **without changing the latent layout**, preserving the coordinate-based rendering pathway.

CQAR enables inter-pixel reasoning at decoding via coordinate-conditioned attention

▷ *Innovation 2: redesign the renderer. A semantically rich latent is wasted if the decoder is a pure pointwise operator.*



Implicit decoder comparison. (A) LIIF: per-pixel MLP. (B) CLIF: shallow CNN. (C) CQAR: coordinate-conditioned windowed attention over output tokens.

(A) LIIF	**(B) CLIF**	**(C) CQAR (Ours)**
per-pixel MLP	shallow CNN	windowed attention over
no output-token attention	**limited receptive field**	**output coordinate tokens**

Quantitative comparison on GenEval at 512×512

▷ *Progressive training (256 → 512) reaches 0.86 overall with a 1.1B denoiser.*

Method	#Params	GenEval ↑						
		Single Obj.	Two Obj.	Counting	Colors	Position	Color Attr.	Overall
▽ *Latent Diffusion Models*								
LDM	1.4B	0.92	0.29	0.23	0.70	0.02	0.05	0.37
SD3	8B	0.98	0.84	0.66	0.74	0.40	0.43	0.68
FLUX.1-dev	12B	0.99	0.81	0.79	0.74	0.20	0.47	0.67
SANA	1.6B	0.99	0.79	0.63	0.88	0.18	0.47	0.66
▽ *Pixel Diffusion Models*								
PixelFlow-XL/4	0.9B	–	–	–	–	–	–	0.60
PixNerd-XXL/16	1.2B	0.97	0.86	0.44	0.83	0.71	0.53	0.73
DeCo-XXL/16	1.1B	**1.00**	0.92	0.72	0.91	**0.80**	**0.79**	**0.86**
PixelGen-XXL/16	1.1B	0.99	0.88	0.59	0.90	0.70	0.70	0.79
PixNerd-XXL/16 (‡)	1.2B	0.98	0.82	0.43	0.81	0.66	0.55	0.71
DeCo-XXL/16 (‡)	1.1B	0.99	0.92	0.65	0.91	0.71	0.77	0.83
PixelGen-XXL/16 (‡)	1.1B	0.99	0.88	0.58	0.91	0.70	0.67	0.79
RaPD	1.1B	0.99	**0.93**	0.72	**0.93**	0.77	0.74	0.85
RaPD (256→512)	1.1B	0.99	**0.93**	**0.77**	0.92	0.79	0.78	**0.86**

Text-to-image generation on GenEval at a 512×512 resolution. ‡ denotes results obtained by running the officially released models with the same random seed and computing environment as RaPD.

Quantitative comparison on DPG-Bench at 512×512

▷ *Progressive training reaches 82.9, above the reproduced pixel-diffusion baselines.*

Method	#Params	DPG-Bench ↑					
		Global	Entity	Attribute	Relation	Other	Average
▽ *Latent Diffusion Models*							
SD v2	0.9B	77.7	78.1	74.9	80.7	80.7	68.1
PixArt-α	0.6B	75.0	79.3	78.6	82.6	77.0	71.1
SANA	1.6B	90.3	91.2	89.0	88.9	92.0	85.5
▽ *Pixel Diffusion Models*							
PixelFlow-XL/4	0.9B	–	–	–	–	–	77.9
PixNerd-XXL/16	1.2B	80.5	87.9	87.2	**91.3**	72.8	80.9
DeCo-XXL/16	1.1B	–	–	–	–	–	81.4
PixNerd-XXL/16 (‡)	1.2B	87.2	87.1	88.9	87.1	**89.8**	81.5
PixelGen-XXL/16 (‡)	1.1B	81.4	88.8	86.7	88.9	88.9	80.4
DeCo-XXL/16 (‡)	1.1B	**88.6**	88.4	88.9	87.6	86.0	81.2
RaPD	1.1B	85.1	87.7	<u>88.9</u>	<u>89.6</u>	84.1	<u>81.6</u>
RaPD (256→512)	1.1B	81.3	**89.9**	**89.7**	88.9	81.2	**82.9**

Text-to-image generation results on DPG-Bench at a 512×512 resolution. ‡ denotes results obtained by running the officially released models with the same random seed and computing environment as RaPD.

Qualitative comparison: visual fidelity and prompt adherence

▷ *Side-by-side with the closest pixel-diffusion baselines on the same prompts.*



Qualitative text-to-image comparison with baselines (PixelGen, DeCo, PixNerd) at 512^2.

Fixed latent → Any Resolution: GenEval and DPG-Bench across scales

▷ *Same model and same denoised latent; only the CQAR query coordinate grid changes.*



An intricately detailed oil painting depicts a raccoon dressed in a black suit with a crisp white shirt and a red bow tie. The raccoon stands upright, donning a black top hat and gripping a wooden cane with a silver handle in one paw, while the other paw clutches a dark garbage bag. The background of the painting features soft, brush-stroked trees and mountains, reminiscent of traditional Chinese landscapes, with a delicate mist enveloping the scene.



An assortment of artisanal cheese wheels, each exhibiting a distinct texture and color palette, ranging from pale creamy whites to rich oranges, are spread across a rough-hewn wooden table brimming with character. The warm rays of the early morning sun filter through a window to the side, lending a natural illumination that highlights the subtle hues of the cheeses. In the background, the soft shadows cast by the window frame ...

(A) Results on GenEval: samples generated from short, compositional prompts and rendered at multiple output resolutions;
(B) Results on DPG-Bench: RaPD faithfully captures fine-grained semantic and stylistic details specified in the prompt.

High-resolution generation: strong composition, remaining prompt-detail gap

▷ *Evaluate the progressively trained* $256 \rightarrow 1024$ *model at matched output resolutions.*

Method	Params	1024^2		2048^2	
		GenEval ↑	DPG ↑	GenEval ↑	DPG ↑
SANA	1.6B	0.66	84.8	0.70	83.1
SANA-v1.5	1.6B	0.82	84.5	0.80	83.8
FLUX.1-dev	12B	0.68	83.1	0.66	82.8
RaPD ($256 \rightarrow 1024$)	1.1B	**0.85**	80.3	**0.85**	80.1

Best GenEval among these methods at both evaluated resolutions.

2048^2 is beyond the 1024^2 training supervision range.

Lower DPG scores show room to improve detailed prompt adherence.

Training setup and the role of semantic supervision

▷ *Stage-1 latent distillation and Stage-2 denoiser alignment play complementary roles.*

Data: approximately 36M BLIP3o images; 60k curated images for fine-tuning.

Stage-1: train the encoder and renderer for 10k steps with semantic distillation.

Stage-2: freeze the encoder and renderer; train the denoiser with flow matching and REPA feature alignment.

Sampling: 25 Euler steps, classifier-free guidance scale 4.0.

Setting	GenEval	DPG
No Stage-1 distillation	0.83	78.9
No Stage-2 REPA	0.61	56.6
Full model	**0.85**	**81.6**

Interpretation

Removing REPA gives the largest drop under the same training budget. Stage-1 distillation provides further gains. These ablations use the base RaPD model, not the progressive-training variant.

Efficiency: Diffusion cost stays fixed as output resolution grows

▷ *Increasing output size leaves RaPD's denoising token count unchanged; rendering still adds cost.*

Method	512×512		1024×1024		2048×2048		4096×4096	
	#Tokens	Time (s)	#Tokens	Time (s)	#Tokens	Time (s)	#Tokens	Time (s)
PixelGen	1,024	0.19	4,096	0.60	16,384	4.72	65,536	61.46 (+61.27)
DeCo	1,024	0.21	4,096	0.67	16,384	5.03	65,536	62.72 (+62.51)
RaPD	1,024	0.23	1,024	0.34	1,024	0.68	1,024	**2.13** (+1.9)

Computational scaling with output resolution. Per-image inference time on an NVIDIA B200 GPU is reported at 1× to 8× output scales. RaPD keeps the expensive denoising stage nearly fixed, while fixed-grid pixel diffusion methods grow with the output resolution.

Take-away

Diffusion cost is fixed regardless of output size. Only **CQAR rendering** grows linearly with the queried output token count.

Ablation Study: Each Component is Necessary

▷ *The paper studies semantic supervision, timestep shift, and coordinate-based rendering.*



Ablation study of representation alignment and CQAR.

(a) Semantic supervision:
distillation and renderer ablations



Effect of timestep shift on a test prompt.

(b) Timestep shift: small s degrades structure; $s = 32$ best

Conclusions: Contributions, Limitations, and Future work

▷ *We end with an honest summary: what we have shown, where the boundaries lie, and what comes next.*

Contributions

- **(i) RaPD:** a resolution-agnostic pixel diffusion framework that directly generates continuous image signals in a NIF latent space.
- **(ii) *SRG*:** transforms reconstruction-oriented continuous latents into generation-aware representations via semantic distillation + density-aware timestep shift.
- **(iii) *CQAR*:** coordinate-conditioned, scale-aware Transformer renderer for spatially coherent, arbitrary-resolution rendering with fixed diffusion cost.

Limitations

- Training data scale (36M; vs commercial billion-scale).
- Compact text encoder (Qwen3-1.7B).
- Inference stack not yet hardware-optimized.

Future Work

- Scale training data and denoiser capacity.
- Editing & control (ControlNet, instruction editing).
- Hardware-aware inference (FlashAttention, FP8, fused kernels).

Paper: arxiv.org/abs/2605.15908 Paper ID: 5904