UniFlow: A Unified Pixel Flow Tokenizer for Visual Understanding and Generation

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

Zhengrong Yue, Haiyu Zhang, Xiangyu Zeng, Boyu Chen, Chenting Wang, Shaobin Zhuang, Lu Dong, Yi Wang, Limin Wang, Yali Wang

Challenges in Complex Video Stylization

Existing tokenizers face a fundamental performance trade-off:

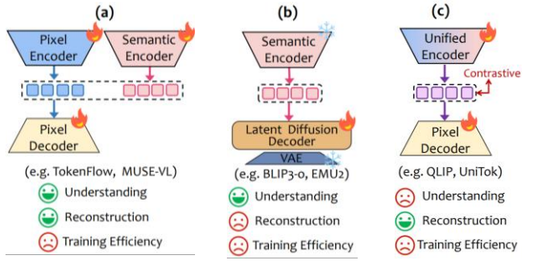
- High-level semantic abstraction (Understanding)
- Low-level pixel reconstruction (Generation)

Prior Approaches:

(1) Dual Encoders;

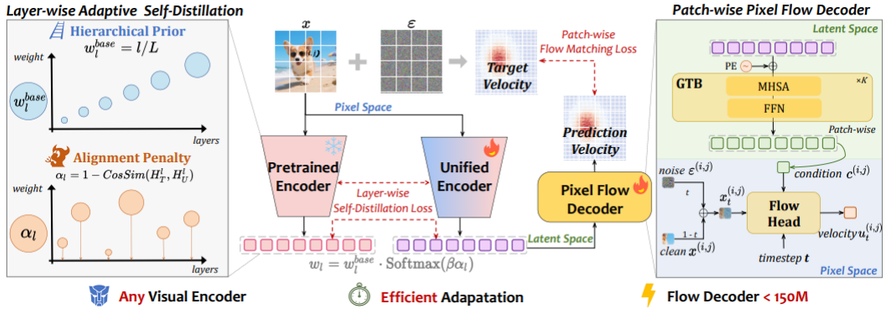
(2) Frozen Semantic + Diffusion;

(3) Unified + Pixel Decoder;



*How can we efficiently unify visual representations within a single tokenizer to achieve both **powerful semantic understanding** and **high-fidelity reconstruction**?*

UniFlow Framwork



Two Core Designs:

(1) LAYER-WISE ADAPTIVE SELF-DISTILLATION

Dynamic weight for each layer:

◇Deep layers: Strong semantic preservation

◇Shallow layers: Flexible detail adaptation

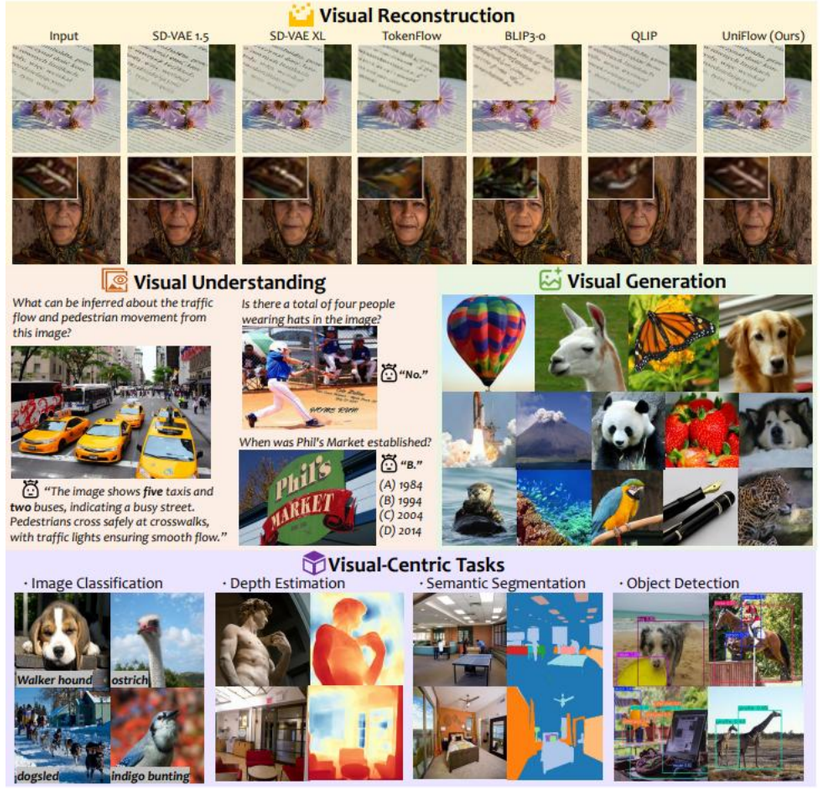
(2) PATCH-WISE PIXEL FLOW DECODER

Flow Matching in pixel space (not latent)

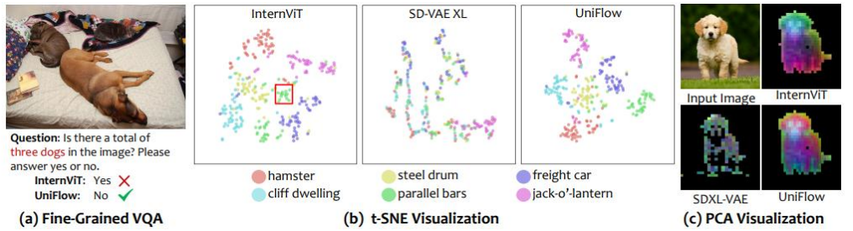
◇Global Transformer Blocks eliminates grid artifacts

◇Single-step Euler inference

Downstream Task Comparison



What Does UniFlow Learn?



Quantitative Results

Table 1: **Comparison of reconstruction quality on the 256 × 256 ImageNet-1K and MS-COCO 2017 validation sets.** "Ratio" denotes downsampling ratio; "Type" indicates tokenizer traits (VQ usage and decoder type). UniFlow achieves state-of-the-art (SOTA) performance in unified tokenizers while also being competitive with the best generative tokenizers. See Appendix B.1 for data details.

Method	Type	Training Data	Ratio	ImageNet-1K			MS-COCO 2017		
				PSNR ↑	SSIM ↑	rFID ↓	PSNR ↑	SSIM ↑	rFID ↓
Generative Only Tokenizer									
Cosmos-DI (Agarwal et al., 2025)	Discrete-Pixel	–	16	19.98	0.54	4.40	19.22	0.48	11.97
LlamaGen (Sun et al., 2024a)	Discrete-Pixel	MS+IN-1K	16	20.65	0.54	2.47	20.28	0.55	8.40
Open-MAGVIT2 (Luo et al., 2024)	Discrete-Pixel	Mixed100M	16	22.70	0.64	1.67	22.31	0.65	6.76
BSQ-ViT (Yang et al., 2021)	Discrete-Pixel	IN-1K	16	28.14	0.81	0.45	–	–	–
SD-VAE 1.x (Rombach et al., 2022)	Continuous-Pixel	OImg	8	23.54	0.68	1.22	23.21	0.69	5.94
SD-VAE 2.x (Rombach et al., 2022)	Continuous-Pixel	OImg+LAae	8	23.54	0.68	1.22	26.62	0.77	4.26
OmniTokenizer (Wang et al., 2024a)	Continuous-Pixel	IN-1K+K600	8	26.74	0.82	1.02	26.44	0.83	4.69
SD-VAE XL (Podell et al., 2023)	Continuous-Pixel	OImg+LAae++	8	27.37	0.78	0.67	27.08	0.80	3.93
Qwen-Image (Wu et al., 2025a)	Continuous-Pixel	–	8	32.18	0.90	1.459	32.01	0.91	4.62
SD-VAE 3 (Esser et al., 2024)	Continuous-Pixel	–	8	31.29	0.87	0.20	31.18	0.89	1.67
Wan2.1 (Wan et al., 2025a)	Continuous-Pixel	–	8	31.34	0.89	0.95	31.19	0.90	3.45
FLUX-VAE (Labs, 2024)	Continuous-Pixel	–	8	**32.74**	**0.92**	**0.18**	**32.32**	**0.93**	**1.35**
Cosmos-CI (Agarwal et al., 2025)	Continuous-Pixel	–	16	25.07	0.70	0.96	24.74	0.71	5.06
VA-VAE (Yao et al., 2025)	Continuous-Pixel	IN-1K	16	27.96	0.79	0.28	27.50	0.81	2.71
Wan2.2 (Wan et al., 2025b)	Continuous-Pixel	–	16	31.25	0.88	0.749	31.10	0.89	3.28
SelfTok (Luo et al., 2024)	Discrete-Diffusion	IN-1K	–	24.14	0.71	0.70	–	–	–
FlowMo-Hi (Shaulov et al., 2025)	Discrete-Diffusion	IN-1K	–	26.93	0.79	0.56	–	–	–
l-DeTok (Yang et al., 2025a)	Continuous-Diffusion	IN-1K	16	–	–	0.68	–	–	–
Unified Tokenizer									
Show-o (Xie et al., 2024b)	Discrete-Pixel	-	16	21.34	0.59	3.50	20.90	0.59	9.26
QLIP-B (Zhao et al., 2025b)	Discrete-Pixel	DC-1B	16	23.16	0.63	3.21	–	–	–
VILA-U (Wu et al., 2024b)	Discrete-Pixel	WL-10B+CY-1B	16	–	–	1.80	–	–	–
Tokenflow (Qu et al., 2025)	Discrete-Pixel	LA+CY	16	21.41	0.69	1.37	–	–	–
DualViTok (Huang et al., 2025)	Discrete-Pixel	Mixed-63M	16	22.53	0.74	1.37	–	–	–
DualToken (Song et al., 2025)	Discrete-Pixel	CC12M	16	23.56	0.74	0.54	–	–	–
MUSE-VL (Xie et al., 2024c)	Discrete-Pixel	IN-1K+CC12M	16	20.14	0.646	2.26	–	–	–
SemHiTok (Chen et al., 2025i)	Discrete-Pixel	CY-50M	16	–	–	1.16	–	–	–
UniTok (Ma et al., 2025)	Discrete-Pixel	DC-1B	16	27.28	0.77	0.41	–	–	–
SeTok (Wu et al., 2025d)	Discrete-Pixel	IN-1K+OImg	–	–	–	2.07	–	–	–
UniLIP (Tang et al., 2025)	Continuous-Pixel	BP-32M	32	22.99	0.747	0.79	–	–	–
EMU2 (Sun et al., 2024b)	Continuous-Diffusion	LA-CO+LAae	14	13.49	0.42	3.27	–	–	–
BLIP3-o (Chen et al., 2025f)	Continuous-Diffusion	BP-32M	16	14.71	0.58	3.18	–	–	–
UniFlow(*CLIP*)	Continuous-Diffusion	IN-1K	14	28.66	0.91	0.67	29.61	0.92	3.69
UniFlow(*SigLIP2*)	Continuous-Diffusion	IN-1K	16	29.38	0.93	0.62	26.38	0.86	3.44
UniFlow(*DINOv2*)	Continuous-Diffusion	IN-1K	14	31.01	0.94	0.54	30.66	0.94	2.81
UniFlow(*InternViT*)	Continuous-Diffusion	IN-1K	14	**33.23**	**0.96**	**0.26**	**32.48**	**0.95**	**1.88**

Ablation Study

Key Findings: • Understanding Enhancement • Adaptive distillation (β=2) >> Uniform • GTB eliminates grid artifacts