ICLR
International Conference On Learning Representations

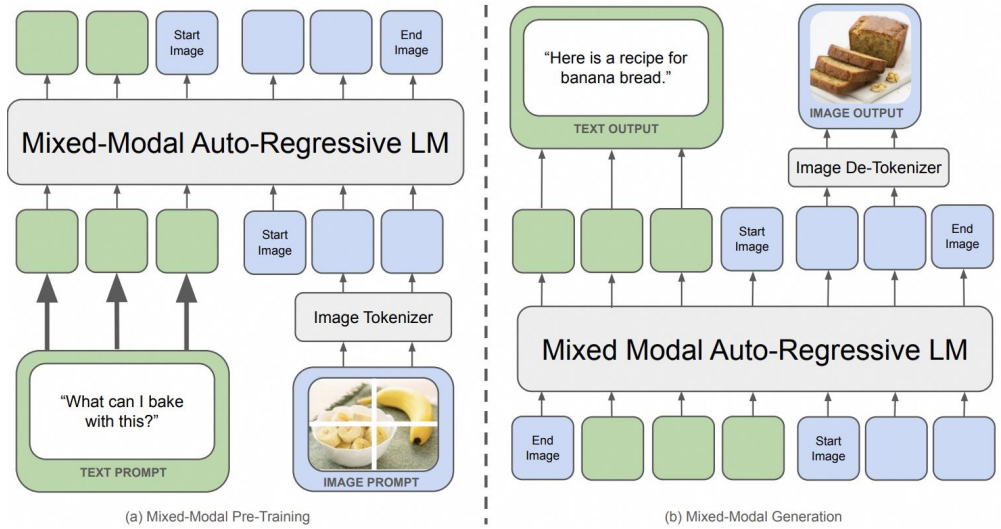
UniLIP: Adapting CLIP for Unified Multimodal Understanding, Generation and Editing

Hao Tang

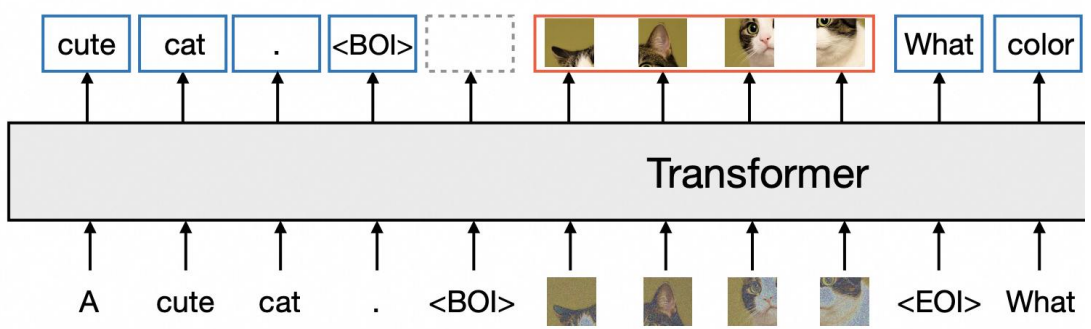
Motivation

- Why we need unification?
  - More tasks: image editing, interleave generation, in-context generation...
  - Better performance on both tasks: task synergy

- Naive unification leads to poor performance
  - Chameleon, Transfusion
  - VQVAE/VAE lack semantics



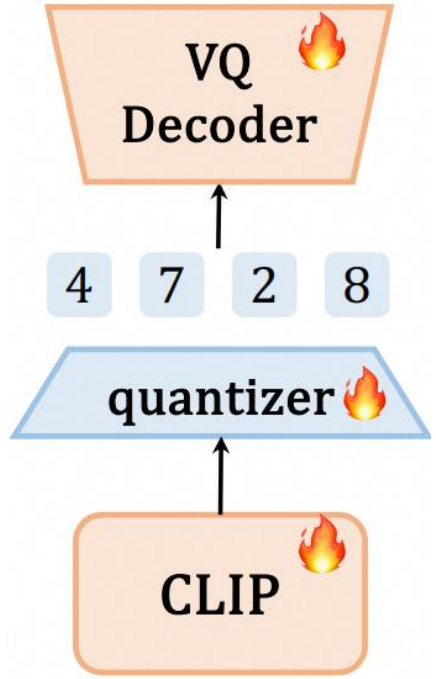
Chameleon

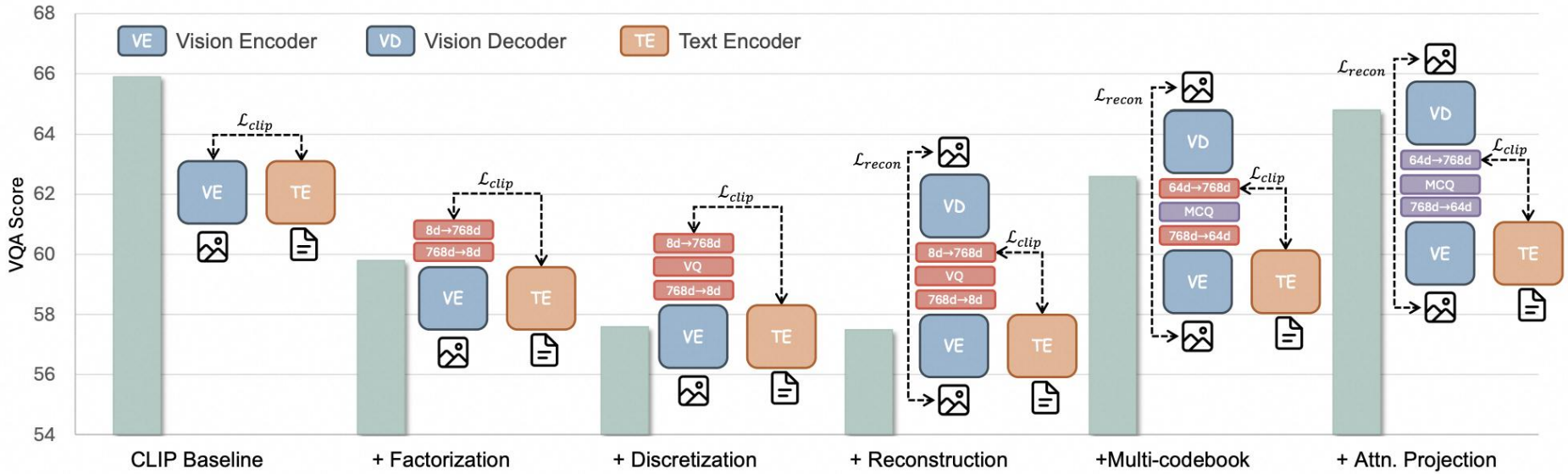


Transfusion

CLIP-based Unified Tokenizer

- Semantic Encoder + Vector Quantization
  - VILA-U, TokLIP, UniTok, TokenFlow, ILLUME, Tar
- Quantification limits the expressive power of representations
  - preformance degradation in understanding
- Best generation model is still based on diffusion/flow matching

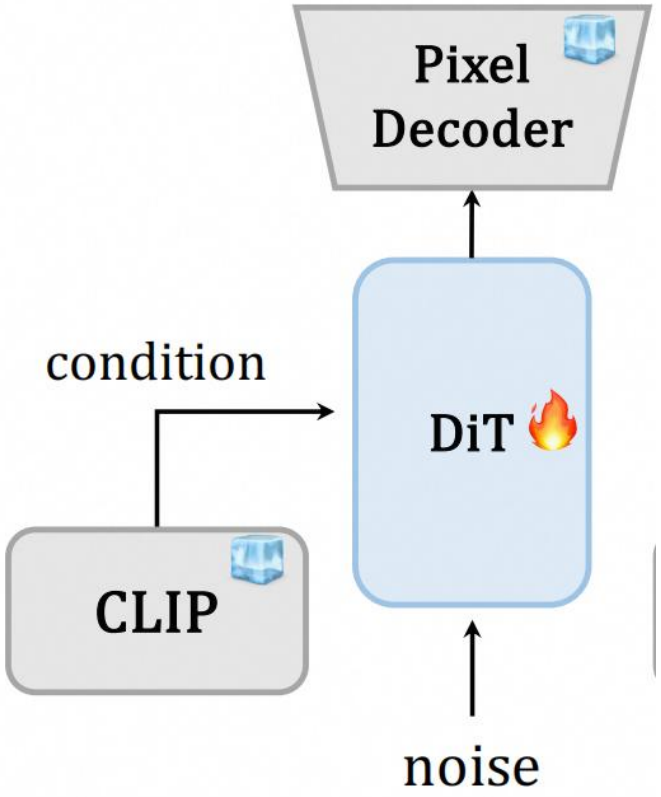




UniTok

Frozen CLIP + Diffusion Decoder

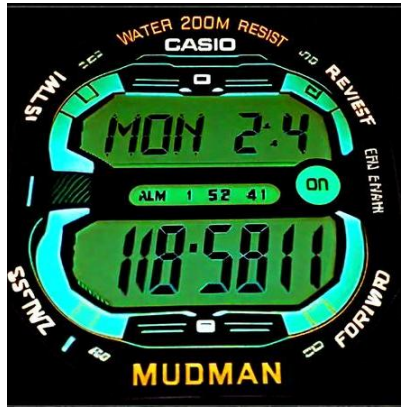
- CLIP as diffusion condition: Emu2, BLIP3o
- Frozen CLIP loses pixel details
  - incosistent reconstruction
  - hard to support editing
- Blip3o: two diffusion process
  - cond: query, noise -> CLIP features
  - cond: CLIP features, noise -> VAE features



Origin

Emu2

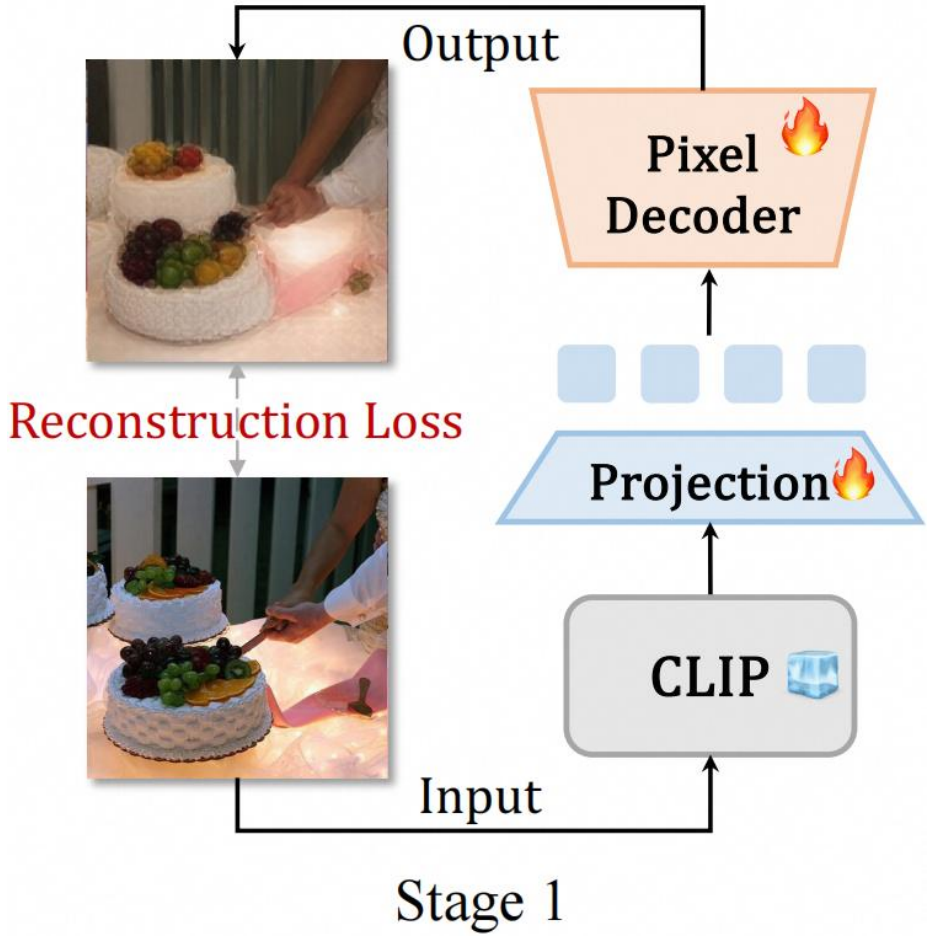
Origin



Emu2

How to adapt CLIP for **reconstruction without degradation?**

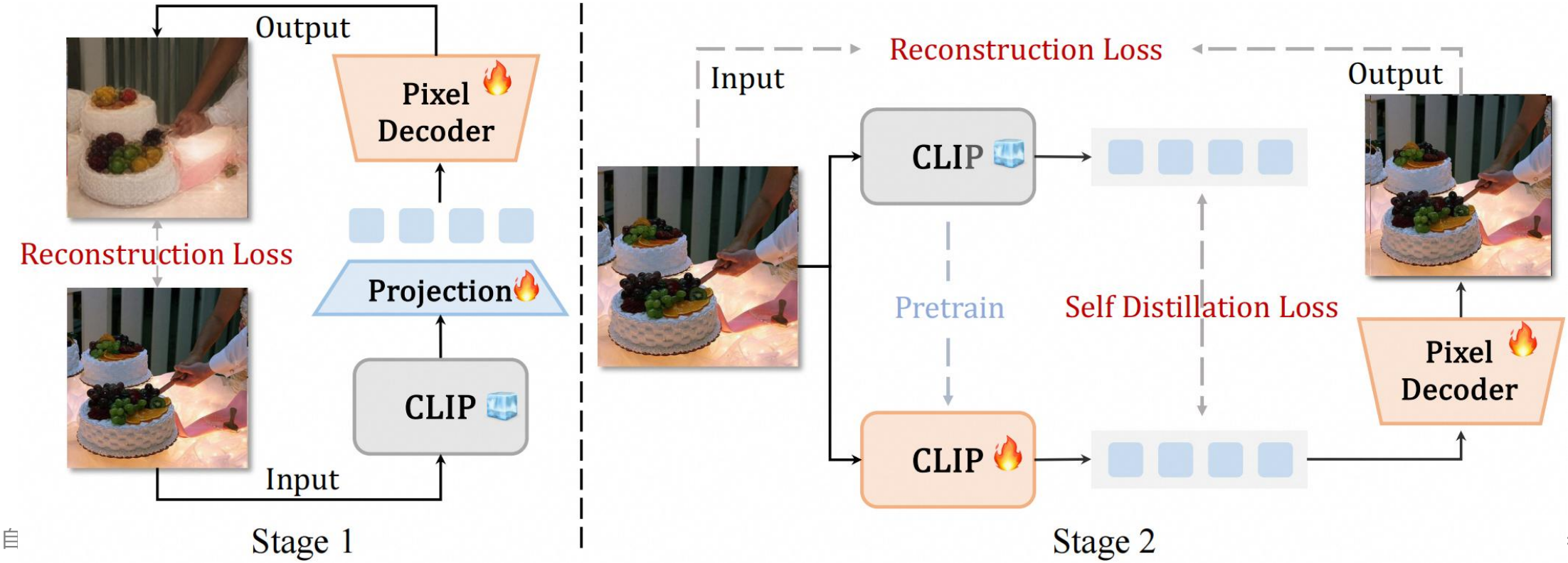
- Check reconstruction quality of frozen CLIP
- Blurry but keep main semantic content
- Information: CLIP >> VAE
  - CLIP dimension >= 768
  - VAE dimension: 4, 8, 16, 32
- Maybe a minor change is enough for recon?
- Und performance maybe robust to minor change?



How to minimize parameter updates?

- Stage 1: train pixel decoder only
  - Adapt to the original clip features as much as possible

- Stage2: self distillation, restrict feature updates
  - mse loss is more effective than cosine, kl



Reconstruction and Understanding

北京大学 PEKING UNIVERSITY

- Better reconstruction than previous CLIP-based unfied tokenizer

Model	Res.	ratio	rFID ↓	PSNR↑	SSIM↑
VILA-U (Wu et al., 2024)	256	16	1.80	-	-
Tokenflow (Qu et al., 2024)	256	16	1.37	21.41	0.687
DualViTok (Huang et al., 2025)	256	16	1.37	22.53	0.741
UniLIP	256	32	**0.79**	**22.99**	**0.747**
Emu2 (Sun et al., 2024b)	448	14	3.27	13.49	0.423
UniLIP	448	32	**0.31**	**24.62**	**0.788**

- Keep original Und. performance

Model	Reconstruction			Understanding				
	rFID↓	PSNR↑	SSIM↑	MME-P↑	MMBench↑	MMVP↑	AI2D↑	TextVQA↑
Frozen CLIP (Zhu et al., 2025)	6.14	16.26	0.572	1492	**72.6**	67.3	69.4	74.1
UniLIP	**0.31**	**24.62**	**0.788**	**1499**	**72.6**	**68.7**	**70.7**	**74.7**

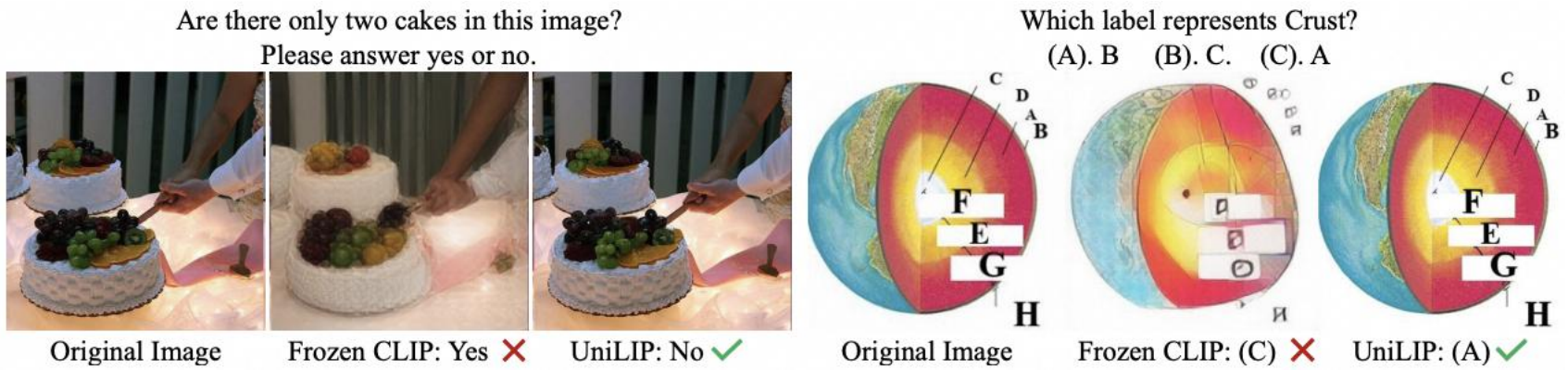
Reconstruction helps Understanding?

北京大学 PEKING UNIVERSITY

- Better performance on MME-P, MMVP, AI2D

Model	Reconstruction			Understanding				
	rFID↓	PSNR↑	SSIM↑	MME-P↑	MMBench↑	MMVP↑	AI2D↑	TextVQA↑
Frozen CLIP (Zhu et al., 2025)	6.14	16.26	0.572	1492	**72.6**	67.3	69.4	74.1
UniLIP	**0.31**	**24.62**	**0.788**	**1499**	**72.6**	**68.7**	**70.7**	**74.7**

- UniLIP introduces more pixel details

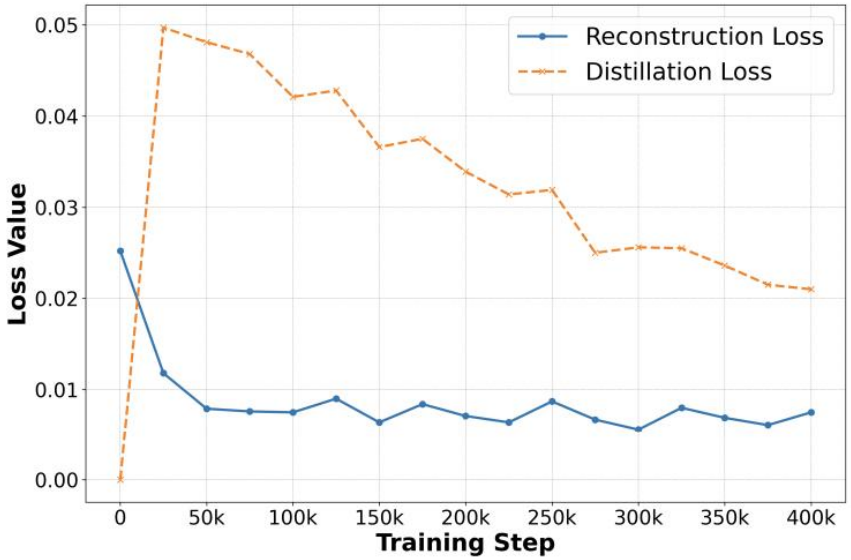


Ablation of Reconstruction Training

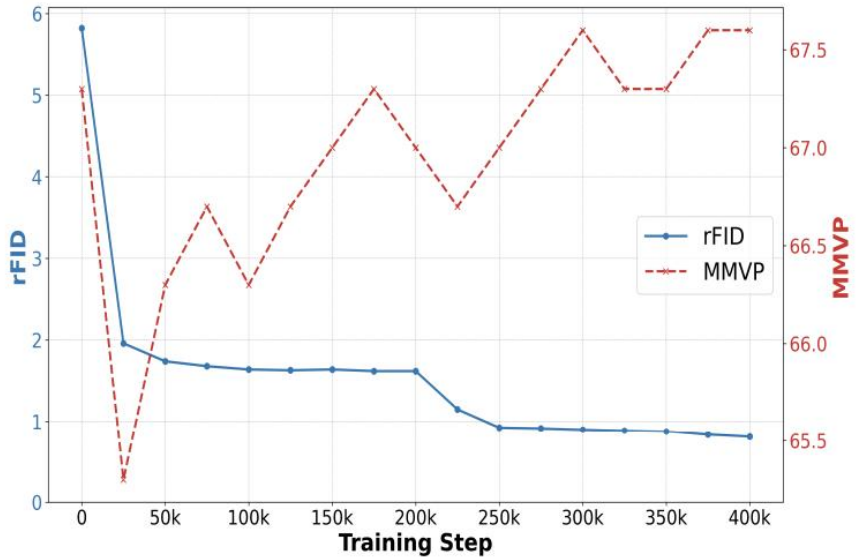
- Training strategies
  - Lr Decay: lr of CLIP is 1e-5, others 1e-4

Two Stage	Self-Distillation	Lr Decay	Reconstruction rFID↓	PSNR↑	SSIM↑	Understanding MME-P↑	MMBench↑	MMVP↑	AI2D↑	TextVQA↑
✗	✗	✗	0.43	26.01	0.819	124	0	47.3	27.9	0
✓	✗	✓	0.29	25.28	0.804	709	18.4	50.0	50.0	7.5
✓	✓	✗	0.28	25.11	0.801	1466	71.4	66.7	68.3	67.3
✗	✓	✓	0.35	24.61	0.782	1478	72.0	67.3	69.5	74.0
✓	✓	✓	0.31	24.62	0.788	1499	72.6	68.7	70.7	74.7

- Training loss



(a) Reconstruction and distillation loss curve.



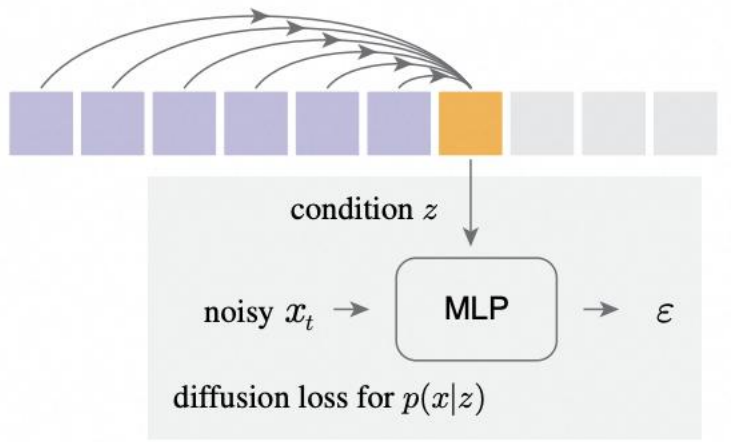
(b) Recon. (rFID) and und. (MMVP) performance.


Original
UniLIP
Frozen CLIP
Emu2
TokenFlow
DC-AE

How to do generation based on UniLIP?

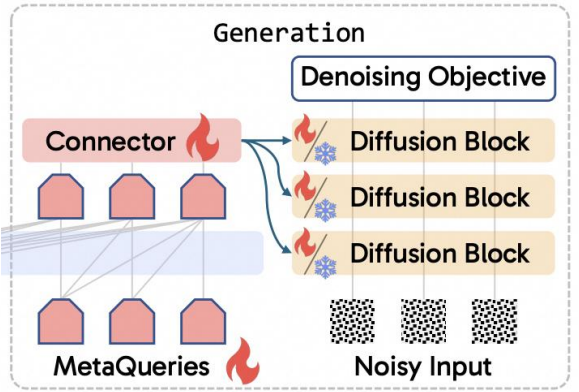
- UniLIP outputs continuous features

- Continuous + Autoregressive: MAR
  - need to train LLM: lose understanding
  - not truely autoregressive: more like MaskGIT
  - not SOTA performance



MAR

- Another way: MetaQuery
  - frozen LLM: keep Und. performance, low cost
  - better generation performance



MetaQueries

MAR	rand	causal	1	CrossEnt	16.22	81.3	4.36	222.7
				Diff Loss	**13.07**	**91.4**	**4.07**	**232.4**
MAR	rand	bidirect	1	CrossEnt	8.75	149.6	3.50	280.9
				Diff Loss	**3.43**	**203.1**	**1.84**	**292.7**

How to do editing based on UniLIP

- MetaQuery-like models seldom extend to image editing
  - No quantitative results

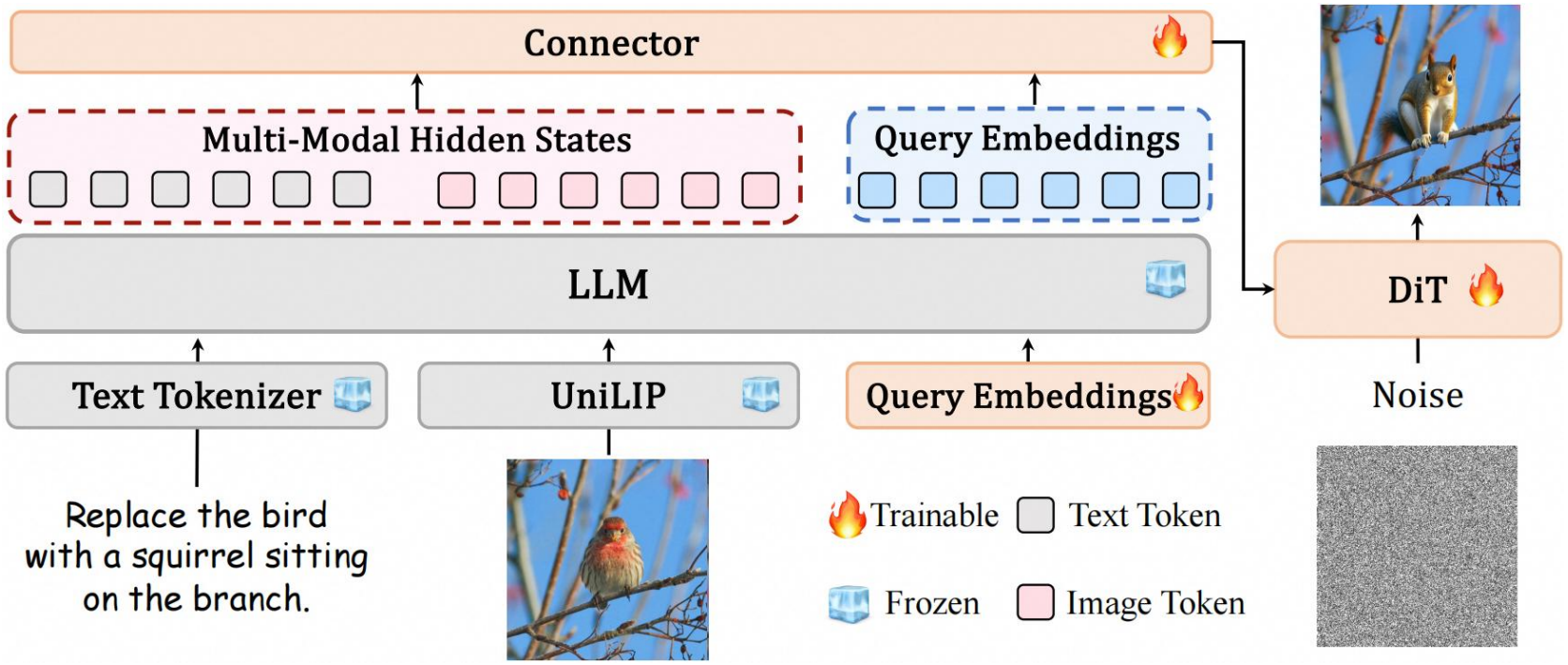
- Tokenizer: Frozen CLIP cannot provide sufficient details, e.g, BLIP3o
  - Diffusion cannot keep consistency

- UniLIP can provied both semantics and pixel details
  - semantics: align with text prompt, easier to identify where and what to edit
  - pixel details: essential to keep consistency

How to do editing based on UniLIP

- Architecture: query forms information bottleneck
- Dual-condition: multimodal context + metaqueries

Multimodal Hidden States	Query Embedding	WISE	ImgEdit
✓	✗	0.47	3.62
✗	✓	0.52	3.38
✓	✓	0.56	3.81



MLLM	Connector	DiT	Pixel Decoder	MetaQuery
InternVL3-1B	6 Layer LLM	SANA-0.6B	DCAE	256

Training

- Generation: BLIP3o-Pretrain (~35M), BLIP3o-60k, Janus4o (~45k)
- Editing: GPT-Image-Edit (1.5M), Janus4o (~45k)
- Batch size 512, Lr 1e-4, Cosine decay

	Stage1	Stage2	Stage3
Trainable	connector	connector + DiT	connector + DiT
Data	BLIP3o–Pretrain	BLIP3o–Pretrain, ImgEdit	BLIP3o–60k Janus4o
Iteration	50k	200k	20k

Understanding

北京大学 PEKING UNIVERSITY

Model	# LLM Params	MME-P	MMB	MMMU	MM-Vet	SEED	AI2D	MMVP
Und. Only								
LLaVA-OV (Li et al., 2024a)	1B	1238	52.1	31.4	29.1	65.5	57.1	-
InternVL2.5 (Chen et al., 2024c)	1B	-	70.7	41.2	48.8	-	69.3	31.3
InternVL3 (Zhu et al., 2025)	1B	1492	72.6	43.4	59.5	71.1	69.4	67.3
InternVL2.5 (Chen et al., 2024c)	1.8B	-	74.7	43.6	60.8	-	74.9	-
InternVL3 (Zhu et al., 2025)	1.8B	1633	80.6	48.2	62.2	75.0	78.5	72.7
Qwen2.5-VL (Bai et al., 2025)	3B	-	79.1	53.1	61.8	-	81.6	-
Emu3-Chat (Wang et al., 2024c)	8B	1244	58.5	31.6	37.2	68.2	70.0	36.6
Und. and Gen.								
Chameleon (Team, 2024)	7B	-	35.7	28.4	8.3	-	-	0.0
VILA-U (Wu et al., 2024)	7B	1336	66.6	32.2	27.7	56.3	-	22.0
MetaMorph (Tong et al., 2024b)	8B	-	75.2	41.8	-	-	-	48.3
SEED-X (Ge et al., 2024)	13B	1457	70.1	35.6	43.0	66.5	-	-
TokenFlow-B (Qu et al., 2024)	13B	1354	55.3	34.2	22.4	60.4	54.2	-
Show-O (Xie et al., 2024b)	1.3B	1097	-	26.7	-	-	-	-
ILLUME (Wang et al., 2024a)	7B	1445	75.1	38.2	37.0	-	71.4	-
Janus-Pro (Chen et al., 2025c)	7B	1567	79.2	41.0	50.0	72.1	-	-
Harmon (Wu et al., 2025c)	1.5B	1155	65.5	38.9	-	67.1	-	
MetaQuery-B (Pan et al., 2025)	1B	1238	58.5	31.4	29.1	66.6	-	-
BAGEL (Deng et al., 2025)	3B	1610	79.2	43.2	48.2	-	-	54.7
BLIP3-o (Chen et al., 2025a)	4B	1528	78.6	46.6	60.1	73.8	-	-
TokLIP (Lin et al., 2025b)	7B	1410	-	42.1	-	65.2	-	-
Tar (Han et al., 2025)	7B	1571	74.4	39.0		73.0	-	-
UniLIP-1B	1B	1499	72.6	43.3	59.4	71.0	70.7	68.7
UniLIP-3B	1.8B	**1636**	**80.7**	**48.7**	**62.2**	**75.0**	**78.6**	**73.0**

Generation

北京大学 PEKING UNIVERSITY

- Best Gen. performance in similar-scale models
  - 1B: GenEval 0.87, WISE: 0.56
  - 3B: GenEval 0.90, WISE: 0.62

Model	# Params	GenEval			WISE		
		Counting	Position	Overall	Cultural	Biology	Overall
Gen. Only							
SDXL (Podell et al., 2023)	2.6B	0.39	0.15	0.55	0.43	0.44	0.43
FLUX.1-dev (Labs, 2023)	12B	0.75	0.68	0.82	0.48	0.42	0.50
PixArt-α (Chen et al., 2024a)	0.6B	0.44	0.08	0.48	0.45	0.49	0.47
Emu3-Gen (Wang et al., 2024c)	8B	0.34	0.17	0.54	0.34	0.41	0.39
SD3-Medium (Esser et al., 2024)	2B	0.72	0.33	0.74	0.42	0.39	0.42
Sana-1.6B (Xie et al., 2024a)	1.6B	0.62	0.21	0.66	-	-	
Und. and Gen.							
VILA-U (Wu et al., 2024)	7B	-	-	-	0.26	0.35	0.31
TokenFlow-XL (Qu et al., 2024)	14B	0.41	0.16	0.55	-	-	-
ILLUME+ (Huang et al., 2025)	3B + 2.6B	0.62	0.42	0.72	-	-	-
Janus-Pro (Chen et al., 2025c)	7B	0.59	0.79	0.80	0.30	0.36	0.35
MetaQuery-B (Pan et al., 2025)	1B + 1.6B	-	-	0.74	0.44	0.41	0.46
MetaQuery-XL (Pan et al., 2025)	7B + 1.6B	-	-	0.80	0.56	0.49	0.55
Harmon (Wu et al., 2025c)	1.5B + 1B	0.66	0.74	0.76	0.38	0.37	0.41
BLIP3-o-4B (Chen et al., 2025a)	3B + 1.4B	-	-	0.81	-	-	0.50
BLIP3-o-8B (Chen et al., 2025a)	7B + 1.4B	-	-	0.84	-	-	0.62
BAGEL (Deng et al., 2025)	7B + 7B	0.81	0.64	0.82	0.44	0.44	0.52
OpenUni-B (Wu et al., 2025b)	1B + 0.6B	0.74	0.77	0.84	0.37	0.39	0.43
OpenUni-L (Wu et al., 2025b)	2B + 1.6B	0.77	0.75	0.85	0.51	0.48	0.52
Tar (Han et al., 2025)	7B	0.83	0.80	0.84	-	-	-
UniLIP-1B	1B + 0.6B	0.83	0.83	0.88	0.54	0.50	0.56
UniLIP-3B	2B + 1.6B	**0.84**	**0.90**	**0.90**	**0.65**	**0.57**	**0.62**

Editing

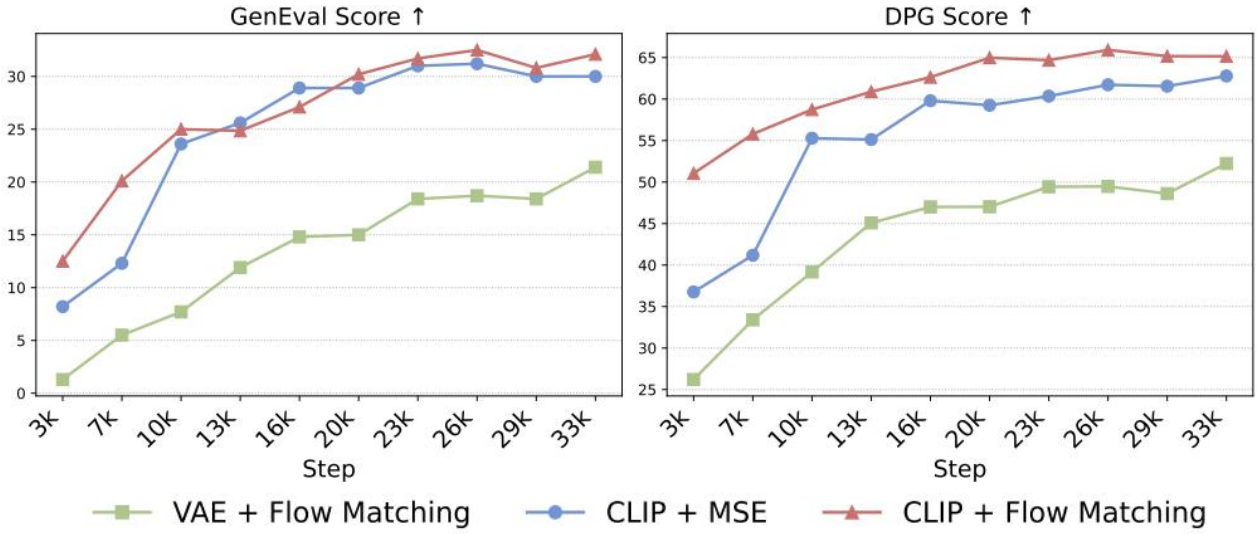
北京大学 PEKING UNIVERSITY

Model	# Params	Add	Adjust	Replace	Remove	Bkg.	Style	Overall
GPT-4o (OpenAI, 2025)	-	4.61	4.33	4.35	3.66	4.57	4.93	4.20
MagicBrush (Zhang et al., 2023)	0.9B	2.84	1.58	1.97	1.58	1.75	2.38	1.90
Instruct-P2P (Brooks et al., 2023)	0.9B	2.45	1.83	2.01	1.50	1.44	3.55	1.88
AnyEdit (Yu et al., 2025)	1.3B	3.18	2.95	2.47	2.23	2.24	2.85	2.45
UltraEdit (Zhao et al., 2024)	2.0B	3.44	2.81	2.96	1.45	2.83	3.76	2.70
OmniGen (Xiao et al., 2025)	3.8B	3.47	3.04	2.94	2.43	3.21	4.19	2.96
Step1X-Edit (Liu et al., 2025)	7B+12B	3.88	3.14	3.40	2.41	3.16	4.63	3.06
ICEdit (Zhang et al., 2025)	12B	3.58	3.39	3.15	2.93	3.08	3.84	3.05
BAGEL (Deng et al., 2025)	7B+7B	3.56	3.31	3.30	2.62	3.24	4.49	3.20
UniWorld-V1 (Lin et al., 2025a)	7B+12B	3.82	3.64	3.47	3.24	2.99	4.21	3.26
Janus-4o (Chen et al., 2025b)	7B	3.60	3.25	3.27	2.28	3.32	4.47	3.26
OmniGen2 (Wu et al., 2025a)	3B+4B	3.57	3.06	3.74	3.20	3.57	4.81	3.44
UniLIP-1B	1B+0.6B	4.11	3.58	4.30	3.97	4.00	**4.87**	3.81
UniLIP-3B	2B+1.6B	**4.11**	**3.81**	**4.48**	**4.20**	**4.11**	4.76	**3.94**

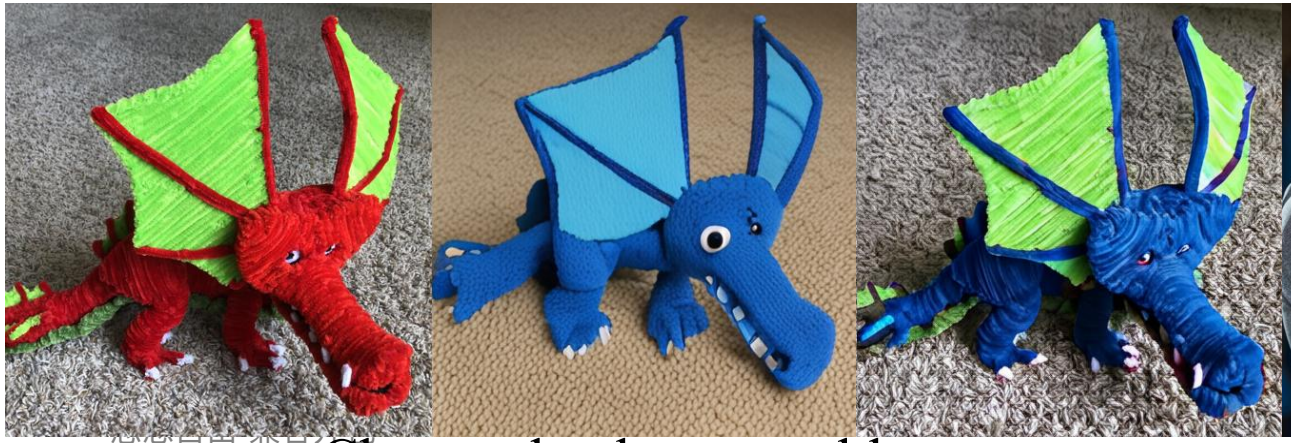
Ablation: Is UniLIP better than CLIP and VAE?

- Better edit fidelity and prompt alignment.

Reference	Target	GenEval	WISE	ImgEdit
CLIP	VAE	0.84	0.46	3.37
UniLIP	VAE	0.85	0.48	3.70
CLIP	UniLIP	0.88	0.53	3.42
UniLIP	UniLIP	0.88	0.56	3.81

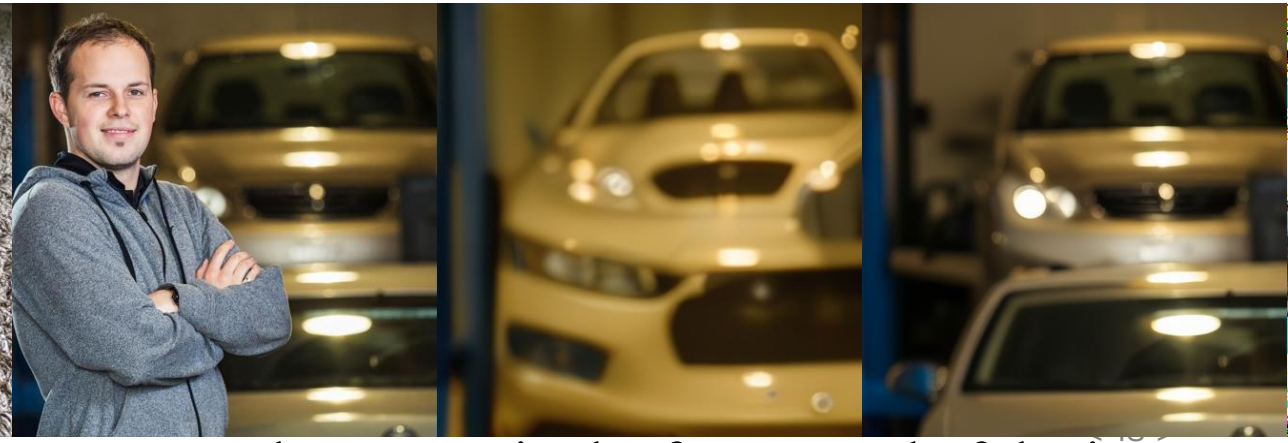


Reference CLIP UniLIP



Change the dragon to blue.

Reference CLIP UniLIP



Remove the person in the foreground of the image.

Visualization