

salesforce

BLIP3-o

A Family of Fully Open Unified Multimodal Architecture, Training and Dataset

Jiuhai Chen^{12*} Zhiyang Xu^{3*} Xichen Pan^{4*} Yushi Hu^{5*} Can Qin¹

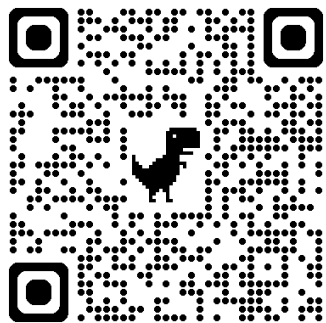
Tom Goldstein² Lifu Huang⁶ Tianyi Zhou² Saining Xie⁴

Silvio Savarese¹ Le Xue¹ Caiming Xiong¹ Ran Xu¹

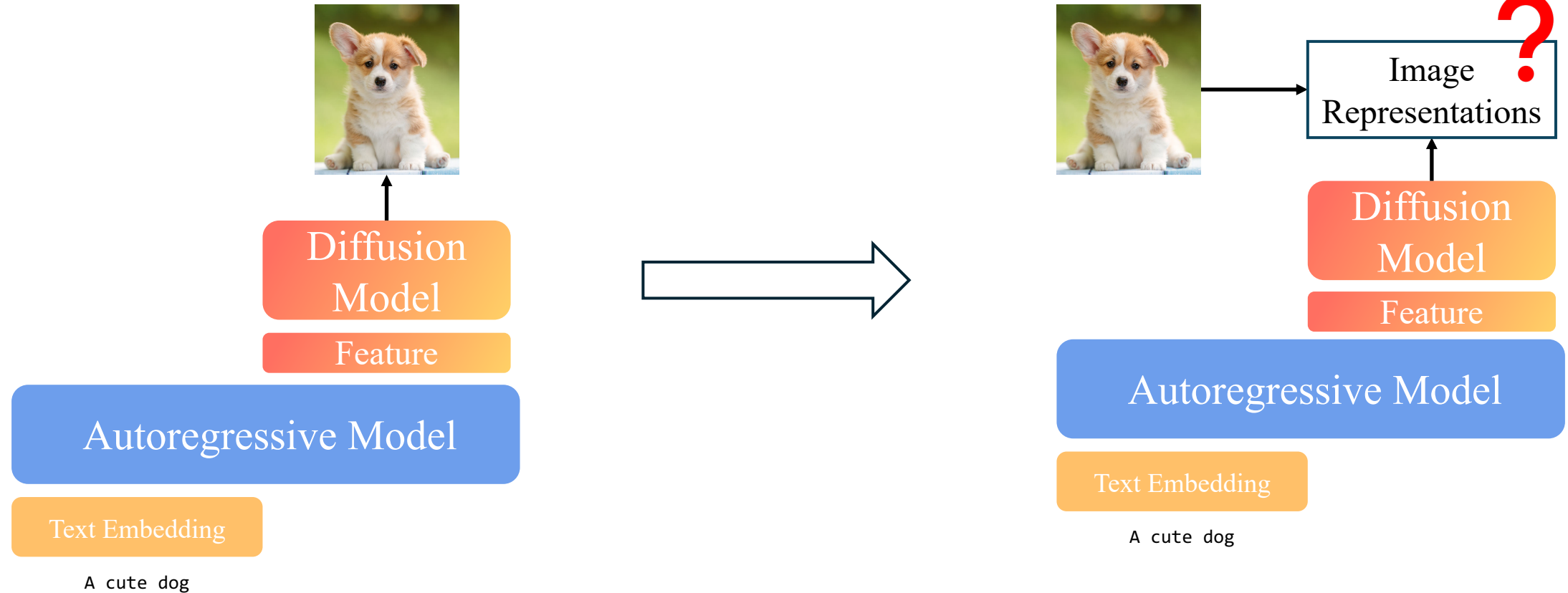
¹Salesforce Research ²University of Maryland ³Virginia Tech

⁴New York University ⁵University of Washington ⁶UC Davis

* Indicates equal contribution



Autoregressive + Diffusion



Encode real images into VAE latent or CLIP embedding ?

Image Encoder and Decoder


VAE: commonly use in diffusion model training

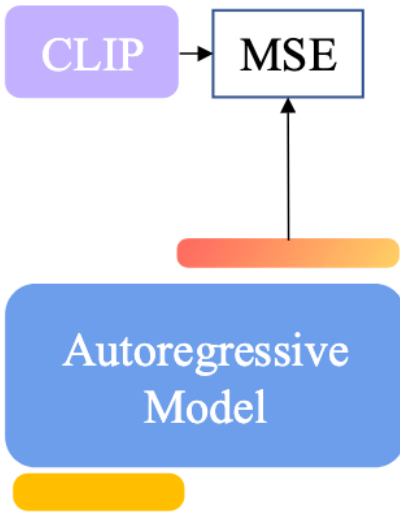
- Low-level pixel features and offer better reconstruction quality

CLIP + Diffusion: Encode images into continuous visual feature by CLIP and reconstructs them via a diffusion model.

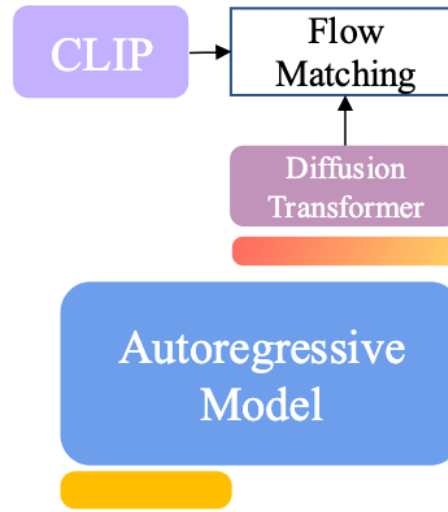
- Unify image understanding and generation in one semantic space

Design Choice

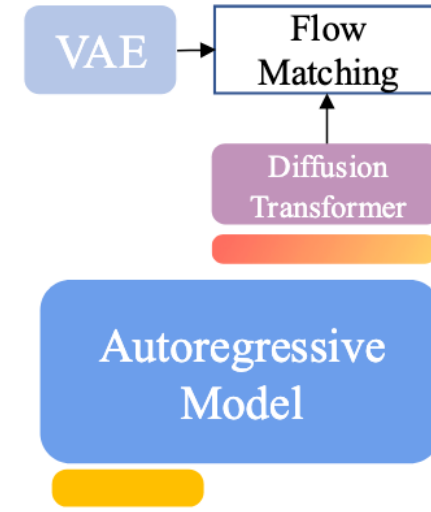
 Prompt  Visual Feature



(a) CLIP + MSE



(b) CLIP + Flow Matching

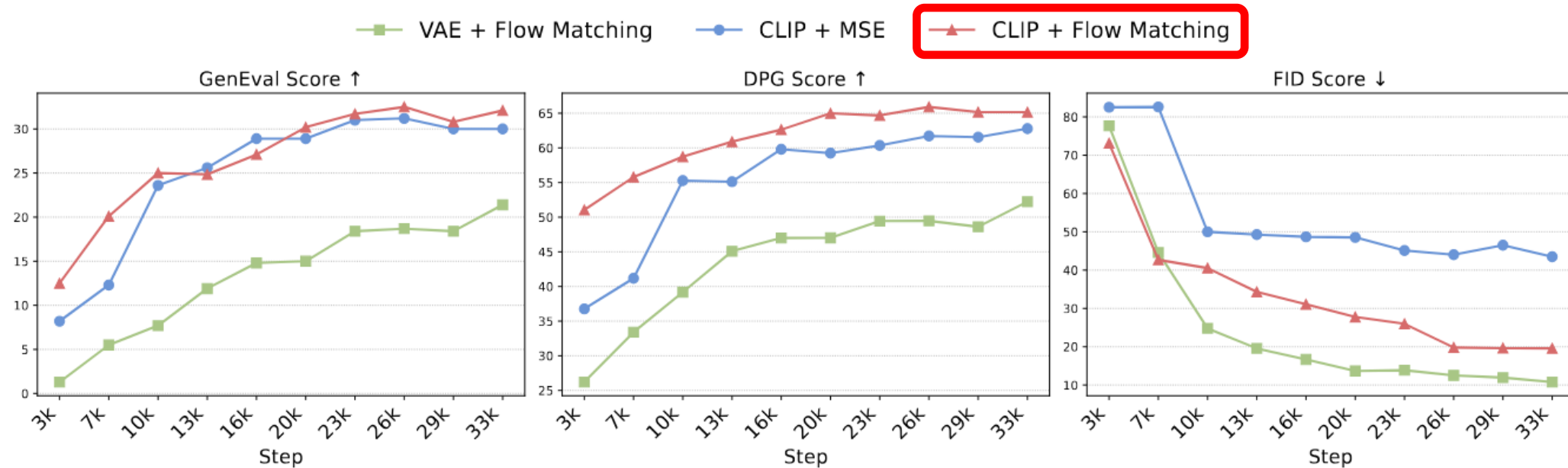


(c) VAE + Flow Matching

(a) AR generates CLIP feature, diffuse CLIP feature to real image (Emu2, SeedX, MetaMorph)

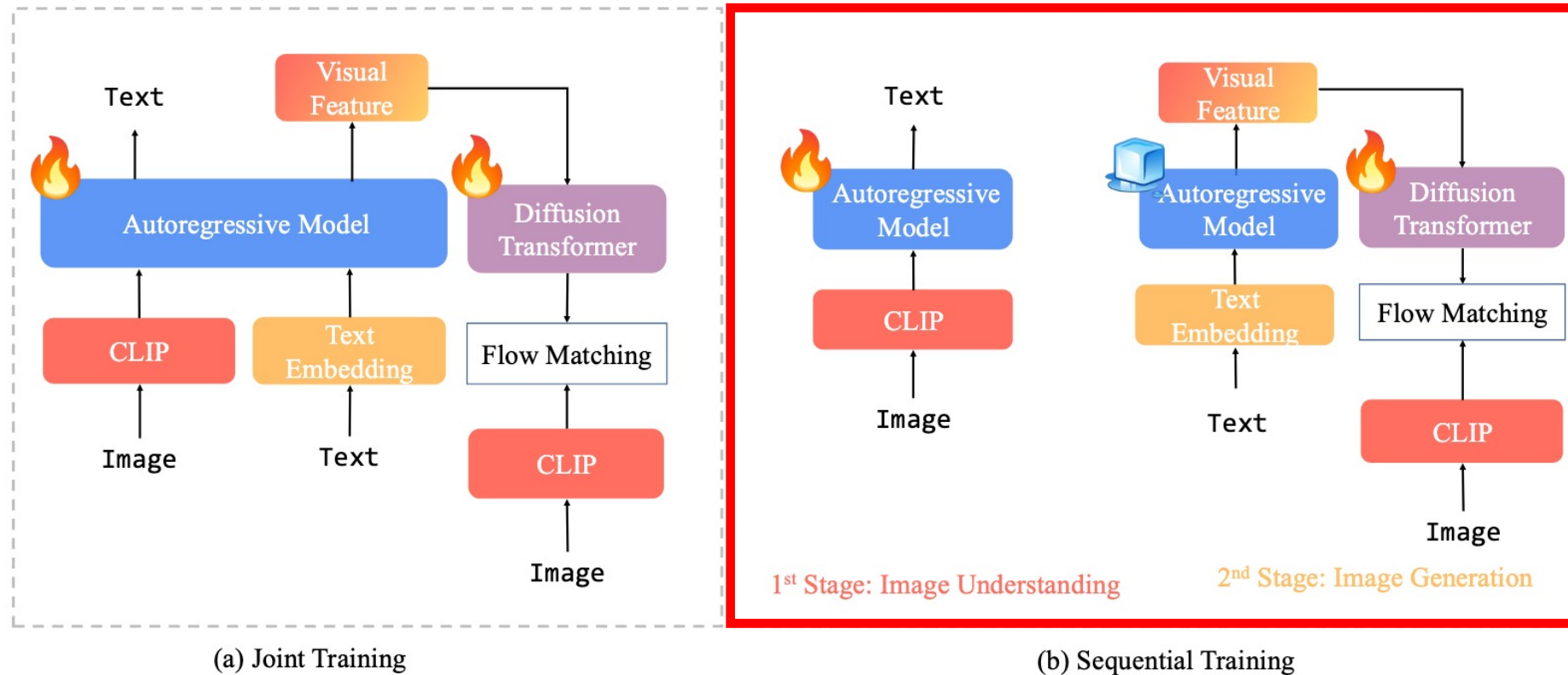
(b) AR generates visual feature, diffuse visual feature to CLIP feature, diffuse CLIP feature to real image

(c) AR generates visual feature, diffuse visual feature to VAE feature (MetaQuery)



- When integrating image generation into a unified model, autoregressive models more effectively learn the semantic-level features (CLIP) compared to pixel-level features (VAE).
- Adopting flow matching as the training objective better captures the underlying image distribution than MSE, resulting in greater sample diversity and enhanced visual quality.

Training Strategy



- Sequential training: freeze the autoregressive backbone and maintain the image understanding capability.
- Joint training: image understanding and generation benefit each other? NO if using diffusion loss

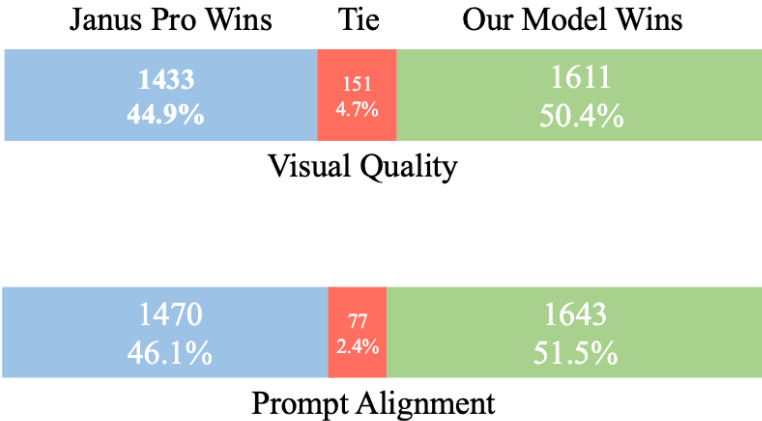
Results

Image Understanding

Model	VQA _{v2}	GQA	MMBench	SEED	MM-Vet	MME-P	MME-C	MMMU	RWQA	TEXTVQA
EMU2 Chat 34B	-	65.1	-	62.8	48.5	-	-	34.1	-	66.6
Chameleon 7B	-	-	19.8	27.2	8.3	202.7	-	22.4	39.0	0.0
Chameleon 34B	-	-	32.7	-	9.7	604.5	-	38.8	39.2	0.0
Seed-X 17B	63.4	49.1	70.1	66.5	43.0	1457.0	-	35.6	-	-
VILA-U 7B	79.4	60.8	66.6	57.1	33.5	1401.8	-	32.2	46.6	48.3
LLaVAFusion 16B	-	-	-	72.1	-	1603.7	367.8	41.7	60.0	-
Show-o 1.3B	69.4	58.0	-	-	-	1097.2	-	27.4	-	-
EMU3 8B	75.1	60.3	58.5	68.2	37.2	1243.8	266.1	31.6	57.4	64.7
MetaMorph 8B	-	-	75.2	71.8	-	-	-	41.8	58.3	60.5
TokenFlow-XL 14B	77.6	62.7	76.8	72.6	48.2	1551.1	371.1	43.2	56.6	77.6
Janus 1.3B	77.3	59.3	75.5	68.3	34.3	1338.0	-	30.5	-	-
Janus Pro 7B	-	62.0	79.2	72.1	50.0	1567.1	-	41.0	-	-
BLIP-3o 4B	75.9	60.0	78.6	73.8	60.1	1527.7	632.9	46.6	60.4	78.0
BLIP-3o 8B	83.1	60.5	83.5	77.5	66.6	1682.6	647.1	50.6	69.0	83.1

Image Generation

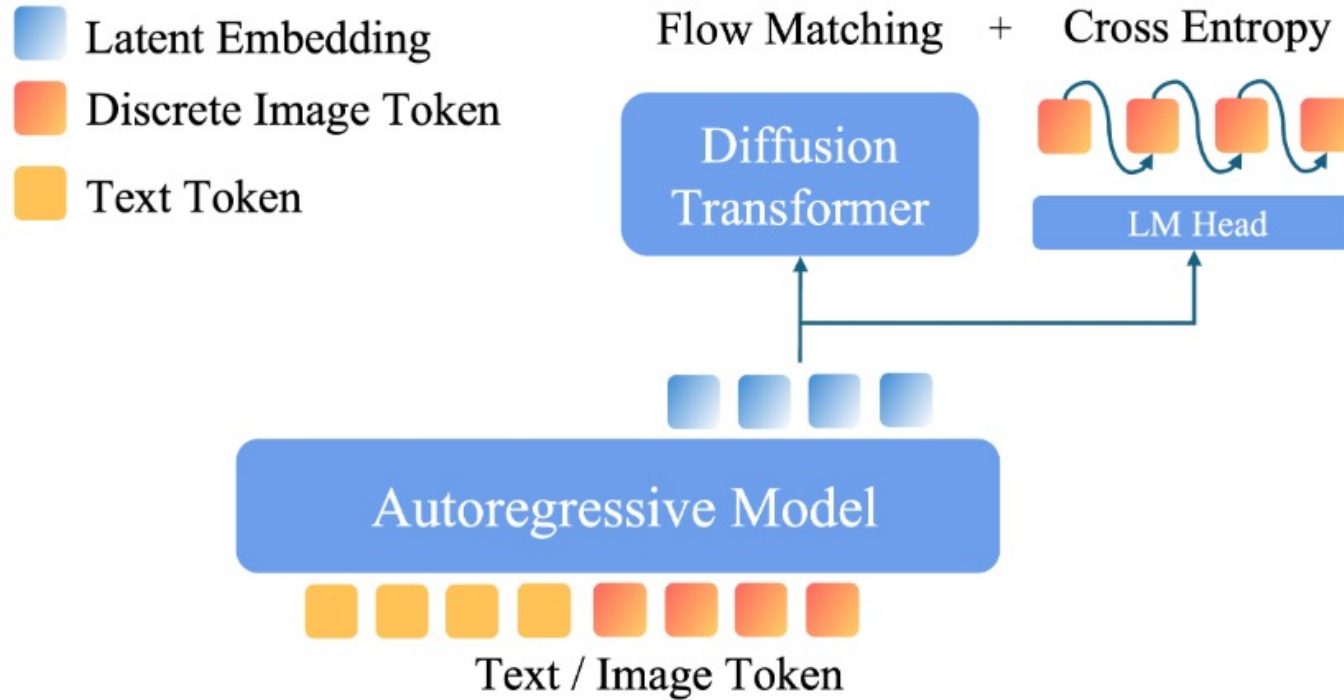
Model	GenEval	DPG-Bench	WISE
Chameleon 7B	0.39	-	-
Seed-X 17B	0.51	-	-
LLaVAFusion 16B	0.63	-	-
Show-o 1.3B	0.68	67.27	0.35
EMU3 8B	0.66	80.60	0.39
TokenFlow-XL 14B	0.63	73.38	0.50
Janus 1.3B	0.61	79.68	0.18
Janus Pro 7B	0.80	84.19	0.35
BLIP-3o 4B	0.81	79.36	0.50
BLIP-3o 8B	0.84	81.60	0.62



Conclusions

- **Diffuses CLIP features** for strong performance and training efficiency in image generation
- **Late Fusion** for image understanding and image generation
- Fully Open-Source model checkpoint, training code, **30 Million Pretraining Data** and **60k GPT-4o Distilled High Quality Instruction Tuning Data**

BLIP3o-Next



AR + Diffusion Architecture

Combines autoregressive and diffusion models for high-quality image generation.

Discrete Image Token Supervision

Uses discrete image tokens as additional supervision.

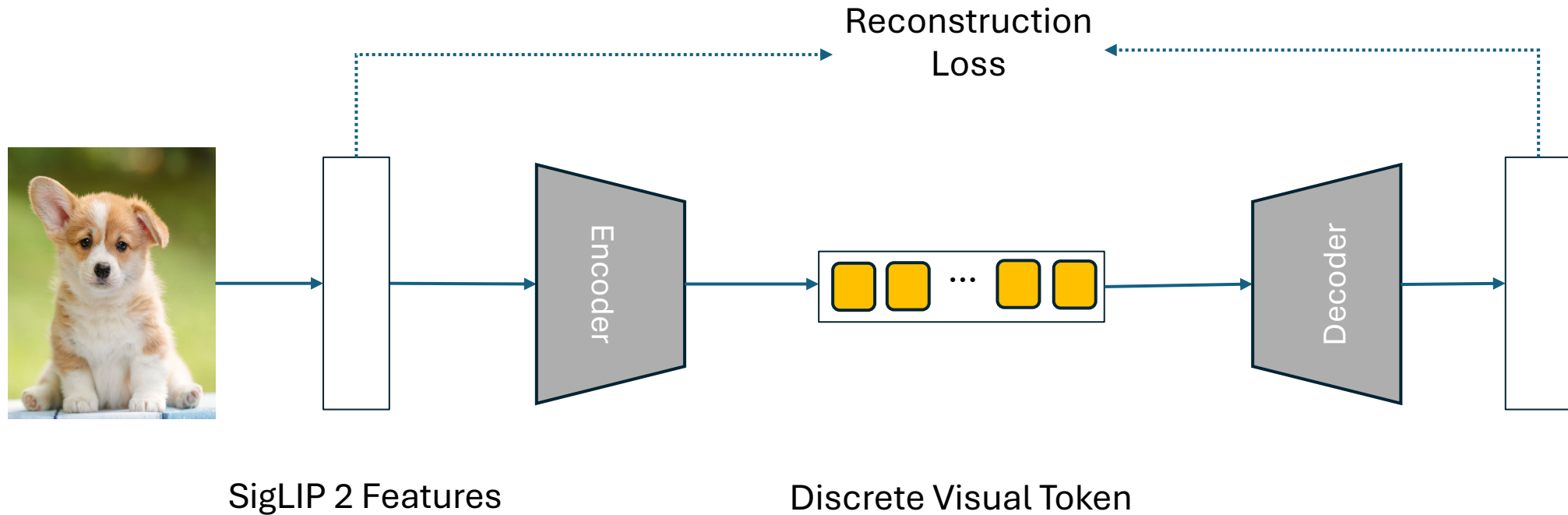
Reinforcement Learning

GRPO training enhances prompt alignment and text rendering accuracy.

Fully Open-Source

All training code, datasets, and models released for full reproducibility.

VQ-SigLIP2



Why VQ-SigLIP2 instead of VQ-VAE ?

- Using SigLIP2 for both image understanding and generation

Continuous V.S. Discrete

Diffusion

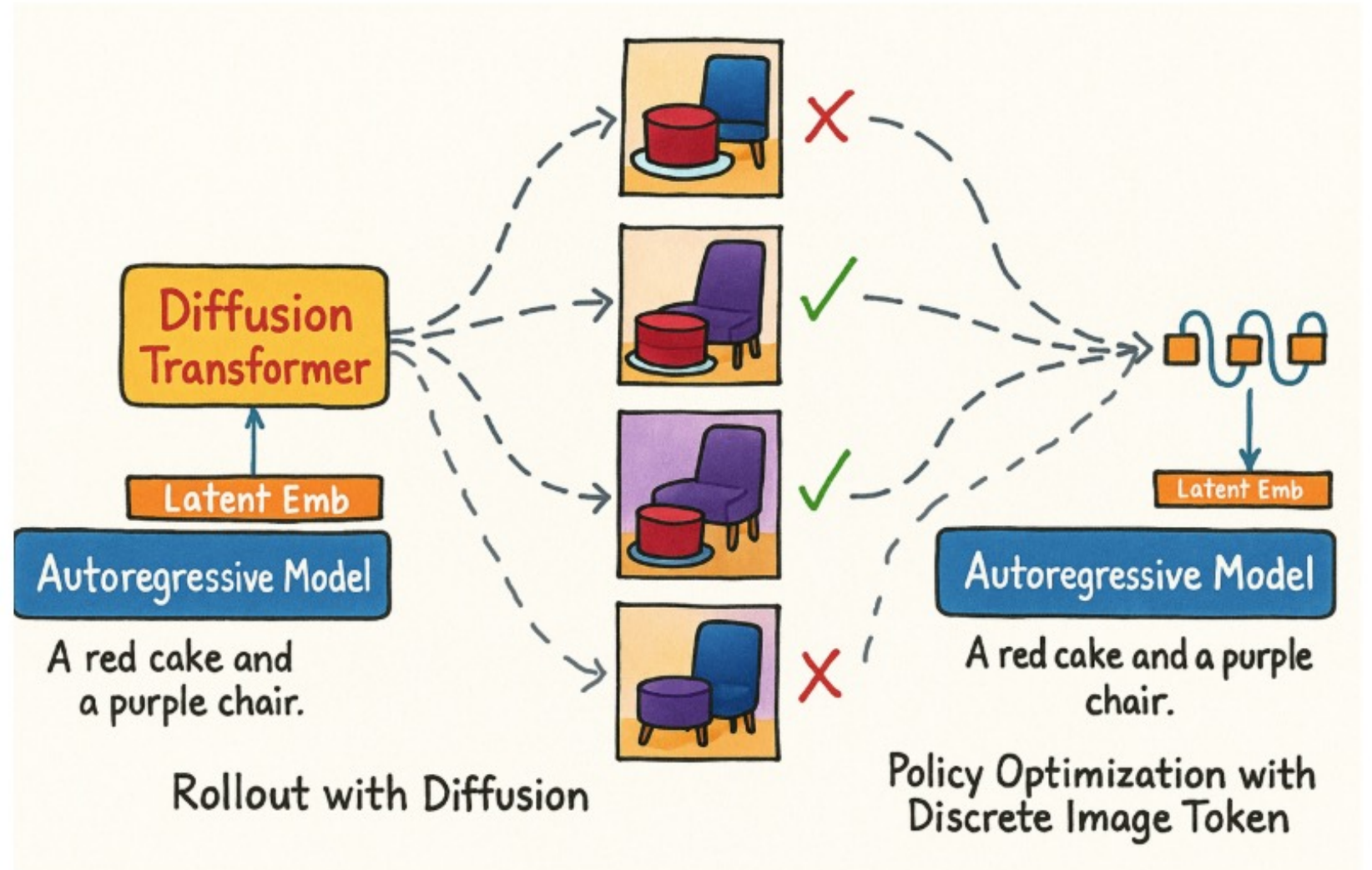
- high visual-fidelity images

Discrete

- Compatible with autoregressive model
- Excel at tasks requiring spatial structures

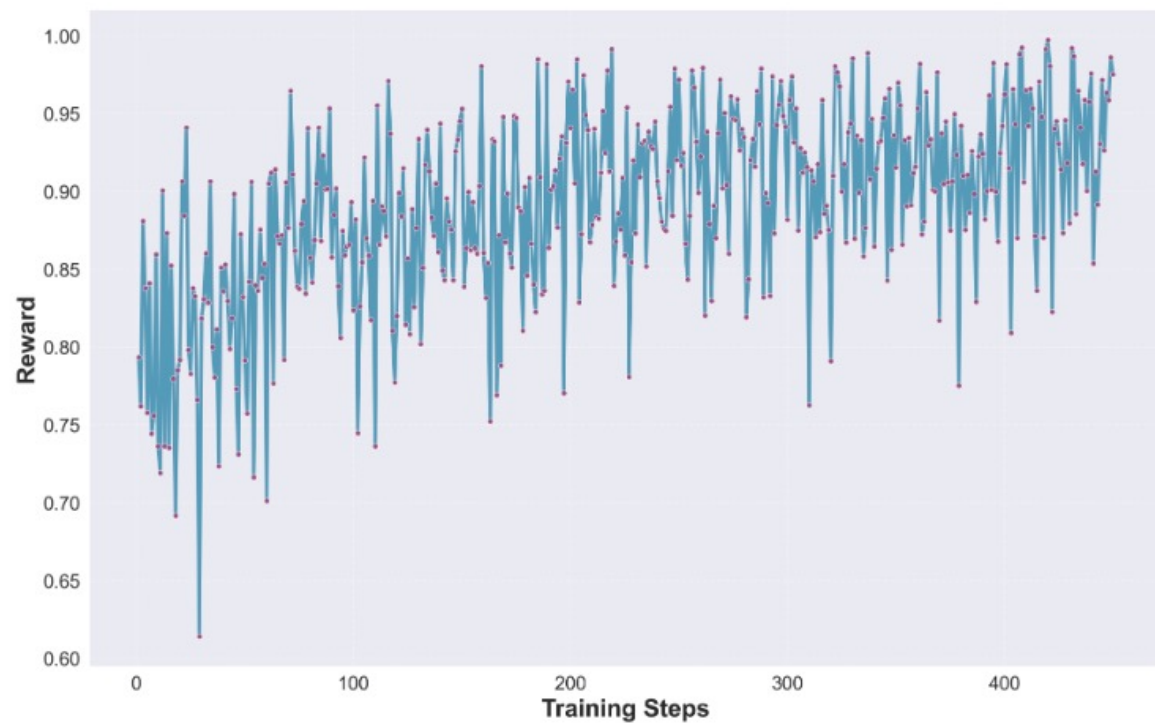
Multimodal RL

- Text Rendering
- Prompt Alignment



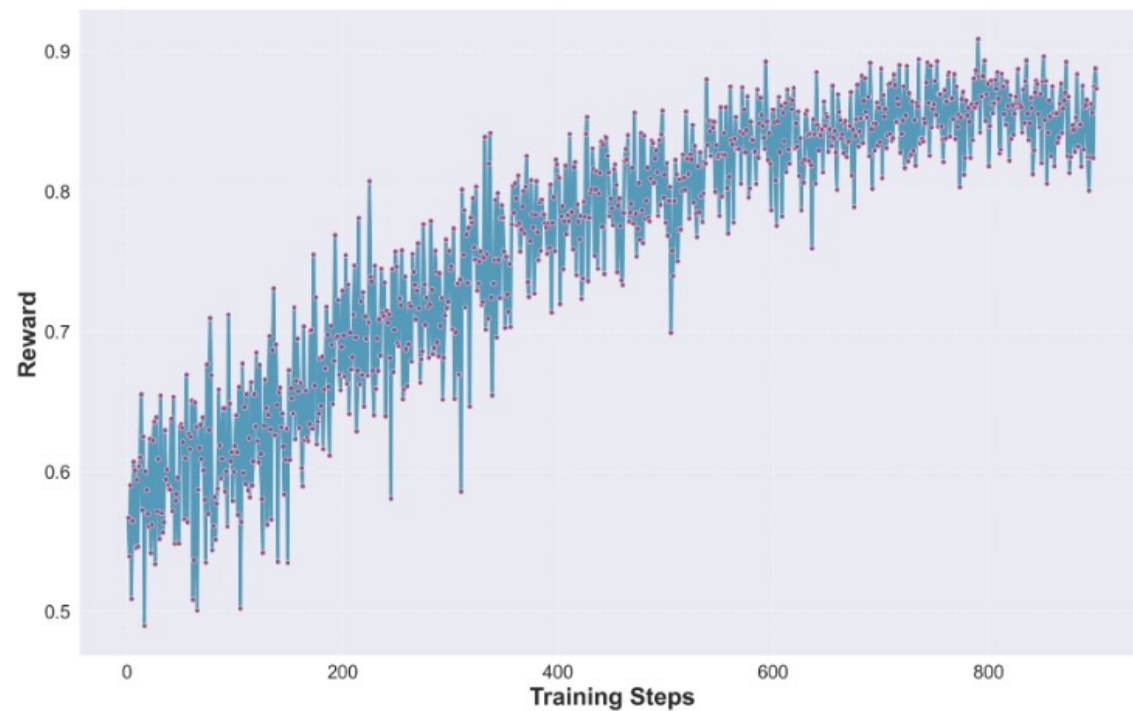
Multimodal RL

GRPO Training Reward



Prompt Alignment

OCR GRPO Training Reward



Text Rendering

Multimodal RL

Before
GRPO



After
GRPO



An astronaut
helmet, a red apple,
and a wooden horse

a teapot, a
skateboard, and a
goldfish in the air

a plastic flamingo,
a chess piece, and a
bread on a marble
staircase

a porcelain cat, a
green apple, and a
polka-dotted soccer
ball on a gym floor

Geneval

Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall
FLUX.1-dev	0.98	0.93	0.75	0.93	0.68	0.65	0.82
Metaqueries XL	-	-	-	-	-	-	0.80 [†]
BAGEL	0.98	0.95	0.84	0.95	0.78	0.77	0.88 [†]
BLIP3o	-	-	-	-	-	-	0.84
BLIP3o-NEXT (3B)	0.99	0.95	0.88	0.90	0.92	0.79	0.91

+ denotes testing with long caption

Multimodal RL

Before
GRPO



After
GRPO



A road sign
warning "Next
Gas 100 Miles".

A supermarket
scene with a sign
that reads "Frozen
Goods Section"

A bustling auto
repair shop with the
sign "Oil Change
Special 29"

A stamp featuring
"Year of the Tiger
2022", with golden
tiger designs and
Chinese motifs

Conclusions

- Autoregressive or Diffusion ? ***Both !***
Combines the strengths of both approaches
- RL beyond text
Expands the RL search space from pure language to rich multimodal outputs.

Thanks for your attention !