Grounded Image Generation & Editing

Week 5 | Sep 24, Thursday

Seminal Paper: ControlNet — Zhang et al. (2023)

Frontier Paper: Skywork UniPic 2.0 — Wei et al. (2025)

Group members:

Yuan Qiu, Yanjie Tong, Zhangding Liu, Jingyuan Zhang

Georgia Tech

Outline

Part 1: ControlNet (Seminal Paper)

01. Problem Definition
Limitations of T2I & need for precise spatial control

02. Related Work
Evolution from pix2pix & LDM to spatial control

03. Methodology
Locked/trainable blocks & zero-initialized 1x1 convs

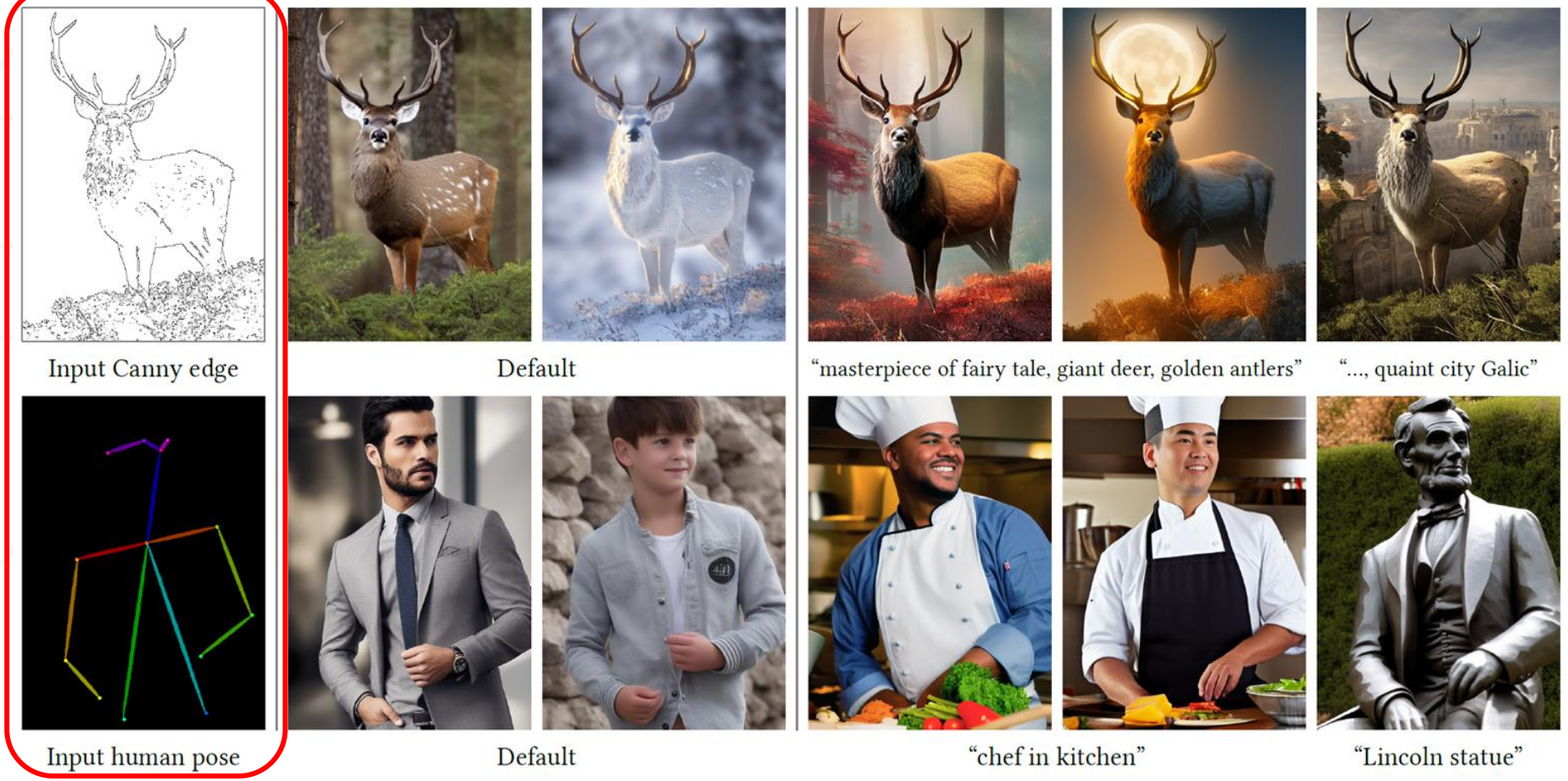
Part 2: Skywork UniPic 2.0 (Frontier Paper)

04. Unified Generation & Editing
Advancements in multimodal grounded control

05. Experiments & Results
Qualitative & quantitative benchmarks evaluation

06. Conclusion
Summary of contributions & open directions

Problem (Precise Spatial Control)



Limitations of text-to-image models:

Generating an image that accurately matches our mental imagery often requires *numerous trial-and-error cycles of editing a prompt, inspecting the resulting images and then re-editing the prompt*.

Can we enable finer grained spatial control by letting users provide additional images that directly specify their desired image composition?

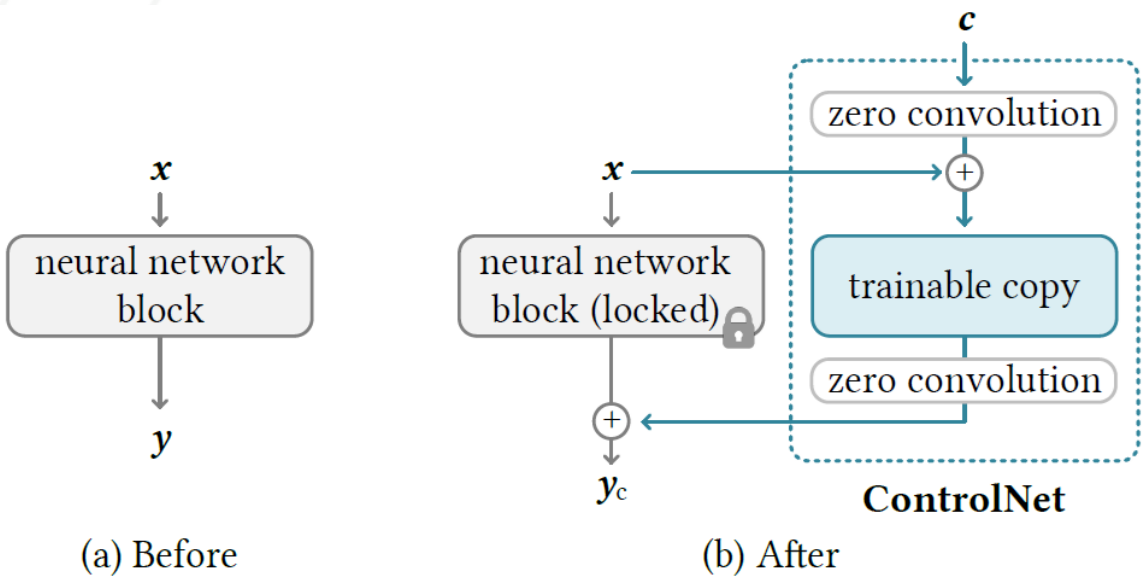
Georgia Tech

Related Work

Work	Main contribution	What remained to be solved
pix2pix (2017) [4], PITI (2022) [6]	Images can condition image-to-image generation	Without a large pretrained text-to-image backbone
Latent diffusion (2022) [3]	A strong, efficient text-to-image backbone	Limited control over precise layout and pose
Prompt-to-Prompt (2022) [5]	Cross attention can guide text-based editing	No direct input for general spatial maps

ControlNet: Learn spatial controls for a frozen, pretrained text-to-image diffusion model.

Method



(a) Before (b) After

Freeze the original block

Retain its pretrained parameters.

Clone a trainable block

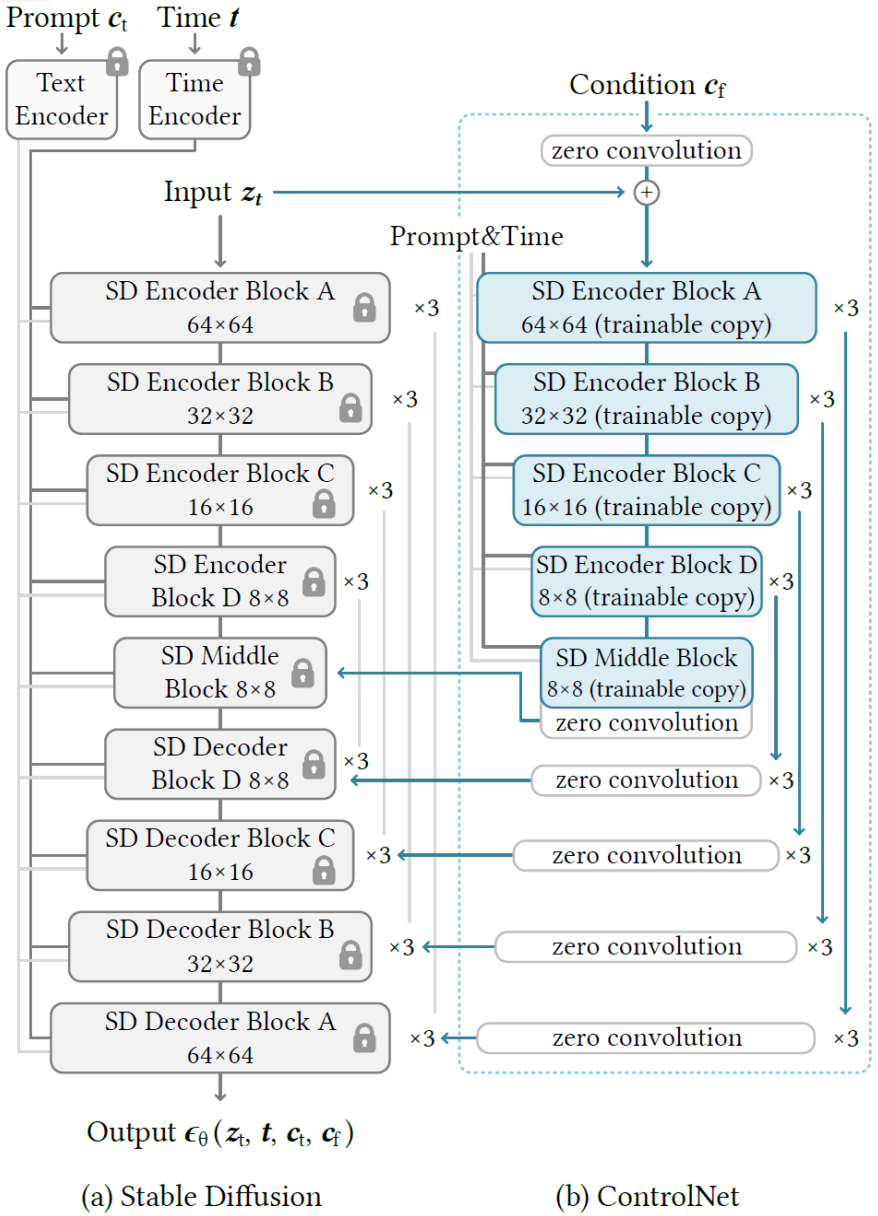
Reuse features to interpret the condition.

Zero-initialized 1 by 1 convolution

Give $y_c = F(x)$ at initialization.

$$\boldsymbol{y}_{\mathrm{c}} = \mathcal{F}(\boldsymbol{x}; \Theta) + \mathcal{Z}(\mathcal{F}(\boldsymbol{x} + \mathcal{Z}(\boldsymbol{c}; \Theta_{\mathrm{z1}}); \Theta_{\mathrm{c}}); \Theta_{\mathrm{z2}})$$

Architecture



Backbone: stable Diffusion

U-Net: encoder + middle + skip-connected decoder

Encoder, Decoder = 3*4 blocks

Down/up-sampling convolution, ViTs, ResNet layer

Lock the parameters in grey blocks

Only train the ones in blue and zero convolution

Georgia Tech

Training

Loss function

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}, \epsilon \sim \mathcal{N}(0,1)} \left[\| \epsilon - \epsilon_\theta(\boldsymbol{z}_t, \boldsymbol{t}, \boldsymbol{c}_t, \boldsymbol{c}_\mathrm{f}) \|_2^2 \right]$$

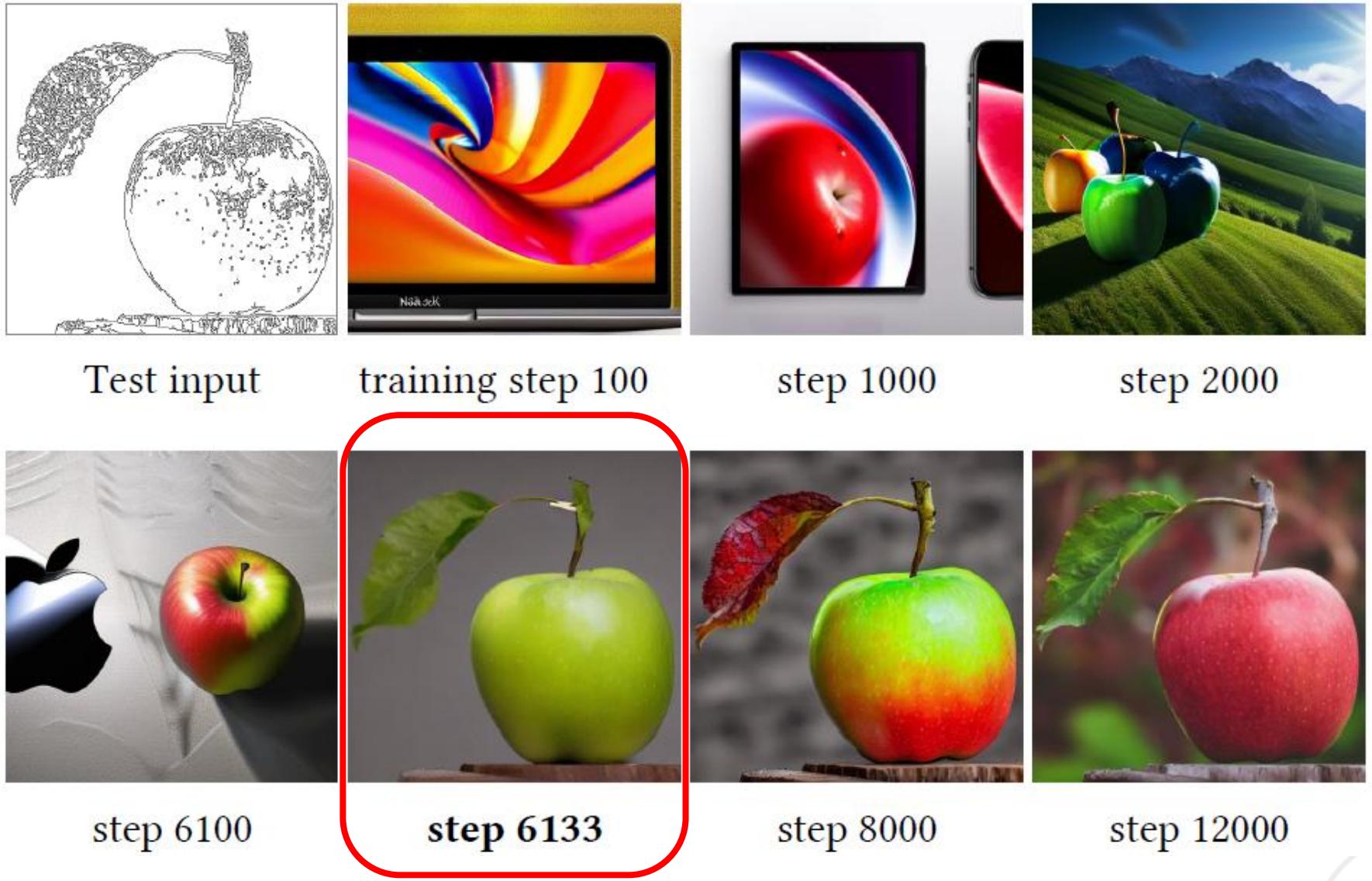
z_0: image image;
t: the number of times noise is added;
z_t: noisy image at time t;
c_t: test prompts;
c_f: task-specific condition $\quad \boldsymbol{c}_\mathrm{f} = \mathcal{E}(\boldsymbol{c}_\mathrm{i})$

Prompt dropout - randomly replaces 50% of text prompts with an empty strings

Increase ability to directly recognize semantics in the input conditioning images (e.g., edges, poses, depth, etc.) as a replacement for the prompt.

Sudden convergence phenomenon

At a certain step in the training process (e.g., the 6133 steps marked in bold), the model suddenly learns to follow the input condition.



Inference (CFG)

ε_{prd}: model's final output
ε_{uc}: unconditional output
ε_c: conditional output
β_{cfg}: user's specified weight

(a) Input Canny map (b) W/o CFG (c) W/o CFG-RW (d) Full (w/o prompt)

SD uses Classifier-Free Guidance (CFG) technique $\epsilon_{\text{prd}} = \epsilon_{\text{uc}} + \beta_{\text{cfg}}(\epsilon_{\text{c}} - \epsilon_{\text{uc}})$

Figure b: no prompts are given, adding it to ε_{uc} and ε_c will completely remove CFG guidance

Figure c: using only ε_c will make the guidance very strong

follows the edge map, but the result is less visually convincing

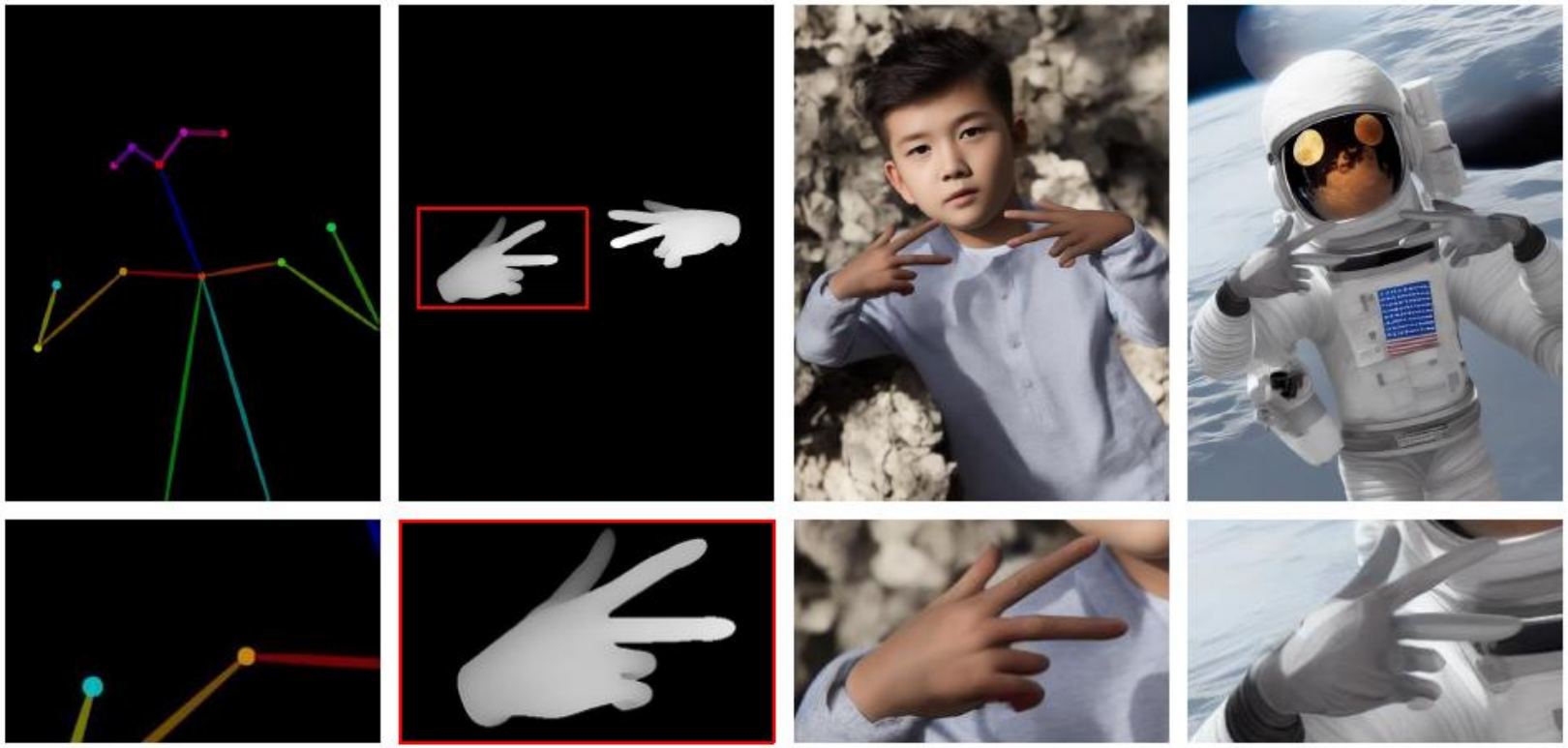
strongly follows the edge map, but over-amplification creates artifacts

Figure d: first add the conditioning image to ε_c and then multiply a weight w_i to each connection between Stable Diffusion and ControlNet according to the resolution of each block $w_i = 64/h_i$, where h_{i} is the size of i-th block

gives relatively more weight to coarse-scale features

Georgia Tech

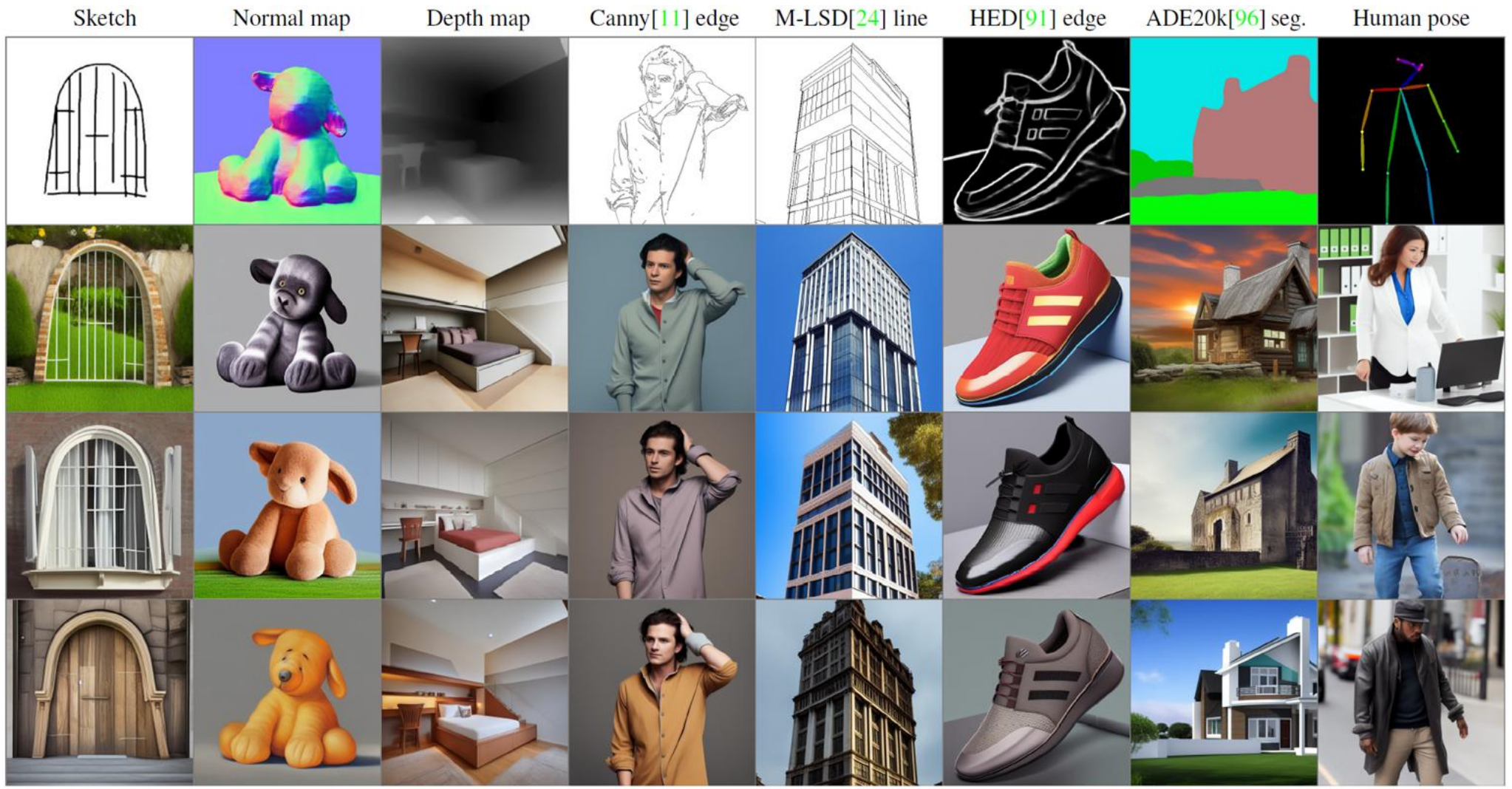
Inference (Composition of multiple conditions)



- Directly add the outputs of the corresponding ControlNets to the Stable Diffusion model
- No extra weighting or linear interpolation is necessary for such composition

Separately trained spatial controls can be combined at inference by adding their contributions, while the text prompt determines the scene’s appearance.

Controlling SD without prompts



Can recognize semantic contents in the input condition images to generate images

Georgia Tech

Ablation study (architectures)

The pretrained copy and zero-initialized connections work together: the proposed architecture follows the sketch across prompt settings, while removing zero convolutions or training the control branch from scratch produces less stable results

Georgia Tech

Quantitative evaluation (user study)

Method	Image quality ↑	Condition fidelity ↑
PITI (sketch)	1.10 ± 0.05	1.02 ± 0.01
Sketch-Guided (β = 1.6)	3.21 ± 0.62	2.31 ± 0.57
Sketch-Guided (β = 3.2)	2.52 ± 0.44	3.28 ± 0.72
ControlNet-lite	3.93 ± 0.59	4.09 ± 0.46
ControlNet	**4.22 ± 0.43**	**4.28 ± 0.45**

12 users, 20 unseen sketches

Increasing Sketch-Guided's edge scale improves sketch fidelity but lowers image quality. ControlNet ranks highest on both criteria, suggesting better control without the same apparent quality tradeoff.

Quantitative evaluation (semantic segmentation)

Method	IoU ↑	FID ↓	CLIP ↑	Aesthetic ↑
Stable Diffusion (without segmentation condition)	—	6.09	0.26	6.32
VQGAN (seg.)*	0.21 ± 0.15	26.28	0.17	5.14
LDM (seg.)*	0.31 ± 0.09	25.35	0.18	5.15
PITI (seg.)	0.26 ± 0.16	19.74	0.20	5.77
ControlNet-lite	0.32 ± 0.12	17.92	0.26	6.30
ControlNet	**0.35 ± 0.14**	15.27	0.26	6.31

SOTA segmentation: OneFormer on real ADE20K images: IoU = 0.58 ± 0.10.

ControlNet achieves the best measured segmentation adherence among the listed generators while largely retaining Stable Diffusion's text alignment and aesthetic score. Spatial control still has a measurable cost in FID and remains imperfect.

Qualitative Comparison to Baselines

Input (sketch) PITI Ours (w/o prompts)

Input (seg.) PITI Ours (default) “golden retriever”

Input (sketch) Sketch-Guided Ours (“electric fan”)

Input (canny) Taming Tran Ours (default) “white helmet

Baselines

- PITI
- Sketch-guided diffusion
- Taming transformers

ControlNet produces recognizable, detailed images from diverse conditioning images (sketches, segmentation maps, and Canny edges); text prompts can further specify what the input should represent.

Georgia Tech

Influence of different training dataset sizes

ControlNet learns recognizable control from limited data, while larger training sets improve detail and visual coherence.

Capability to interpret contents

Input "a high-quality and extremely detailed image"

If the input is ambiguous and the user does not mention object contents in prompts, the results look like the model tries to interpret input shapes.

Transferring to community models without retraining

"house"

SD 1.5

Comic Diffusion

Protogen 3.4

A compatible backbone can change style while retaining spatial control.

Follow-up Work

Work	Main idea	Limitation it addresses
T2I-Adapter (concurrent, 2023)	Add spatial controls with a lightweight adapter	A copied ControlNet branch adds compute and parameters
ControlNet 1.1 (2023)	Improve and broaden the released control models	Coverage and reliability across condition types
ControlNet++ (2024)	Compare the generated image's extracted condition with the requested condition	Noise-prediction loss only indirectly rewards following the condition

ControlNet++ augments the denoising objective with an explicit condition-consistency loss (extracts a condition map from the output and compares it with the input map) during ControlNet fine-tuning.

Main Takeaways

Main idea

Learn a spatial control branch attached to a pretrained text-to-image generator.

Key mechanism

Zero convolutions and train the copied encoder.

Empirical evidence

Better (both qualitatively and quantitatively) conditioned results than previous methods'.

From spatial control to instruction-driven image creation

	ControlNet	UniPic 2.0
User input	Spatial map + optional text	Prompt or reference image + instruction
Must preserve	Requested geometry	Instruction intent and unedited content
Primary output	A newly generated image	Edited/generated image or text
Scope	Conditional generation	Understanding, generation, editing

Shared goal: make pretrained image models follow user intent while preserving image quality.

An example

Suppose you want an astronaut making a peace sign

- ControlNet: Supply a pose skeleton showing where the astronaut's body, arms, and fingers should be, plus the prompt "an astronaut." The pose map specifies *where* and *how* the person is positioned.

- UniPic 2.0: Supply a photo of a person and instruct it: "Turn this person into an astronaut making a peace sign, but keep their face and the background." The instruction specifies *what to change* and *what to preserve* from the reference.

ControlNet shows how to make a pretrained generator obey where content should appear. UniPic 2.0 broadens the question to what should be created or changed, and what should remain consistent with the reference.

Skywork UniPic 2.0

Building Kontext Model with Online RL for Unified Multimodal Model

The challenge

Follow the instruction.
Preserve what should not change.

The question

Can better design and training give a 2B generator strong generation and editing performance?
Kontext: generation + editing.
PDTR: staged reward optimization.

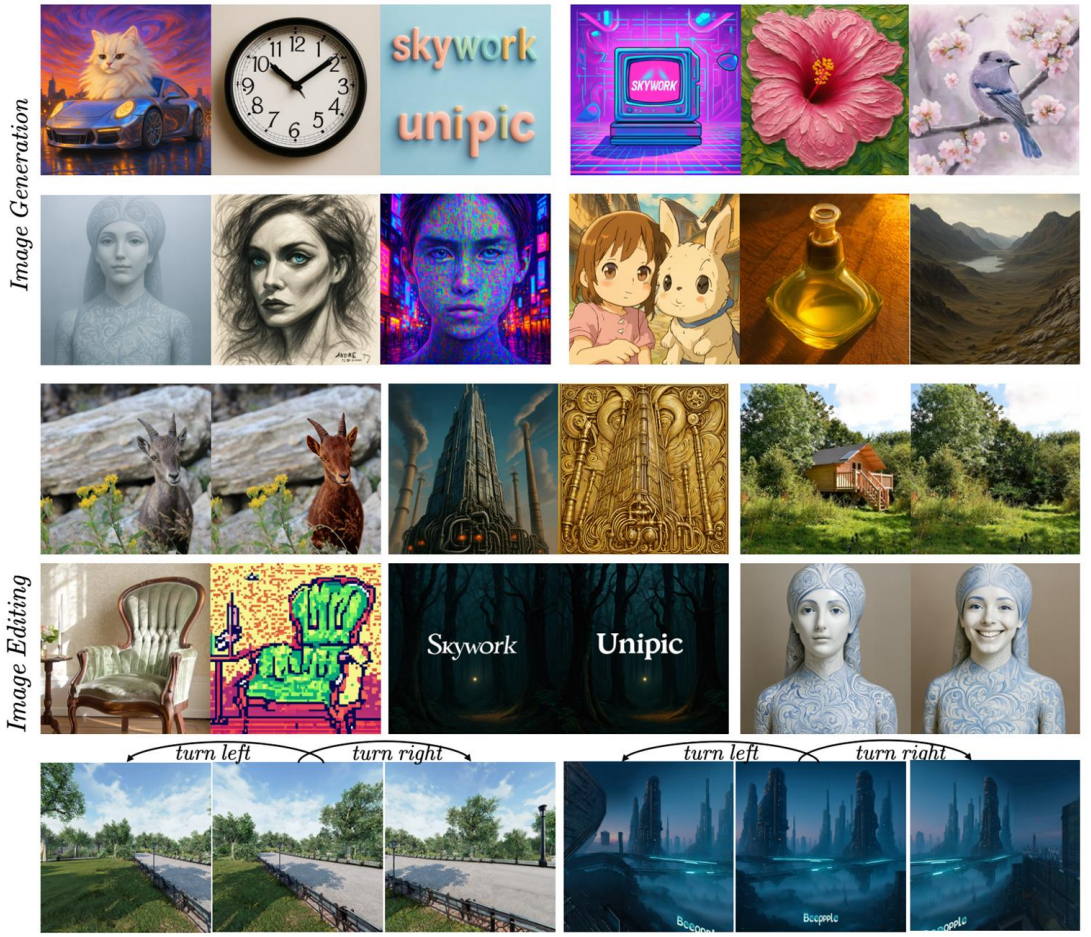


Figure 1. Generation and editing examples

Introduction and Figure 1. Supplied paper version: 22 January 2026.

Georgia Tech

Two Models, Three Tasks

Kontext: generation + editing | 2B generator

MetaQuery: Kontext + frozen MLLM + connector | adds understanding

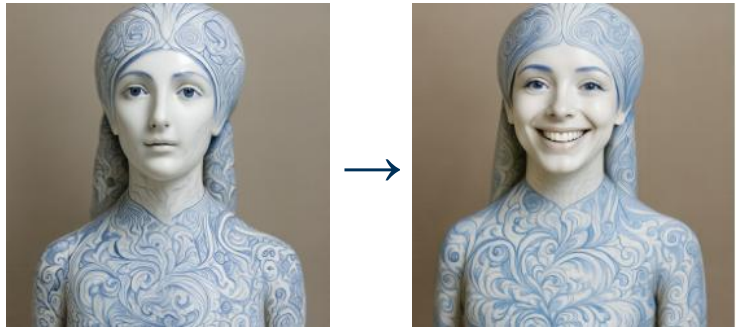
UNDERSTAND

Q: What is shown?
A: A statue of a woman.

GENERATE

“A pale statue with blue patterns.”

EDIT



“Make her smile.”
Keep the subject and style.

The bridge adds understanding; the image generator still creates the pixels.

Figure 1 crops; Sections 2.1–2.2 and 3.1. Prompts and Q&A below are teaching examples.

Georgia Tech

Where UniPic 2.0 Adds Value

UniPic 1.0: a native unified model

UniPic 2.0: connects pretrained models

Starting point	UniPic 2.0 component	What this paper adds
SD3.5-Medium	Kontext	Reference-image conditioning Generation + editing training
MetaQuery-style bridge	UniPic2-MetaQuery	Frozen understanding model connected to Kontext
Flow-GRPO	PDTR	Editing RL first Text-to-image RL second

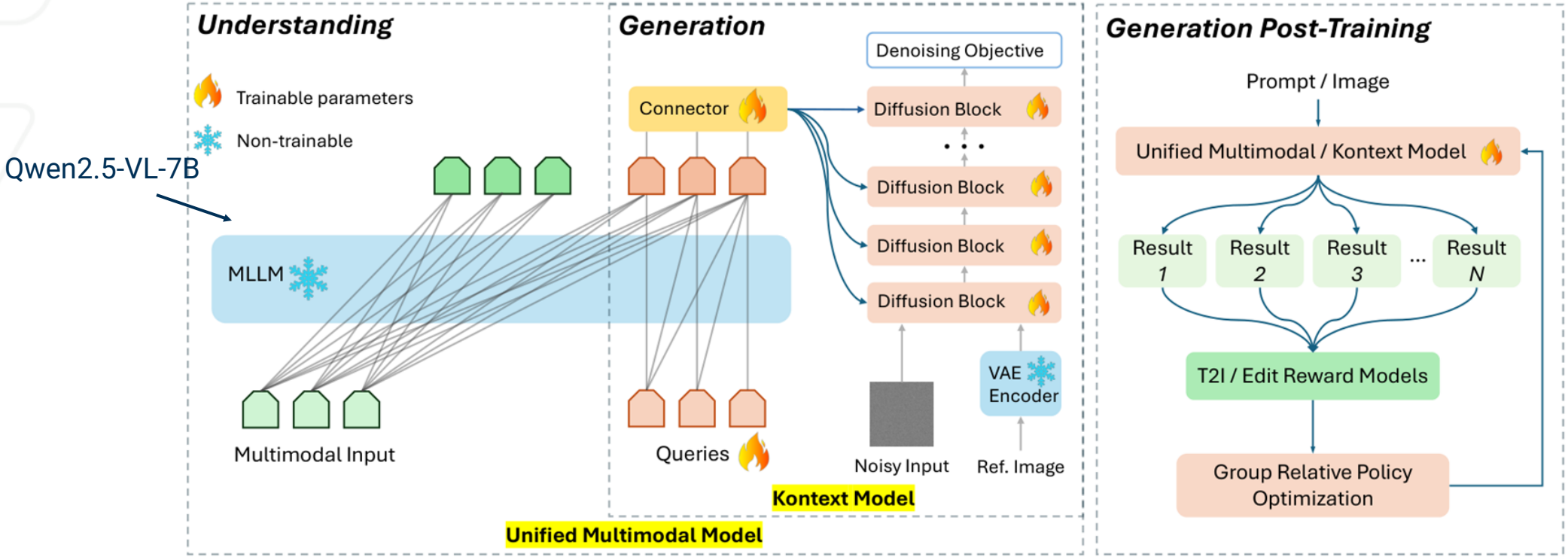
Sections 2.3 and 3. Building blocks and contributions.

Georgia Tech

Overall Architecture

1 Forward path: condition → generate

2 Training feedback



MLLM: Qwen2.5-VL-7B (frozen)
Connector: 24 layers, ~1B parameters

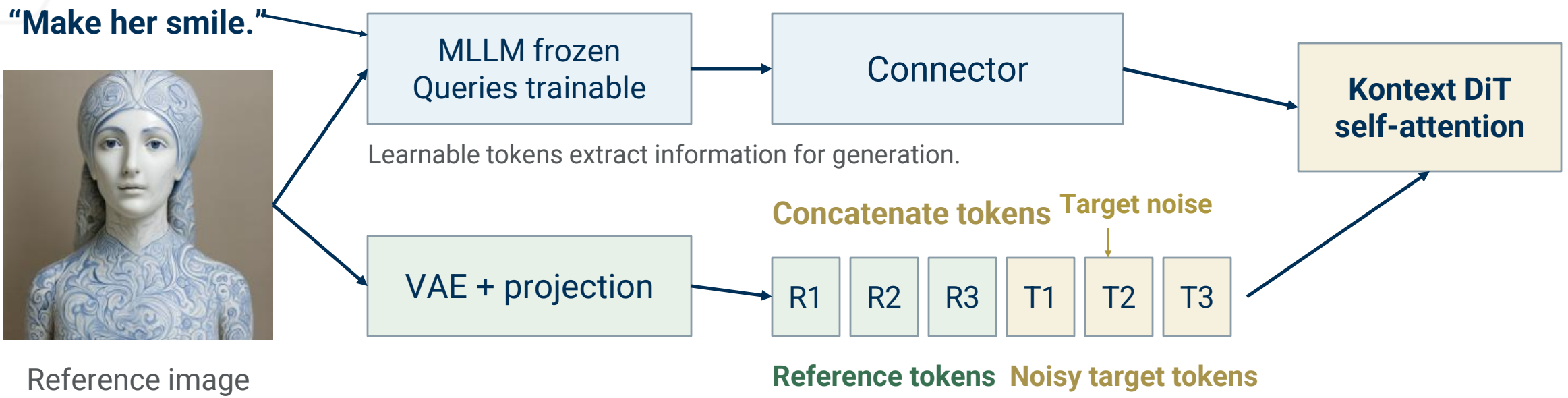
Kontext generator: 2B DiT
Based on SD3.5-Medium

Figure 2 and Sections 2–3. MLLM: Qwen2.5-VL-7B. Generator backbone: SD3.5-Medium.

Georgia Tech

Two Image Paths, One Token Sequence

Reference-image conditioning in UniPic2-MetaQuery



Position encoding separates reference and target tokens.

Section 3.1, p. 5; Figure 1 crop. Colored tokens are a schematic, not actual latent values.

Georgia Tech

Pre-training: What Is Updated?

Paired training examples and an image-generation loss; RL comes later.

Stage	MLLM	Queries + connector	DiT
Build Kontext	Not used	Not used	Train
Align	Frozen	Train	Frozen
Unified tuning **Joint supervised fine-tuning**	Frozen	Train	Train

Why freeze? Preserve understanding and reduce parameter updates.

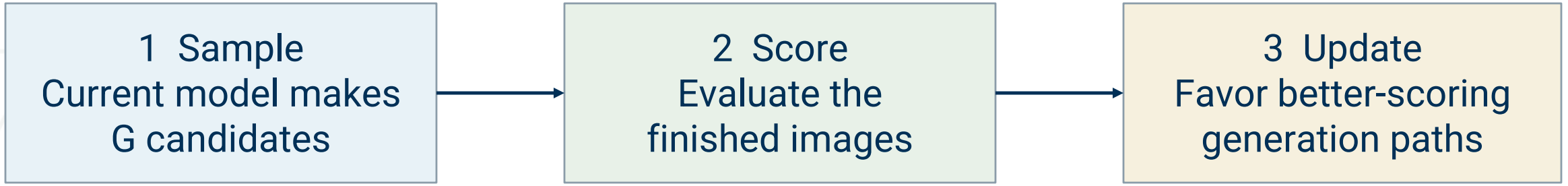
Next stage Reward-based post-training with GRPO

Section 3.1, p. 5. The Introduction describes alignment before replacing SD3.5-Medium with Kontext.

Georgia Tech

Why RL? Feedback on New Outputs

Pre-training learns image modeling; rewards target final task success.



Repeat with the updated model → fresh candidates

Generation rewards

Objects, counts, attributes, relations

Editing rewards

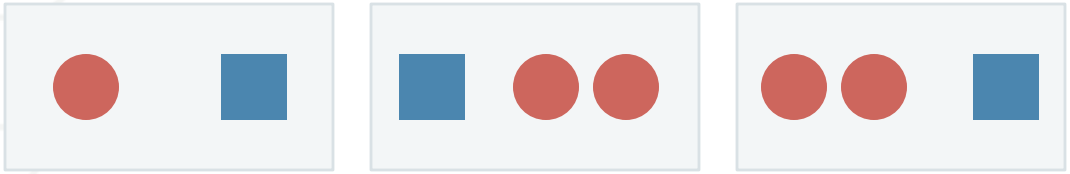
Instruction following and image quality

Introduction and Section 3.2, pp. 5–6. Online sampling repeats as the policy changes.

How RL Uses Feedback: Flow-GRPO

1 Compare candidates

“Two red balls to the left of a blue square.”



Wrong count	Wrong position	Matches prompt
R = 1, A < 0	R = 2, A = 0	R = 3, A > 0

Mean reward = 2. Positive A favors the sampled path.

2 Compute advantage

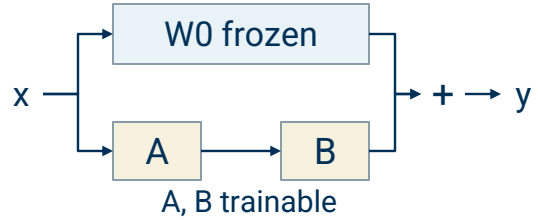
$$A_i = \frac{R\left(x_0^i, h\right) - \operatorname{mean}\left(\left\{R\left(x_0^i, h\right)\right\}_{i=1}^{G}\right)}{\operatorname{std}\left(\left\{R\left(x_0^i, h\right)\right\}_{i=1}^{G}\right)}, \qquad (1)$$

Policy ratio

$$r_t^i(\theta) = \frac{p_\theta\left(x_{t-1}^i | x_t^i, h\right)}{p_{\theta_{\text{old}}}\left(x_{t-1}^i | x_t^i, h\right)}.$$

Old policy: samples this group

LoRA parameter updates



Reference policy: KL penalty

3 Update with clipping and a KL penalty

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{h\sim\mathcal{D},\left\{x_T^i,\ldots,x_0^i\right\}_{i=1}^{G}\sim\pi_\theta}$$

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\left(\min\left(r_t^i(\theta)A_i, \operatorname{clip}\left(r_t^i(\theta), 1-\epsilon, 1+\epsilon\right)A_i\right) - \beta D_{KL}\left(\pi_\theta \| \pi_{\text{ref}}\right)\right), \qquad (2)$$

Sections 3.2 and 4.1. Teaching samples. LoRA schematic: Hu et al. (2021).

Georgia Tech

PDTR: Editing First, Then Generation

Progressive Dual-Task Reinforcement separates the two reward objectives.

1 Editing RL
Reference image + instruction

→

2 Text-to-image RL
Text prompt

EditReward or GPT-4.1

GenEval checks + detectors

Why stages?

Separate competing objectives.
Refine a new task from the model trained on the previous task.

Why this order?

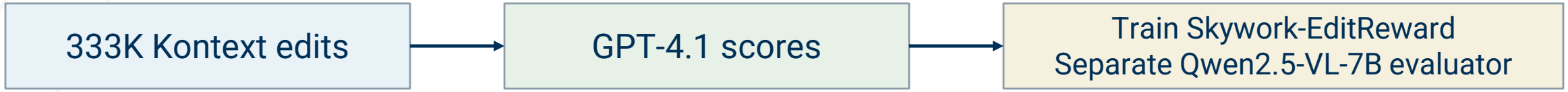
Build editing skill, then refine generation.
The paper reports positive transfer and retained editing performance.

The evidence to look for: does generation improve while editing is retained?

Section 3.2, p. 6. PDTR uses Flow-GRPO; task-order alternatives require experimental comparison.

Georgia Tech

Reward Design for Editing and Generation



Supervised regression trains the evaluator; online GRPO trains the generator.

	Editing	Generation
Evaluate	Instruction + image quality	Counts, attributes, relations
Method	EditReward or GPT-4.1	GenEval + detectors

Next: experiments

Training setup, main results, and ablation studies.

Section 3.2, p. 6. Editing reward learning and generation reward signals.

Georgia Tech

Three stages, three training signals

Kontext task learning

5M edits + 6M text-to-image samples

Adapt the 2B DiT for generation and editing.

MetaQuery alignment

150M image–text pairs; frozen MLLM + DiT

Train queries and connector to bridge the two models.

Unified tuning → online RL

Reuse task data, then sample outputs for reward feedback

Keep the MLLM frozen; use separate editing and generation rewards.

Why add RL after paired training?

Paired training

Learn from a supplied instruction and target image.

The training pair shows one desired output.

Online RL

Sample current outputs and score how well they satisfy the task.

Reward feedback targets errors in the model's own results.

Editing reward
Instruction fidelity + quality

Generation reward
Count, binding, spatial checks

Georgia Tech

Generation gains vary by benchmark

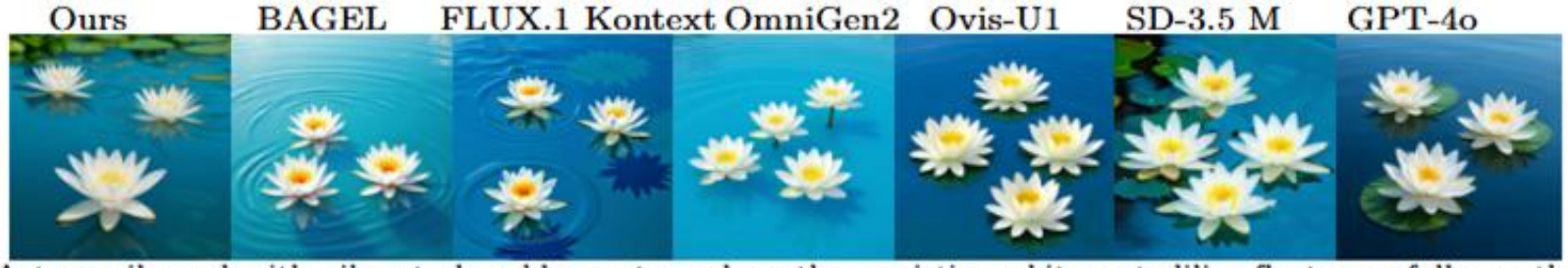
Model	# Params.	Generation		Editing		Understanding		
		GenEval	DPG	GEdit-En	Imgedit	MMBench	MMMU	MM-Vet
FLUX.1 Kontext [2]	-	-	-	6.26	3.52	×	×	×
SD3.5-Medium* [17]	-	0.65	83.86	×	×	×	×	×
BAGEL [14]	7B + 7B	0.88	85.07	6.52	3.20	85.0	55.3	67.2
Ovis-U1 [59]	2.4B + 1.2B	0.89	83.72	6.42	4.00	77.8	51.1	66.7
GPT-4o [22]	-	0.84	85.15	7.53	4.20	86.0	72.9	76.9
UniPic2-SD3.5M-Kontext	2B	0.89	84.23	6.59	4.00	×	×	×
UniPic2-SD3.5M-Kontext †	2B	0.89	84.23	6.74	4.02	×	×	×
UniPic2-Metaquery	7B + 2B	0.90	83.79	6.87	4.03	83.5	58.6	67.1
UniPic2-Metaquery †	7B + 2B	0.90	83.79	7.10	4.06	83.5	58.6	67.1

Kontext gains on GenEval, but the comparison changes architecture, data, and training together.

MetaQuery edges up on GenEval and drops on DPG

Georgia Tech

Generation: Counts, Attributes, Positions



A tranquil pond with vibrant clear blue water, where three pristine white waterlilies float gracefully on the surface. Each waterlily has a perfectly formed shape with delicate petals radiating outwards from a yellow center. The smooth surface of the water reflects the sky above, enhancing the serenity of the scene, as gentle ripples emanate outward from the blossoms.

Three waterlilies
— count

An exquisite mahogany chair with intricate carvings stands prominently in the room. The chair features a high back with elaborate designs and a plush red cushion that provides a stark contrast to the dark wood. Its legs are elegantly curved, adding to the overall sophistication of the piece. The texture of the cushion looks soft and inviting, beckoning one to sit and enjoy the comfort it offers.

Mahogany chair,
red cushion
— attributes

A dog on the left and a cat on the right.

Dog left, cat right
— position

Georgia Tech

Generation: Style, Context, and Detail

Ours BAGEL FLUX.1 Kontext OmniGen2 Ovis-U1 SD-3.5 M GPT-4o

a detailed sketch of a space shuttle, rendered in the intricate, technical style reminiscent of Leonardo da Vinci's famous drawings. The shuttle is depicted with numerous annotations and measurements, showcasing its complex design and structure. The paper on which it is drawn has an aged, yellowed appearance, adding to the historical feel of the artwork.

Aged paper and annotations support the requested Da Vinci style.

A striking orca whale is seen gliding through the blue waters of the Nile River, its black and white pattern contrasting vividly against the river's hues. In the background, the ancient silhouette of an Egyptian pyramid looms, its sandy beige stones bathed in the sunlight. The surface of the water ripples lightly as the majestic creature navigates the unusual setting, with palm trees and desert landscapes visible on the riverbanks.

Orca + Nile + pyramid tests an unusual combination.

A close-up shot of a vibrant sunflower in full bloom, with a honeybee perched on its petals, its delicate wings catching the sunlight.

The bee is near the center; placement specifically on a petal is less clear.

Georgia Tech

Main Results: Image Editing

Model	# Params.	Generation		Editing		Understanding		
		GenEval	DPG	GEdit-En	Imgedit	MMBench	MMMU	MM-Vet
FLUX.1 Kontext [2]	-	-	-	6.26	3.52	×	×	×
SD3.5-Medium* [17]	-	0.65	83.86	×	×	×	×	×
Step1X-Edit [34]	-	×	×	6.97	3.06	×	×	×
BAGEL [14]	7B + 7B	0.88	85.07	6.52	3.20	85.0	55.3	67.2
Ovis-U1 [59]	2.4B + 1.2B	0.89	83.72	6.42	4.00	77.8	51.1	66.7
GPT-4o [22]	-	0.84	85.15	7.53	4.20	86.0	72.9	76.9
UniPic2-SD3.5M-Kontext	2B	0.89	84.23	6.59	4.00	×	×	×
UniPic2-SD3.5M-Kontext †	2B	0.89	84.23	6.74	4.02	×	×	×
UniPic2-Metaquery	7B + 2B	0.90	83.79	6.87	4.03	83.5	58.6	67.1
UniPic2-Metaquery †	7B + 2B	0.90	83.79	7.10	4.06	83.5	58.6	67.1

Kontext is competitive on both editing tests with a 2B generator.

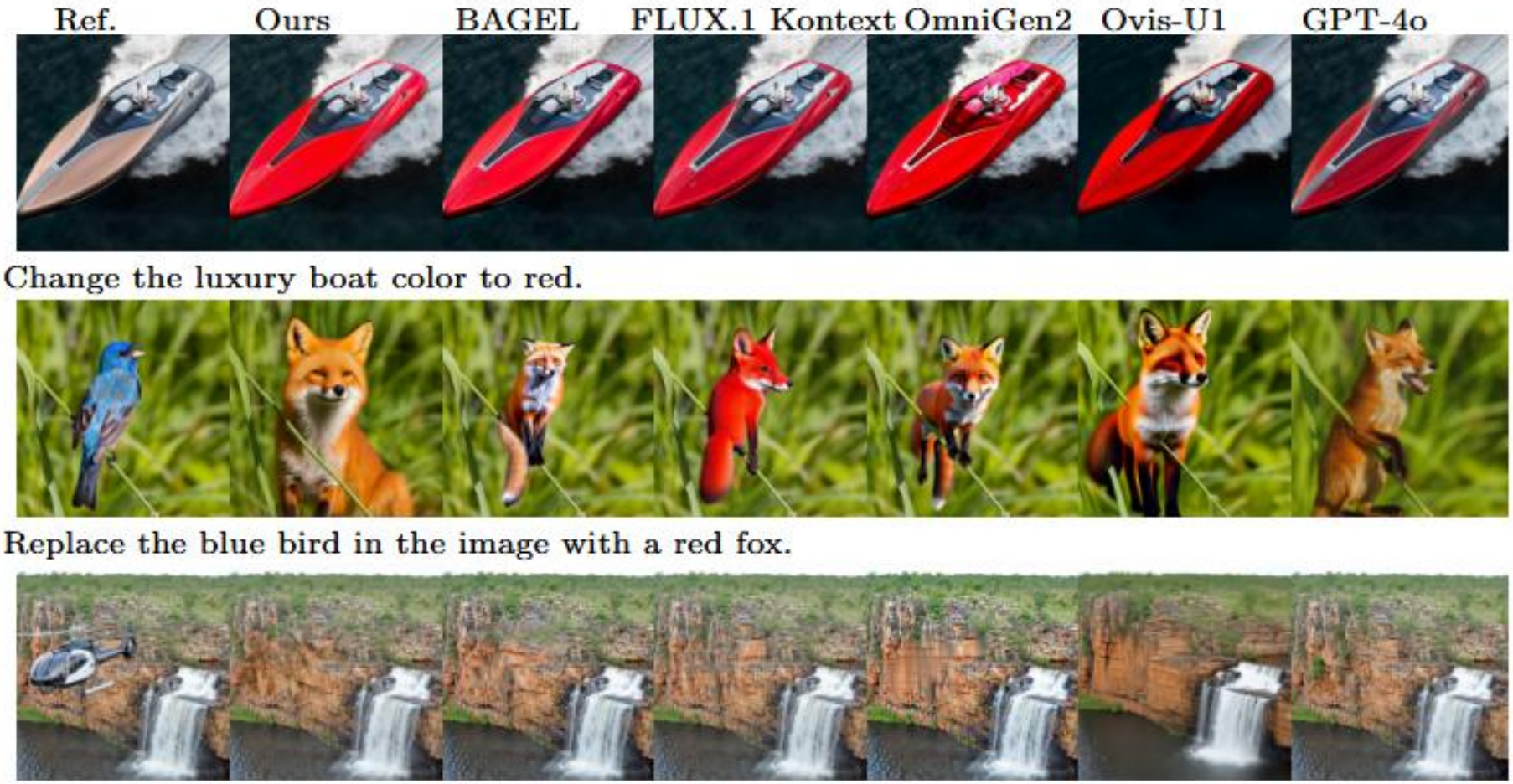
Step1X-Edit ranks above MetaQuery on GEdit-EN and below it on ImgEdit. The test changes the ranking.

A 2B generator outperforms 12B FLUX.1 Kontext and 7B BAGEL on editing benchmarks

Note: '×' indicates the model is incapable of performing the task. '†' indicates results obtained using the official OpenAI API. '*' denotes generation results for SD3.5-Medium are reported based on our evaluation.

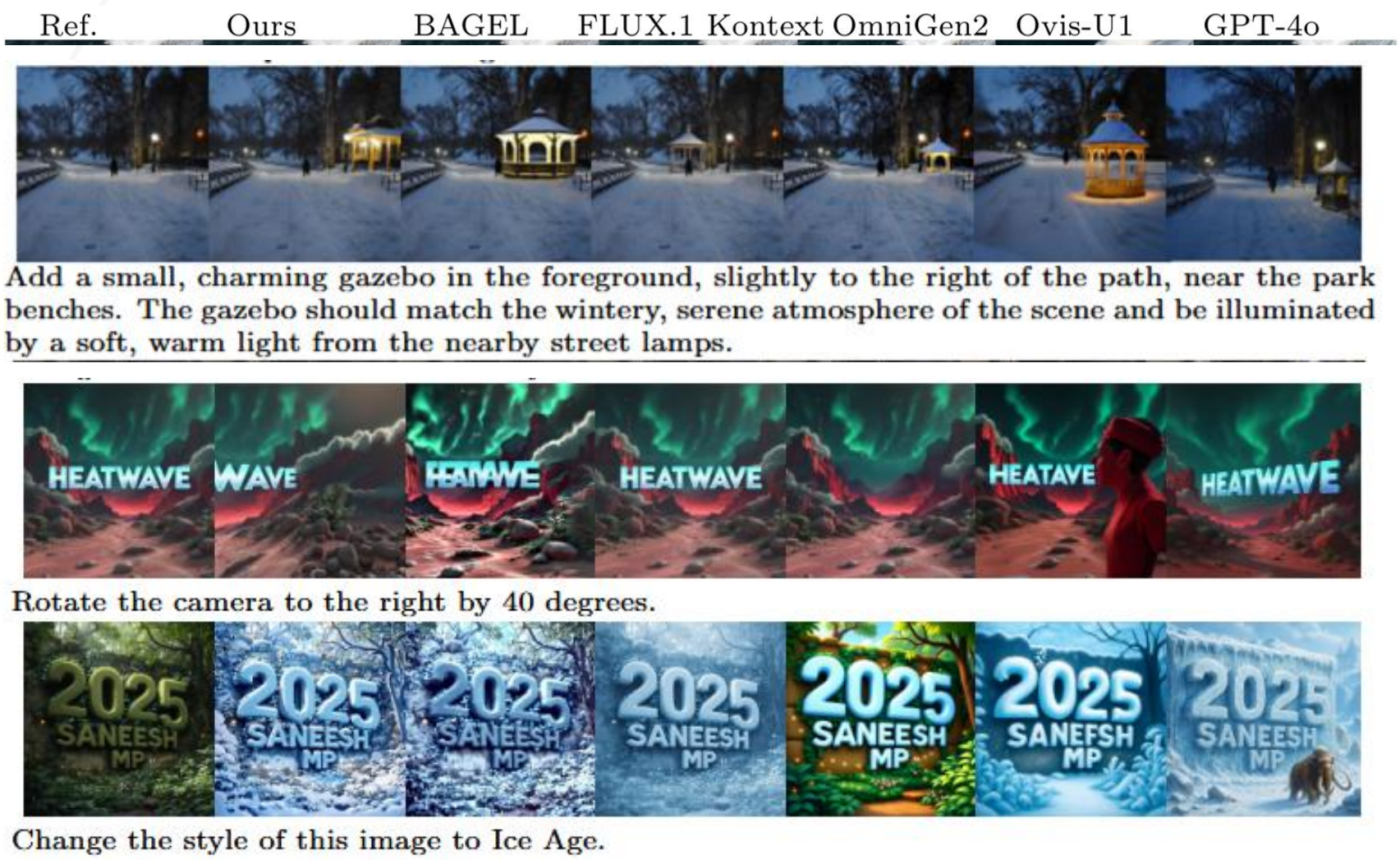
Georgia Tech

Successful edits preserve useful context



Georgia Tech

Editing: Add, Rotate, and Restyle



- Rotate the camera by 40°
- HEATWAVE becomes WAVE: an unrelated detail is lost during the edit.

Main Results: Understanding

Model	# Params.	Generation		Editing		Understanding		
		GenEval	DPG	GEdit-En	Imgedit	MMBench	MMMU	MM-Vet
MetaQuery-XL [38]	7B + 1.6B	0.80	82.05	-	-	83.5	58.6	66.6
UniWorld-V1 [30]	7B + 12B	0.84	81.38	4.85	3.26	83.5	58.6	67.1
Blip3-o-8B [7]	7B + 1.4B	0.84	81.60	×	×	83.5	58.6	66.6
OmniGen2 [67]	3B + 4B	0.86	83.57	6.42	3.44	79.1	53.1	61.8
BAGEL [14]	7B + 7B	0.88	85.07	6.52	3.20	85.0	55.3	67.2
Ovis-U1 [59]	2.4B + 1.2B	0.89	83.72	6.42	4.00	77.8	51.1	66.7
GPT-4o [22]	-	0.84	85.15	7.53	4.20	86.0	72.9	76.9
UniPic2-SD3.5M-Kontext	2B	0.89	84.23	6.59	4.00	×	×	×
UniPic2-SD3.5M-Kontext †	2B	0.89	84.23	6.74	4.02	×	×	×
UniPic2-Metaquery	7B + 2B	0.90	83.79	6.87	4.03	83.5	58.6	67.1
UniPic2-Metaquery †	7B + 2B	0.90	83.79	7.10	4.06	83.5	58.6	67.1

MetaQuery combines generation and editing with Qwen2.5-VL's understanding capabilities.

Georgia Tech

Understanding Examples

User: What kind of bird is this? Please give its name.
Model Output: The bird in the picture is a Blue-throated Bee-eater (Merops superciliosus). It is known for its vibrant colors, including a bright blue throat and a colorful plumage that includes yellow, green, and orange. These birds are native to parts of Africa and are often found in open woodlands and savannas.

User: What are these attractions? Please give their names.
Model Output : The images in the picture depict the following famous attractions

1.The Great Wall of China - located in northern China, it is one of the most iconic and longest man-made structures in the world.
2.Eiffel Tower - situated in Paris, France, it is an architectural landmark and one of the most recognizable symbols of France.
3.Statue of Liberty - located in New York Harbor, United States, it is a colossal neoclassical sculpture on Liberty Island.
4.Terracotta Army - located in Xi'an, Shaanxi, China, it is a collection of terracotta sculptures depicting the armies of Qin Shi Huang, the first Emperor of China. These landmarks are some of the most visited and recognized sites around the world.

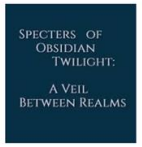
User: Locate the elephant and lion, report the bbox coordinates in JSON format.
Model Output:
[{"bbox_2d": [150, 34, 227, 116], "label": "elephant"},
{"bbox_2d": [89, 58, 150, 126], "label": "lion"}]

User: Where is the scene? Describe the scene in detail.
Model Output:

The scene depicts a picturesque view of a canal, likely in Venice, Italy, during the evening or early night. The sky is dark with a mix of deep blues and purples, indicating that it is either dusk or dawn. The canal is lined with historic buildings on both sides, featuring traditional Venetian architecture with ornate facades, balconies, and arched windows. The buildings are illuminated by warm, golden lights from street lamps and the interiors of the establishments, creating a cozy and inviting atmosphere.

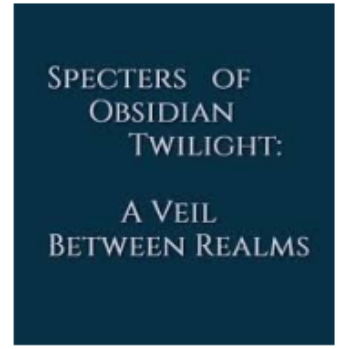
Several boats are visible on the water, including gondolas, which are traditional Venetian rowboats. One gondola in the foreground is occupied by people, suggesting that it might be a tourist ride. The water reflects the lights from the buildings and boats, adding to the serene and enchanting ambiance of the scene. On the left side of the canal, there are several restaurants with outdoor seating, as indicated by the red awnings and the presence of tables and chairs. The overall setting is tranquil and romantic, capturing the essence of Venice's charm and beauty.

User: What the text is in the image?
Model Output: The text in the image reads: "Specters of Obsidian Twilight: A Veil Between Realms."



User: Locate the elephant and lion, report the bbox coordinates in JSON format.
Model Output:
[{"bbox_2d": [150, 34, 227, 116], "label": "elephant"},
{"bbox_2d": [89, 58, 150, 126], "label": "lion"}]

User: What the text is in the image?
Model Output: The text in the image reads: "Specters of Obsidian Twilight: A Veil Between Realms."



Recognition · Scene description· Grounding · OCR

Georgia Tech

ABLATION

Where Should the Reference Image Go?

Table 2: Different image conditioning strategies.

#	Model	Image Condition MLLM	Image Condition DiT	GEdit-EN
1	UniPic2-SD3.5M-Kontext	✘	✓	6.31
2	UniPic2-MetaQuery	✓	✘	5.00
3		✘	✓	6.40
4		✓	✓	**6.90**

Which Components Should Be Fine-tuned?

Table 3: Freezing or fine-tuning (FT) the connector and DiT for UniPic2-MetaQuery.

#	Connector	DiT	GenEval	GEdit-EN
1	Freeze	FT	0.84	6.75
2	FT	Freeze	0.85	6.46
3	FT	FT	**0.86**	**6.90**

- Both reference-image paths give the best editing score.
- Joint connector–DiT tuning performs best on both reported metrics.

Georgia Tech

ABLATION

What Does Each Reward Add?

Table 4: Ablation Study of Reward Signals on UniPic2-SD3.5M-Kontext.

Model	RL		Generation		Editing	
	T2I	Edit	GenEval	DPG	GEdit-EN	ImgEdit
w/o GRPO	✘	✘	0.83	83.68	6.31	3.95
w/ GPT-4.1	✘	✓	0.83	84.15	6.55	4.04
w/ EditReward	✘	✓	0.83	83.99	6.59	4.00
w/ GenEval	✓	✘	0.87	83.97	6.40	3.95
w/ GPT-4.1 + GenEval	✓	✓	0.89	84.23	6.54	3.99
w/ EditReward + GenEval	✓	✓	0.89	84.36	6.59	4.00

Baseline: no RL

One reward

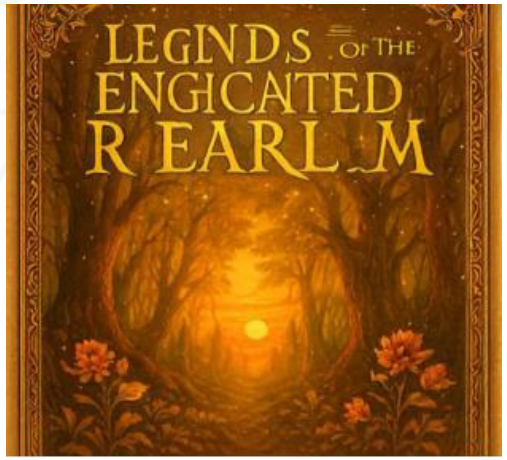
Both rewards

Adding generation RL after EditReward improves GenEval while retaining the reported editing scores.

Georgia Tech

Fine-grained constraints still fail

Book title



Requested: “Legends of the Enchanted Realm”. Letters are malformed.

Reflection

Requested dog reflection; the output still reflects a tiger.

Extraction

Model Output:

Reference above, output below: extraction leaves a blurred region.

Clock text

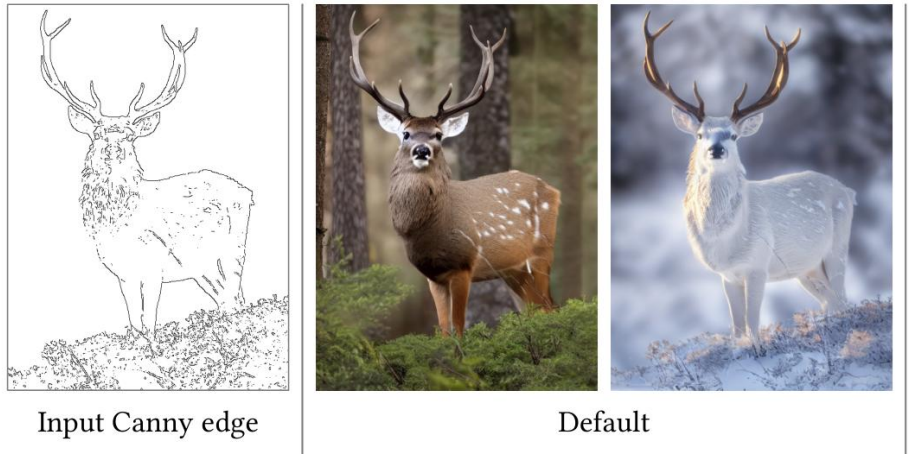
Model Output:

Requested: SKYWORKUNIPI. Output lettering is incomplete.

Georgia Tech

Control inputs express different intentions

ControlNet



Edge, depth, or pose maps make spatial structure explicit; text specifies desired content.

UniPic 2.0 editing

A reference image supplies content; the instruction specifies what should change and remain.

Shared goal: extend a pretrained generator with visual conditions the user can supply.

Georgia Tech

Reuse stays; adaptation changes

ControlNet

Freeze the generator; learn a control branch.

A copied encoder reads spatial conditions. Zero-initialized links initially preserve the pretrained U-Net output.

Motivation: preserve pretrained capabilities and reduce overfitting when condition-specific data are limited.

UniPic 2.0 / MetaQuery

Adapt the generator and its multimodal bridge.

Reference VAE tokens enter the DiT. A frozen MLLM adds conditioning through queries and a connector.

During unified tuning, the MLLM stays frozen. Table 3 supports updating the connector and DiT together.

Task scope shapes training and evaluation

Dimension	ControlNet	UniPic 2.0
Supervision	Spatial condition–image pairs	Image–text alignment, edits, generation
Optimization	Conditioned denoising; frozen base U-Net	Paired task training + online task rewards
Evaluation	Image quality + condition fidelity	Separate generation, editing, understanding tests

What is convincing, and what remains open

Supported strengths

Complementary inputs

Both reference paths score 6.90, above either route alone.

Joint adaptation

Updating connector and DiT works best among the tested settings.

Combined rewards

Generation improves while editing holds in the EditReward route .

Limitations and open questions

Local precision

Viewpoint edit loses text ; extraction and unusual relations fail.

Reward coverage

Composition checks and learned judges cover selected aspects of output quality.

Training order

Does editing-generation training order matter? Would the reverse training order give similar results?

Georgia Tech

Takeaways

Both papers extend pretrained image generators with visual control.

ControlNet adds explicit spatial guidance. UniPic brings generation, editing, and understanding into one system.

UniPic’s ablations support its design choices, while exact text and unusual relations remain challenging.

Full benchmark comparison

Type	Model	# Params.	Generation		Editing		Understanding		
			GenEval	DPG	GEdit-En	Imgedit	MMBench	MMMU	MM-Vet
Generation	SDXL [40]	-	0.55	74.65	×	×	×	×	×
	DALL-E 3 [3]	-	0.67	83.50	×	×	×	×	×
	FLUX.1-dev [25]	-	0.67	84.00	×	×	×	×	×
	FLUX.1 Kontext [2]	-	-	-	6.26	3.52	×	×	×
	SD3.5-Medium* [17]	-	0.65	83.86	×	×	×	×	×
Editing	AnyEdit [77]	-	×	×	3.21	2.45	×	×	×
	Instructuct-P2P [5]	-	×	×	3.68	1.88	×	×	×
	MagicBrush [79]	-	×	×	4.52	1.90	×	×	×
	Step1X-Edit [34]	-	×	×	6.97	3.06	×	×	×
Unified	Emu3 [62]	8B	0.66	80.60	-	-	58.5	31.6	37.2
	Janus-Pro [11]	7B	0.80	84.19	×	×	75.5	36.3	39.8
	MetaQuery-XL [38]	7B + 1.6B	0.80	82.05	-	-	83.5	58.6	66.6
	UniWorld-V1 [30]	7B + 12B	0.84	81.38	4.85	3.26	83.5	58.6	67.1
	Blip3-o-8B [7]	7B + 1.4B	0.84	81.60	×	×	83.5	58.6	66.6
	OmniGen2 [67]	3B + 4B	0.86	83.57	6.42	3.44	79.1	53.1	61.8
	BAGEL [14]	7B + 7B	0.88	85.07	6.52	3.20	85.0	55.3	67.2
	Ovis-U1 [59]	2.4B + 1.2B	0.89	83.72	6.42	4.00	77.8	51.1	66.7
	GPT-4o [22]	-	0.84	85.15	7.53	4.20	86.0	72.9	76.9
Ours	UniPic2-SD3.5M-Kontext	2B	0.89	84.23	6.59	4.00	×	×	×
	UniPic2-SD3.5M-Kontext †	2B	0.89	84.23	6.74	4.02	×	×	×
	UniPic2-Metaquery	7B + 2B	0.90	83.79	6.87	4.03	83.5	58.6	67.1
	UniPic2-Metaquery †	7B + 2B	0.90	83.79	7.10	4.06	83.5	58.6	67.1

Original Table 1. † Official OpenAI API; * authors' evaluation of SD3.5-Medium.

Georgia Tech

Detailed training settings

Stage	Data / updates
Kontext task learning	5M edits + 6M generations; train DiT
MetaQuery alignment	150M pairs: 30M public + 120M internal; queries + connector
Unified tuning	Reuse task data; update queries, connector, DiT; freeze MLLM
Editing RL	ShareGPT4o-Image; GPT-4.1 or EditReward
Generation RL	GenEval training set used by Flow-GRPO; GenEval reward

RL settings: 512 × 512; group size G = 16; sampling steps T = 10; LoRA rank 32, α = 64.

References

[1] L. Zhang, A. Rao, and M. Agrawala.
Adding Conditional Control to Text-to-Image Diffusion Models. ICCV, 2023. arXiv:2302.05543v3.

[2] Skywork Multimodality Team.
Skywork UniPic 2.0: Building Kontext Model with Online RL for Unified Multimodal Model. arXiv:2509.04548v2, 22 January 2026.

[3] R. Rombach et al.
High-Resolution Image Synthesis with Latent Diffusion Models. CVPR, 2022. arXiv:2112.10752.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros.
Image-to-Image Translation with Conditional Adversarial Networks. CVPR, 2017.

[5] A. Hertz et al.
Prompt-to-Prompt Image Editing with Cross-Attention Control. ICLR, 2023. arXiv:2208.01626.

[6] T. Wang et al.
Pretraining is All You Need for Image-to-Image Translation. arXiv:2205.12952, 2022.

References (cont.)

[7] C. Mou et al. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. arXiv:2302.08453, 2023.

[8] M. Li et al. ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback. ECCV, 2024. arXiv:2404.07987.

[9] L. Zhang and contributors. ControlNet: official code and documentation. github.com/lllyasviel/ControlNet

[10] L. Zhang and contributors. ControlNet 1.1: official release, 2023. github.com/lllyasviel/ControlNet-v1-1-nightly

Why Zero Convolutions Can Learn

Consider $y = wx + b$, then

$$\partial y/\partial w = x \qquad \partial y/\partial x = w \qquad \partial y/\partial b = 1$$

and if $w = 0$ and $x \neq 0$, then

$$\partial y/\partial w \neq 0 \qquad \partial y/\partial x = 0 \qquad \partial y/\partial b \neq 0$$

which means as long as $x \neq 0$, one gradient descent iteration will make w non-zero.

Then

$$\partial y/\partial x \neq 0$$

so that the zero convolutions will progressively become a common conv layer with non-zero weights.