Generative Visual Chain-of-Thought for Image Editing

arXiv 2603

Zijin Yin, Tiankai Hang, Yiji Cheng, Shiyi Zhang, Runze He, Yu Xu,
Chunyu Wang, Bing Li, Zheng Chang, Kongming Liang, Qinglin Lu, Zhanyu Ma
Beijing University of Posts and Telecommunications,
Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance,
Tencent Hunyuan, King Abdullah University of Science and Technology

STRUCT Group Seminar
Presenter: Lan Xicheng
2026.5.10

Outline

- Background
- Methods
- Experiments
- Conclusion

Background

Instruction-Guided Image Editing

- Input: source image + textual instruction
- Output: edited image satisfying the instruction

- **Key requirements:**
  - Instruction adherence
  - Background preservation

Background

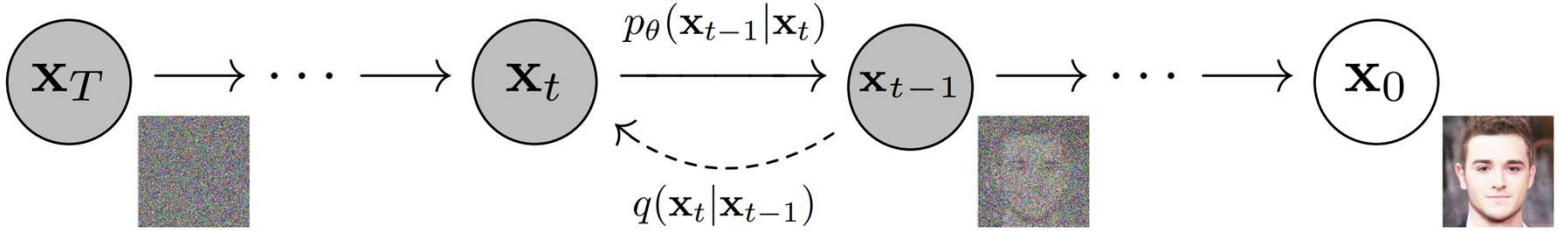
Editing Categories

- Content: Add / Remove / Replace objects
- Attribute: Color / Material / Local style
- Spatial: Move / Resize / Rotate / Pose change
- Background change, multi-target complex edits

- This paper covers 19 sub-tasks

Background

Diffusion Models



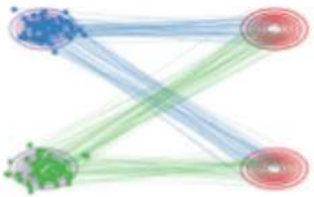
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (1)$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad (2)$$

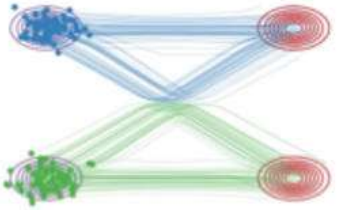
Background – Flow Matching vs. Diffusion

- Diffusion: Models curved stochastic paths (SDE)
- Flow Matching: Models straight velocity fields (ODE)
- Faster sampling, better couplings

$$\min_v \int_0^1 \mathbb{E}\left[\|(X_1 - X_0) - v(X_t, t)\|^2\right] \mathrm{d}t. \quad X_t = tX_1 + (1-t)X_0, \quad t \in [0,1].$$



Linear interpolations



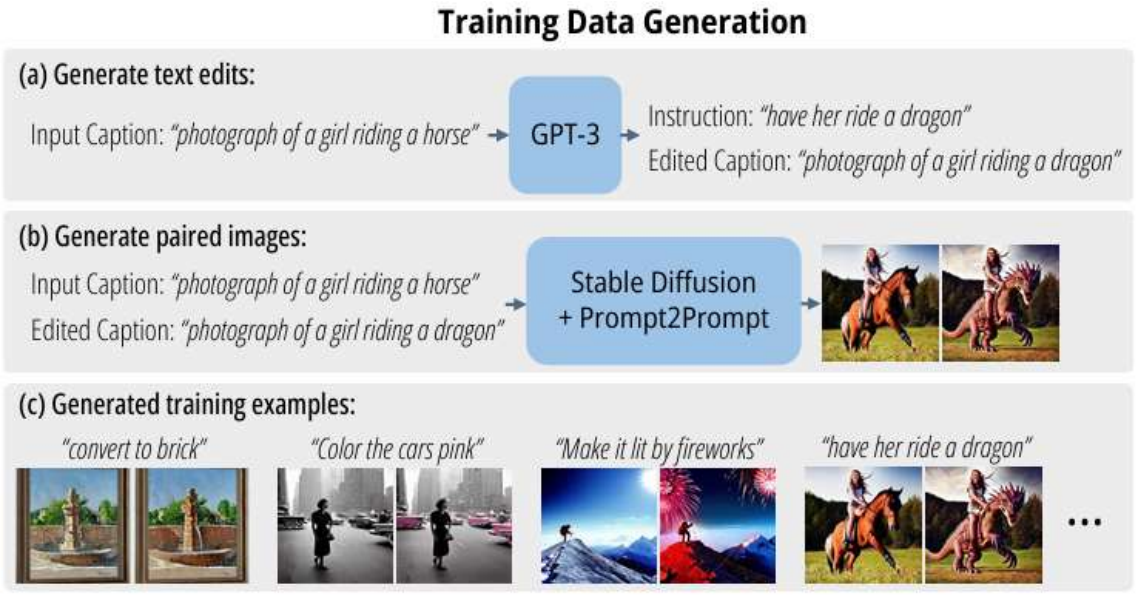
Learned vector field

Background

InstructPix2Pix (CVPR 23)

- First instruction-driven editing
- Synthetic dataset via GPT-3 + Stable Diffusion
- Treat instructions as direct conditioning input

- Limitation: poor on complex / spatial instructions
- Triggered a wave of follow-up editing models



Background

Recent Editing Models (2023-2025)

- MagicBrush (NeurIPS 2023): manually annotated dataset
- Step1X-Edit (2025): unified DiT-based pipeline
- FLUX.1 Kontext (2025): in-context flow matching
- Qwen-Image-Edit (2025): SOTA open-source

——Capability grows rapidly, but a core problem remains

Background

"Where to Edit?"

- Existing methods struggle:
  - Complex scenes with multiple similar entities
  - Fine-grained spatial referring
  - Implicit / context-dependent references

Input

"Make them look like flight attendants"

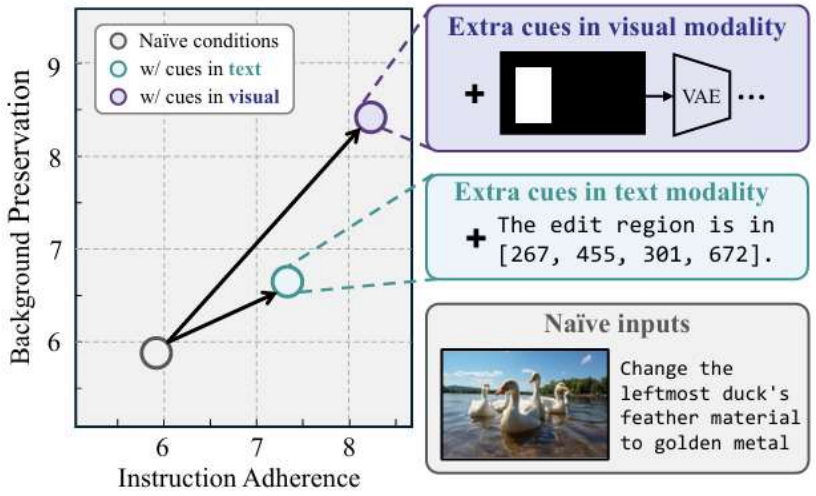
"Make them look like doctors"

Background

A Motivating Example

- Instruction: "Change the leftmost duck's feather material to golden metal"
- Without spatial reasoning, the model cannot reliably localize "leftmost"

→Spatial reasoning is required before pixel-level editing



Background

From LLMs to MLLMs

- **Large Language Models (LLMs)**
  - GPT, LLaMA — emergent reasoning capability

- **Multimodal LLMs (MLLMs)**
  - LLaVA, Qwen-VL, GPT-4V — accept image + text input
  - Strong perception, but cannot natively generate images

Background

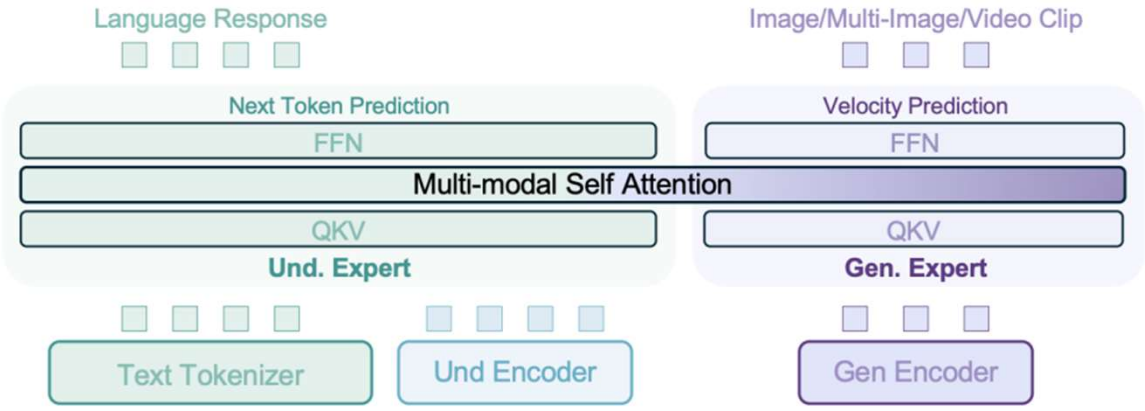
From LLMs to MLLMs – Image Generation

- **A. LLM + Diffusion**
  - Prompt → LLM → Prompt → Diffusion
  - DALL·E 2/Imagen
- **B. Auto Regressive**
  - Prompt → LLM → Auto Regressive Transformer → Tokens → Image
  - GPT 4
- **C. Combined**
  - Prompt → LLM → Auto Regressive Transformer → Rectified Flow
  - JanusFlow/BAGEL

Background

From LLMs to MLLMs – Image Generation

- A. LLM + Diffusion
- B. Auto Regressive
- **C. Combined**
  - Prompt → LLM → Auto Regressive Transformer → Rectified Flow
  - JanusFlow/BAGEL



Background

MLLMs as Multi-Role Agents

- In this paper, MLLMs play 3 roles:
  - Open-vocabulary visual grounding
  - Data annotators — generate instructions, predict bboxes
  - Evaluators — MLLM-as-a-Judge for reward & evaluation

Background

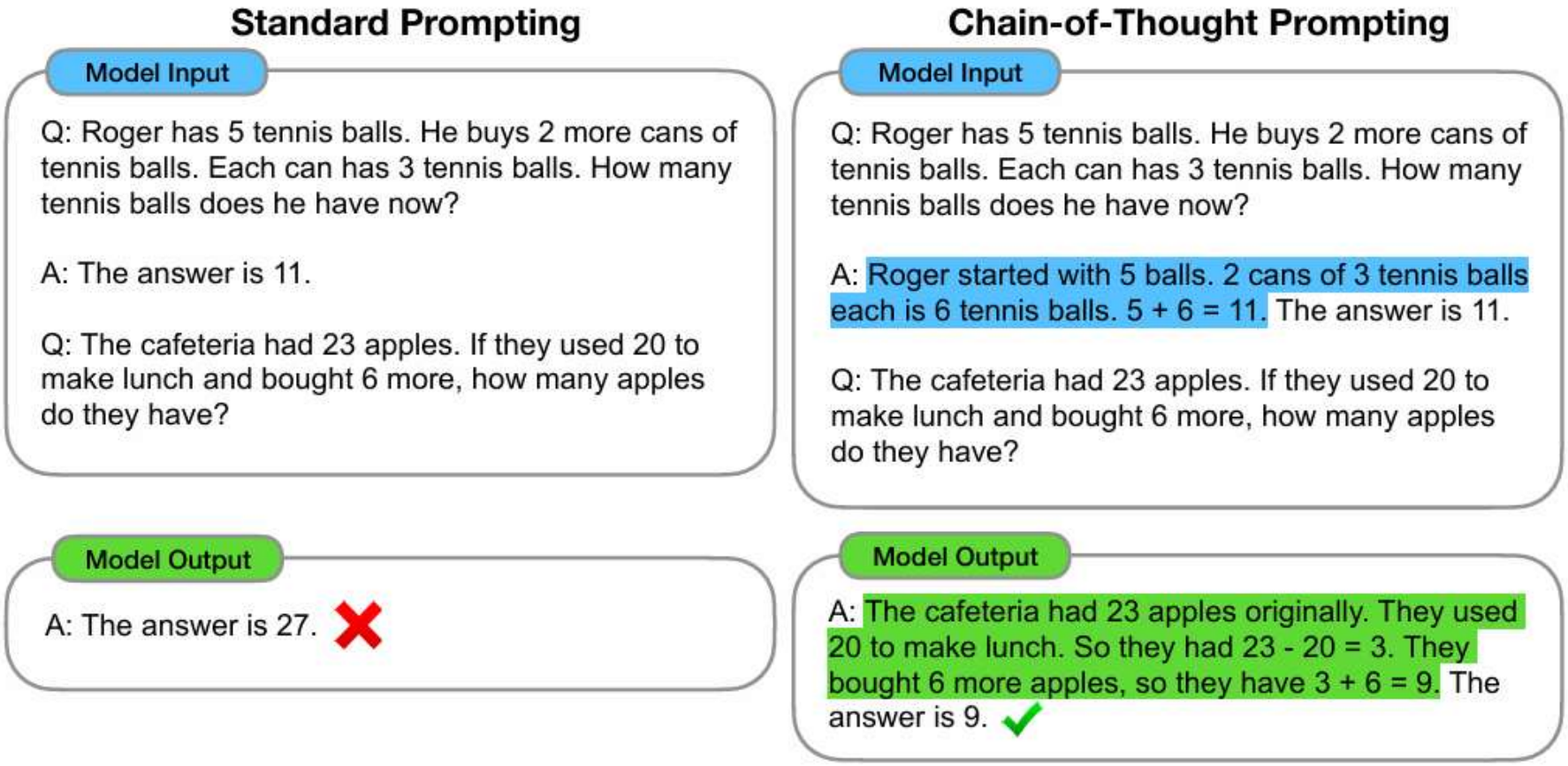
Chain-of-Thought (NeurIPS 2022)

- Let the model "think step-by-step" before answering
- Major boost on math, commonsense, reasoning

- Inference-time scaling:
  - More thinking → better results
  - Foundation of modern reasoning paradigms

Background

Chain-of-Thought (NeurIPS 2022)

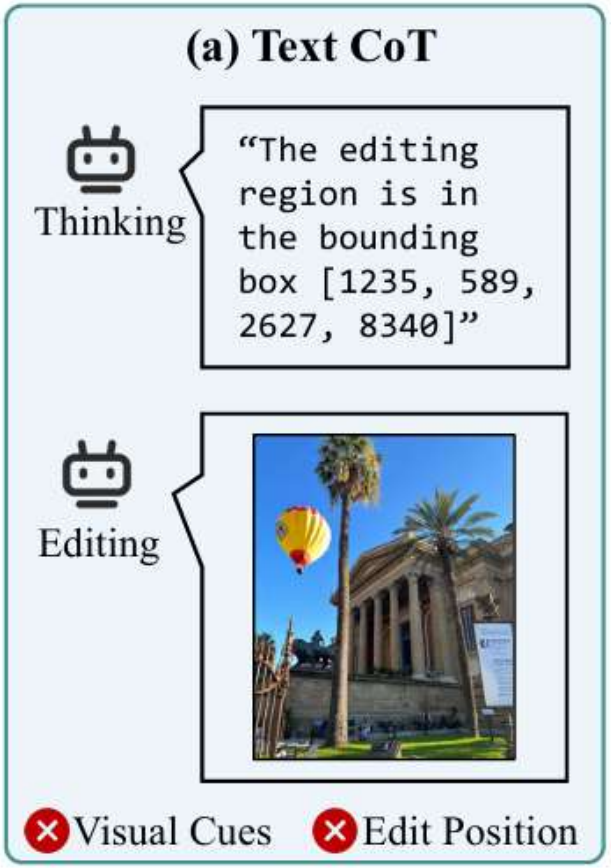


Background

Text CoT for Image Generation

- GoT / GoT-R1 (2025)
  - Generate textual reasoning trace before image
  - e.g., "The editing region is in [1235, 589, ...]"

- Limitation:
  - Spatial info compressed into a few coordinates
  - Reasoning happens entirely in token space



Background

"Painting Twice" Hypothesis

- Cognitive Science (Barsalou, 1999):
  - Visual reasoning is modality-specific
  - A skilled artist first imagines, then paints

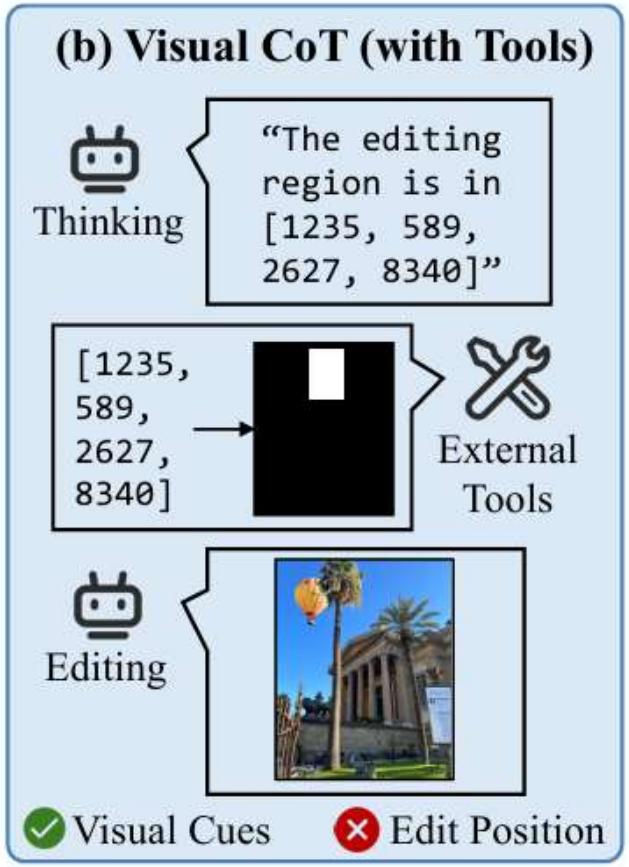
- Implication for AI:
  - Reasoning in vision space may help visual tasks
  - Not just describing — actually drawing the thought

Background

Visual CoT (with Tools)

- Visual Sketchpad (NeurIPS 2024): drawing auxiliary lines
- Pixel Reasoner (2025): zoom-in via crops
- Refocus / Visual Abstract Thinking: region highlighting

- Limitation:
  - Reasoning still text-driven; tools just render visuals

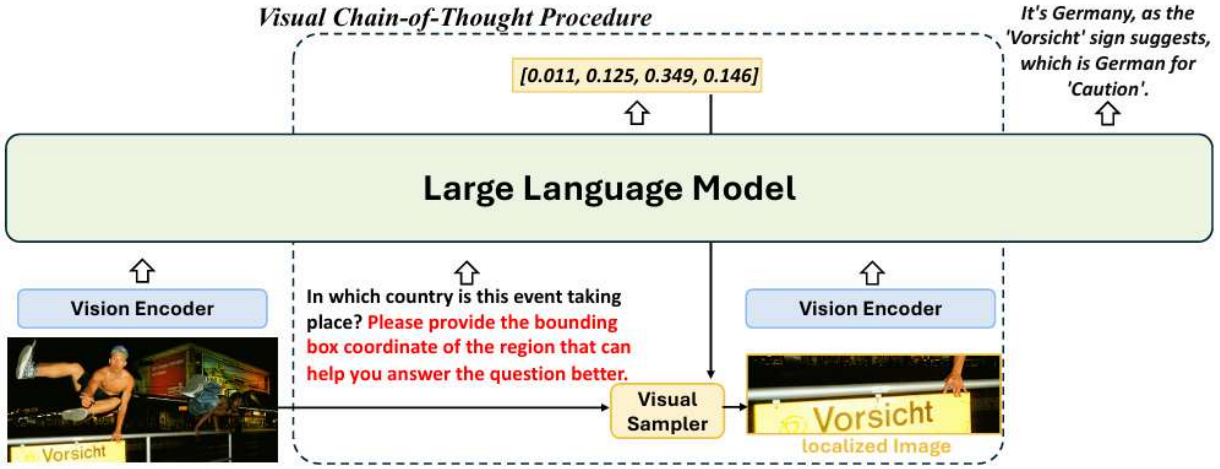


Background

Native Visual CoT

- Model itself generates intermediate visual artifacts

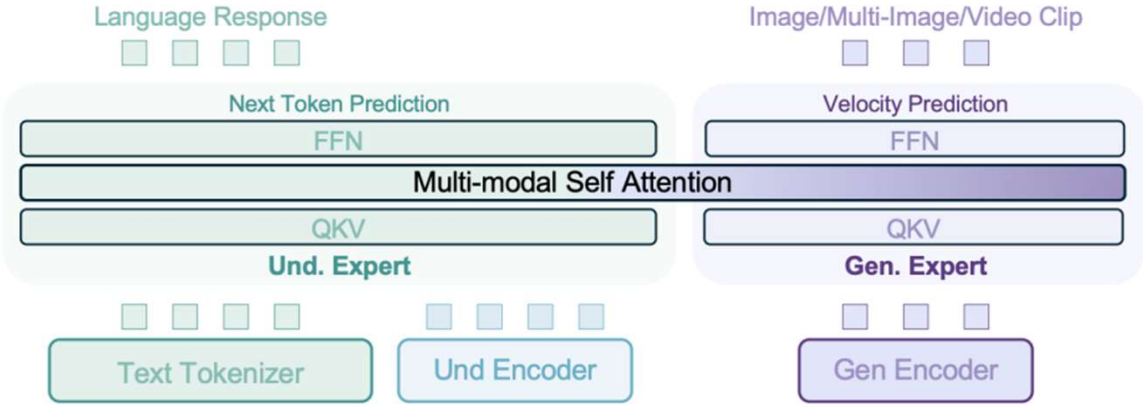
- Largely unexplored in image editing — gap this paper fills



Background

Bagel (Bytedance, 2025)

- Mixture-of-Transformers (MoT) architecture
- Two experts:
  - Understanding Expert (with ViT encoder)
  - Generation Expert (with VAE encoder)
- 7B parameters, supports interleaved I/O



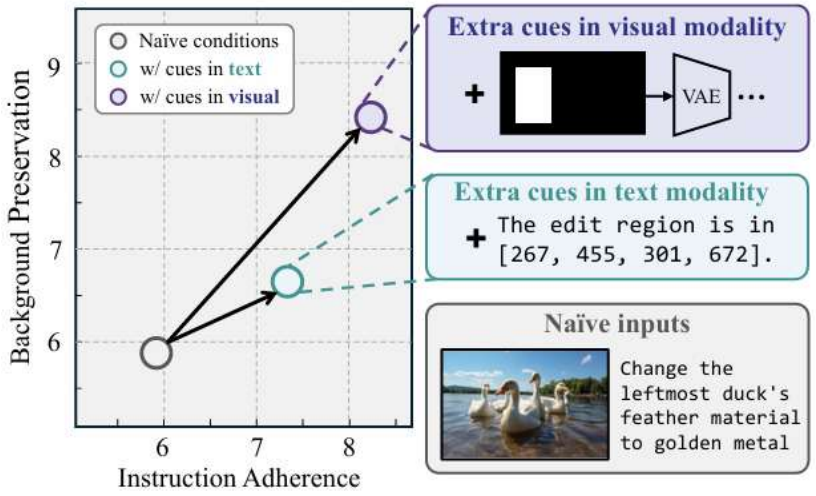
Background

Preliminary Study: Cue Modality

- How to provide spatial cues to editing models?
  - Text modality — bounding box coordinates
  - Visual modality — binary mask overlay

- Visual cues yield much better instruction adherence and background preservation

→ Motivates Generative Visual CoT



Outline

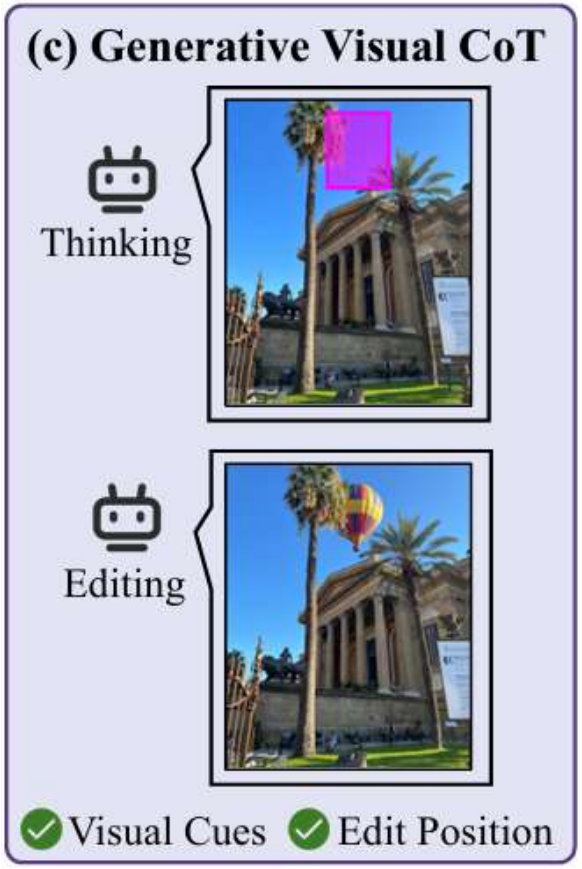
- Authors
- Background
- Methods
- Experiments
- Conclusion

Methods

Generative Visual Chain-of-Thought

- Step 1: generate visual thought $\mathbf{x}_{cot}$ — masks drawn on source
- Step 2: generate edited image $\mathbf{x}_{edit}$ — using $\mathbf{x}_{cot}$ as condition

- Both steps in a single autoregressive sequence
- Trained end-to-end with rectified-flow loss



Methods

GVCoT Formulation

$$\mathbf{x}_{cot} = f_\theta(\mathbf{x}_{src}, \mathbf{t})$$
$$\mathbf{x}_{edit} = f_\theta(\mathbf{x}_{src}, \mathbf{t}, \mathbf{x}_{cot})$$

- $\mathbf{x}_{src}$: source image
- $\mathbf{t}$: text instruction
- $\mathbf{x}_{cot}$: visual thought (mask overlay)
- $\mathbf{x}_{edit}$: final edited image
- f_θ : unified model (Bagel-based)

Methods

Two-Phase Training Recipe

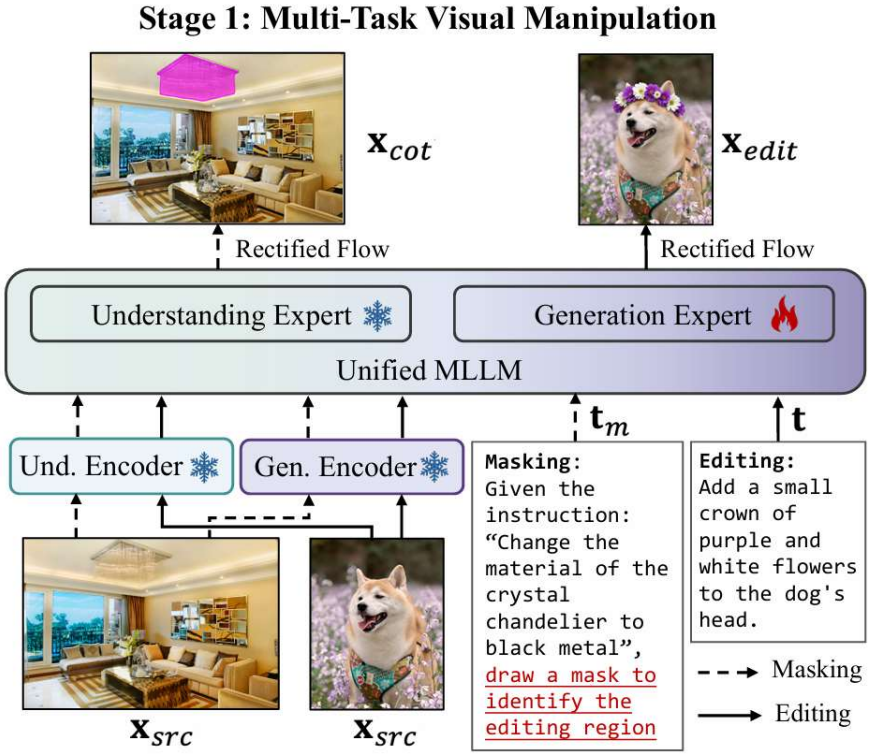
- Phase 1: Progressive Supervised Fine-Tuning (SFT)
  - Step 1 — Multi-task visual manipulation
  - Step 2 — Visual reason-aided editing

- Phase 2: Reinforcement-based Refining (Flow-GRPO)
  - Step 1 — Refine reasoning with verified rewards
  - Step 2 — Refine editing with MLLM-as-Judge rewards

Methods

SFT Stage 1: Multi-Task Visual Manipulation

- Goal: inject "masking" skill into generation expert
- Multi-task objective:
  - Masking — generate $\mathbf{x}_{cot}$ from $\mathbf{x}_{src}$+ masking instruction $\mathbf{t}_m$
  - Editing — generate $\mathbf{x}_{edit}$ from $\mathbf{x}_{src}$+ edit instruction $\mathbf{t}$
- Freeze understanding expert to prevent forgetting



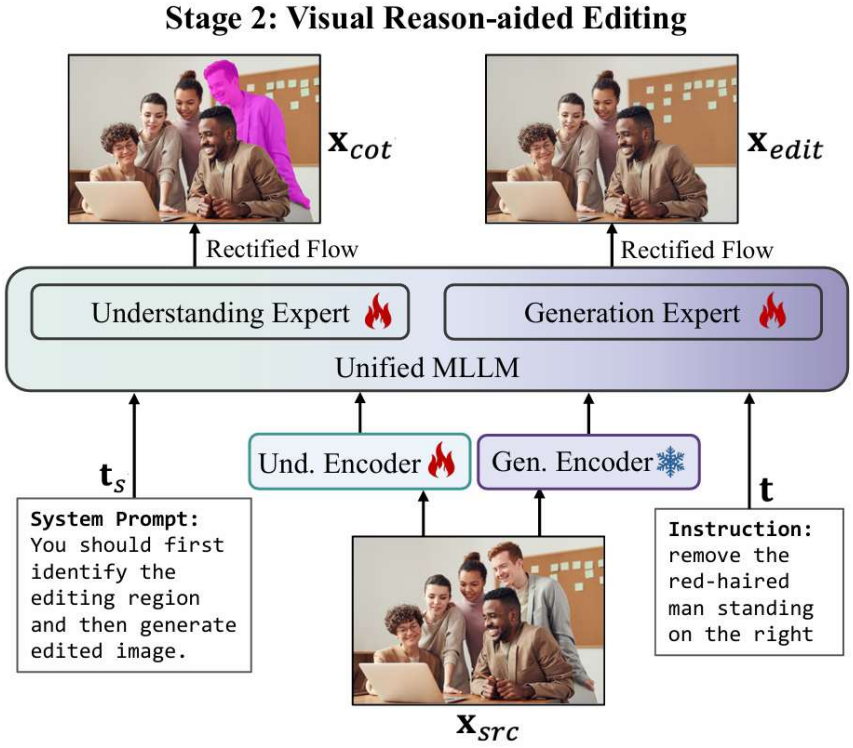
Methods

SFT Stage 2: Visual Reason-aided Editing

- Goal: integrate reasoning + editing in one sequence

- Generate $\mathbf{x}_{cot}$, then $\mathbf{x}_{edit}$, sequentially
- All components unfrozen except VAE encoder
- Teacher-forcing visual thought used for editing supervision

- Loss covers both reasoning and editing flow-matching



Methods

RL Stage 1: Visual Reasoning Rewards

- Refine the localization quality of $\mathbf{x}_{cot}$
- Format Reward
  - Binary classifier separates "thought" vs "edit" frames
  - Ensures model follows reasoning → editing order
- IoU Reward
  - Predicted mask = pixel-wise diff ($\mathbf{x}_{src}$, $\mathbf{x}_{cot}$)
  - IoU against ground-truth edit-region mask

Methods

RL Stage 2: Editing with MLLM-as-Judge

- Refine the final edited image with MLLM scores

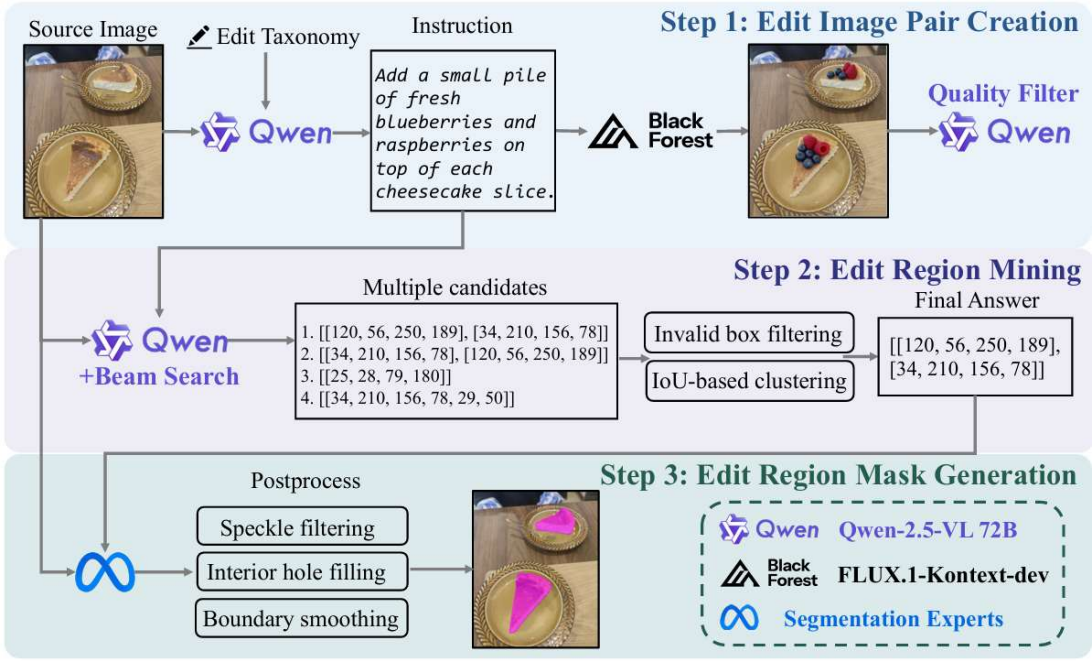
- CoT-Edit Consistency Reward
  - Faithfully translate teacher-forcing thought into edit
  - Scored by Qwen2.5-VL-72B

- Image Quality Reward
  - Visual realism, naturalness, lack of artifacts
  - Also scored by Qwen2.5-VL-72B

Methods

GVCoT-Edit-Instruct Data

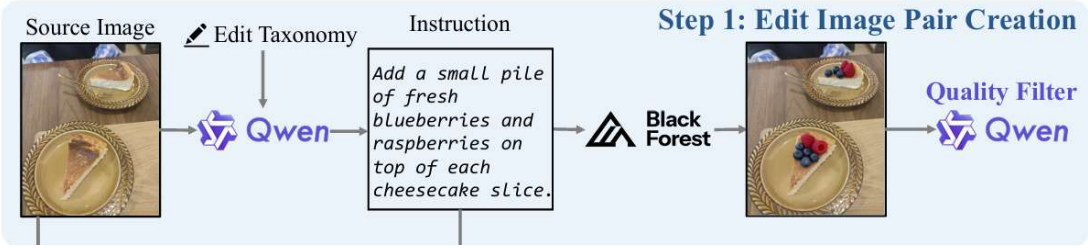
- Quadruple per sample: source / instruction / region annotation / target
- Three-step scalable pipeline:
  - Edit image pair creation
  - Edit region mining
  - Edit region mask generation



Methods

GVCoT-Edit-Instruct Data

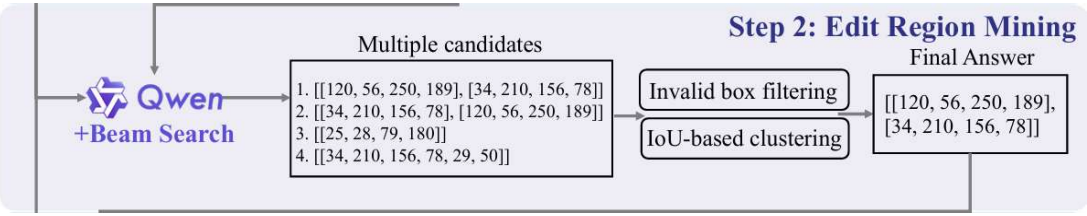
- Three-step scalable pipeline:
  - Edit image pair creation
    - Qwen2.5-VL → natural user-style instructions
    - FLUX.1 Kontext-Dev → synthesize edited images
    - MLLM-based verifier → quality filter
  - Edit region mining
  - Edit region mask generation



Methods

GVCoT-Edit-Instruct Data

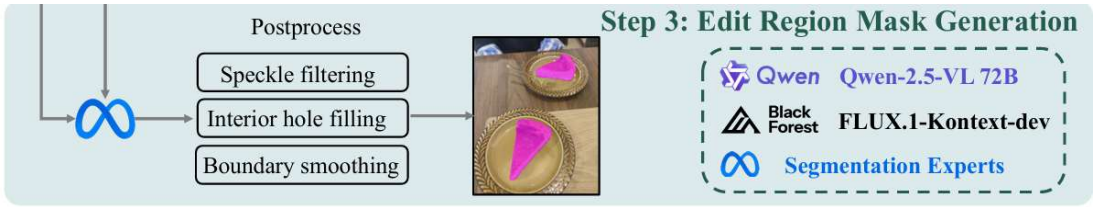
- Three-step scalable pipeline:
  - Edit image pair creation
  - Edit region mining
    - Qwen2.5-VL predicts bbox coordinates
    - Multiple candidates via beam search
    - Filter invalid boxes + IoU-based clustering
  - Edit region mask generation



Methods

GVCoT-Edit-Instruct Data

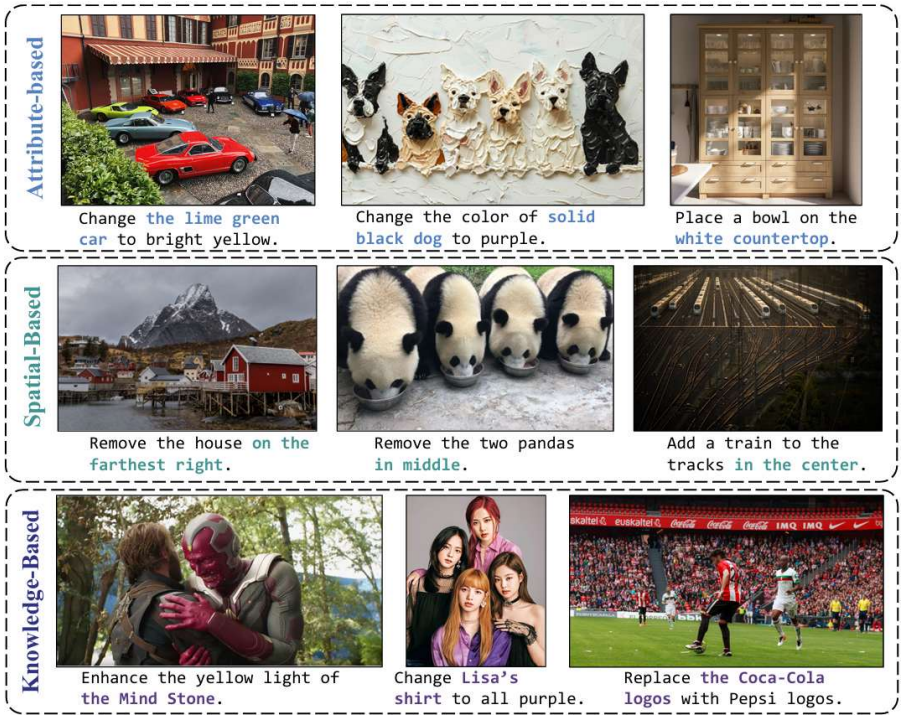
- Three-step scalable pipeline:
  - Edit image pair creation
  - Edit region mining
  - Edit region mask generation
    - Insertion: directly use box mask (boundary unknown)
    - Modification / removal: SAM2 + BiRefNet for instance mask
    - Post-process: hole filling, speckle removal, smoothing



Methods

SREdit-Bench

- Two design principles:
  - Sophisticated scenes — multi-entity, non-object-centric
  - Fine-grained referring — spatial / property / knowledge

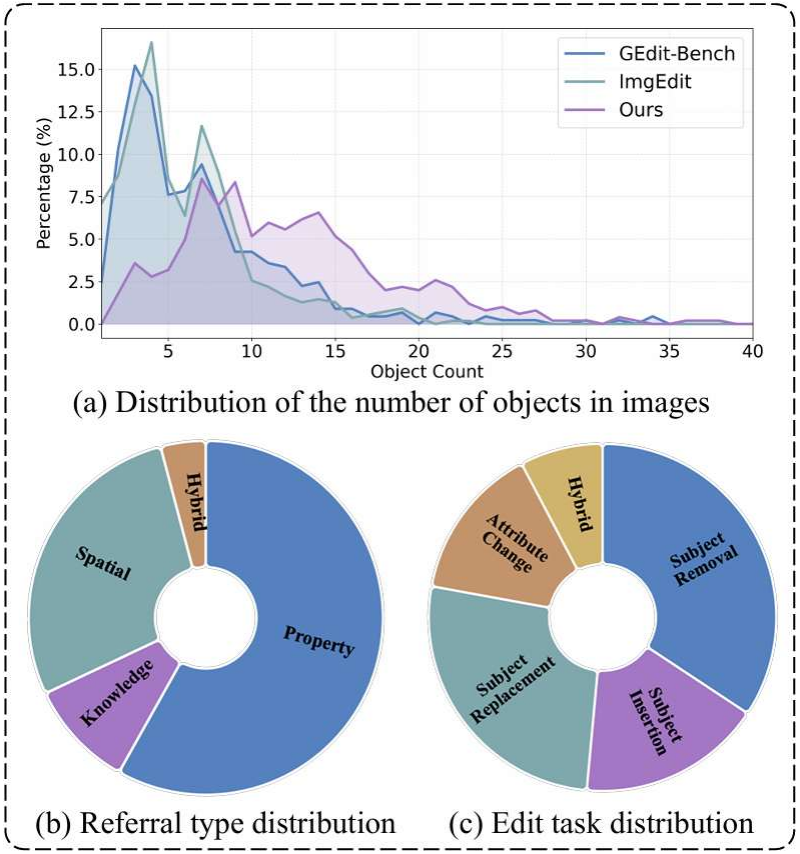


Methods

SREdit-Bench

- More objects per scene
- Referral types and edit tasks balanced

- Three referral modes:
  - Spatial — explicit location ("on the right")
  - Property — appearance ("the yellow one")
  - Knowledge — implicit cues (named entities)



(a) Distribution of the number of objects in images

(b) Referral type distribution (c) Edit task distribution

Outline

- Authors
- Background
- Methods
- Experiments
- Conclusion

Experiments

Setup

- Base model: BAGEL-7B-MoT
- 64 × NVIDIA H20 GPUs

- Compared with 17 editing methods:
  - GPT Image 1, FLUX.1 Kontext Pro, Qwen-Image, Step1X-Edit, OmniGen2, …

Experiments

Quantitative Results on SREdit-Bench

- Bagel-GVCoT achieves Overall = 8.53
- Best among open-source models
  - Surpasses Qwen-Image (8.32) — diffusion specialist
  - Surpasses FLUX.1 Kontext Pro (8.41) — closed-source

Model	SREdit-Bench ↑		
	SC_g	PQ_g	O_g
Product-level models			
GPT Image 1 [High] [40]	9.02	8.42	8.56
FLUX.1 Kontext [Pro]	8.69	8.40	8.41
Qwen-Image [54]	8.57	8.43	8.32
Generation Only			
Instruct-Pix2Pix [4]	2.10	6.40	2.58
MagicBrush [61]	2.99	5.91	3.42
ICEdit [63]	4.41	7.71	4.81
OmniGen [56]	7.05	7.38	6.68
FLUX.1 Kontext [Dev] [27]	7.52	8.17	7.27
Step1X-Edit [35]	6.96	7.87	6.98
Unified Understanding and Generation			
GoT [11]	5.78	7.59	5.83
Bagel [9]	8.02	7.90	7.75
Bagel-think [9]	<u>8.13</u>	8.01	<u>7.82</u>
UniWorld-v1 [31]	5.92	<u>8.47</u>	6.02
OmniGen2 [55]	6.52	7.54	6.44
Ming-UniVision [21]	6.05	6.85	5.96
Bagel-GVCoT (Ours)	**8.87**	**8.76**	**8.53**
Δ Over Base Model	+0.85	+0.86	+0.78

Experiments

Quantitative Results on ImgEdit

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall ↑
Product-level models										
FLUX.1 Kontext [Pro] [27]	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
GPT Image 1 [High] [40]	4.61	4.33	2.90	4.35	3.66	4.57	4.93	3.96	4.89	4.20
Qwen-Image [54]	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27
Generation Only										
MagicBrush [61]	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
Instruct-Pix2Pix [4]	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
AnyEdit [60]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit [64]	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
OmniGen [56]	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
ICEdit [63]	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Step1X-Edit [35]	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
FLUX.1 Kontext [Dev] [27]	**4.12**	3.80	2.04	4.22	3.09	3.97	4.51	**3.35**	4.25	3.71
Unified Understanding and Generation										
GoT [11]	3.74	3.06	1.33	2.72	2.46	2.33	3.45	1.77	2.50	2.65
Bagel [9]	3.56	3.31	1.70	3.3	2.62	3.24	4.49	2.38	4.17	3.20
Bagel-think [9]	3.65	3.53	2.03	3.60	3.03	3.45	4.43	2.59	4.22	3.39
UniWorld-V1 [31]	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
OmniGen2 [55]	3.57	3.06	1.77	3.74	3.20	3.57	**4.81**	2.52	**4.68**	3.44
Ming-UniVision [21]	3.55	3.14	1.52	3.25	3.29	2.77	3.99	2.74	3.91	3.06
BLIP3o-NEXT [6]	4.00	3.78	2.39	4.05	2.61	**4.30**	4.64	2.67	4.13	3.62
Bagel-GVCoT (Ours)	4.02	**4.07**	**2.92**	**4.23**	**3.74**	4.16	3.83	2.82	4.48	**3.82**
Δ Over Base Model	+0.46	+0.76	+1.22	+0.93	+1.12	+0.92	-0.66	+0.44	+0.31	+0.62

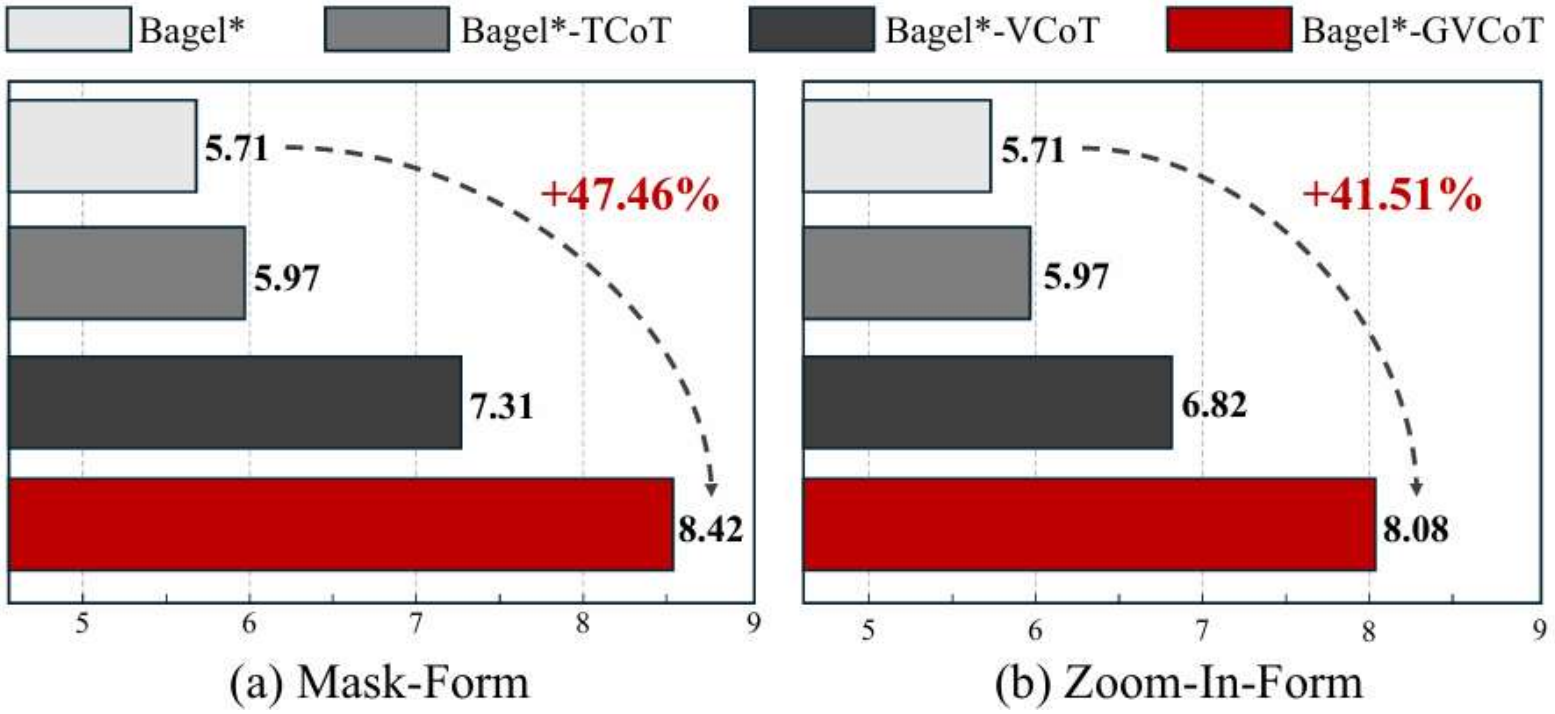
Experiments

Quatitative Results vs. Mask-Based Editing on Human

Method	CLIP-I↑	LPIPS↓	PSNR↑	SSIM↑
SmartEdit [20]	0.8841	0.2915	17.1728	0.6828
BrushNet [24]	0.8986	0.1830	19.2172	0.7877
MagicQuill [37]	<u>0.9381</u>	<u>0.1162</u>	22.2380	**0.8981**
MIND-Edit [52]	0.9310	0.1245	<u>22.2714</u>	0.8517
Bagel [9]	0.9124	0.1721	22.1843	0.8640
Bagel-GVCoT (Ours)	**0.9451**	**0.1066**	**23.0161**	<u>0.8943</u>
Δ Over Base Model	+0.0327	-0.0655	+0.8341	+0.0303

Experiments

Quatitative Results



Experiments

Quatitative Results: GVCoT > VCoT

	Model	IoU↑	SC_g↑	PQ_g↑	O_g↑
	Bagel*	–	6.21	6.09	5.71
	TCoT	–	6.54	6.22	5.97
Mask	VCoT	0.68	7.55	7.92	7.31
Mask	**GVCoT**	0.60	8.53	8.62	8.42
Mask	VCoT w/ TF-thought	–	8.03	7.96	7.75
Mask	**GVCoT w/ TF-thought**	–	8.79	8.95	8.72
Zoom-In	VCoT	–	7.05	7.14	6.82
Zoom-In	**GVCoT**	–	8.12	8.30	8.08
Zoom-In	VCoT w/ TF-thought	–	7.78	7.67	7.54
Zoom-In	**GVCoT w/ TF-thought**	–	8.39	8.58	8.33

Experiments

Ablations

Step 1	Step 2	IoU↑	SC_g↑	PQ_g↑	O_g↑
✓		0.64	7.92	8.01	7.75
	✓	0.61	8.21	8.33	8.10
✓	✓	**0.66**	**8.53**	**8.62**	**8.42**

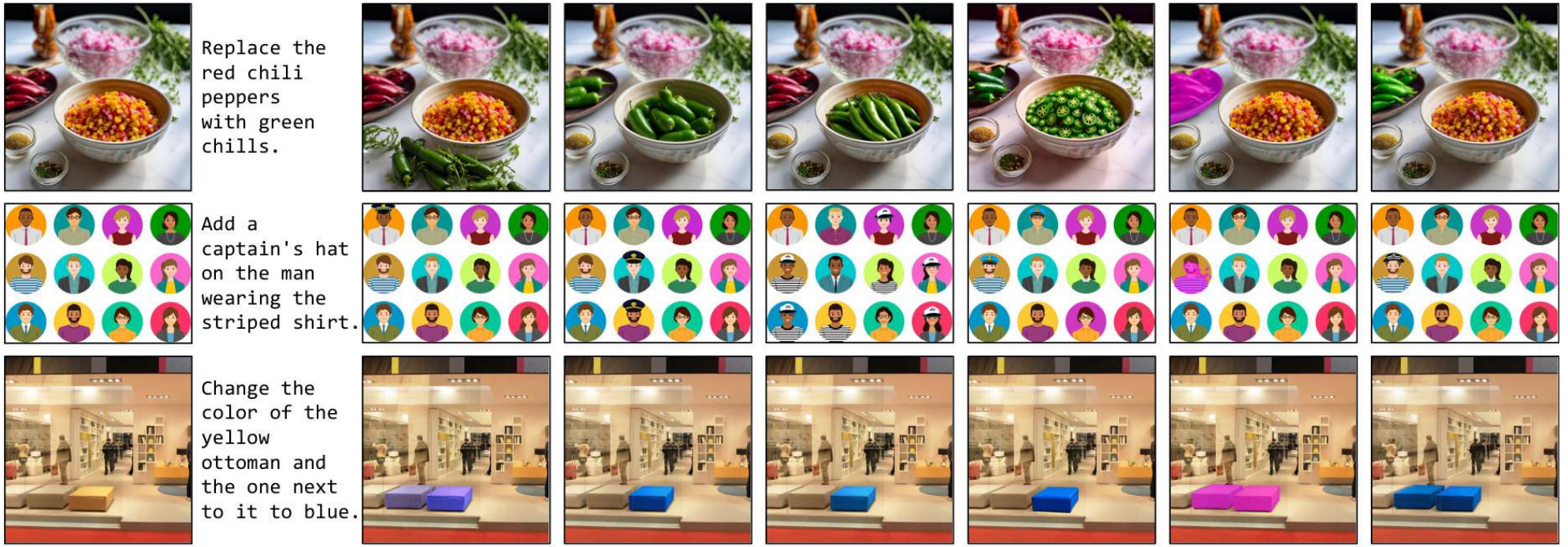
SFT Stage

Setting	IoU↑	SC_g↑	PQ_g↑	O_g↑
SFT Only	0.60	8.53	8.62	8.42
SFT+RL	**0.67**	**8.57**	**8.76**	**8.53**
(1) Multi-stage RL training				
w/o Stage 2	0.67	8.32	8.47	8.25
(2) Visual thought reward design				
Full reward set	**0.67**	**8.32**	**8.47**	**8.25**
w/o IoU reward	0.50	8.32	8.39	8.13
w/o Format reward	0.62	8.13	8.24	8.19
(3) Editing reward design				
Full reward set	**0.67**	**8.57**	**8.76**	**8.53**
w/o CoT-Edit consistency	0.67	8.49	8.63	8.45
w/o Image quality reward	0.67	8.45	8.75	8.49

RL Stage

Experiments

Qualitative Results



Experiments

Qualitative Results

Experiments

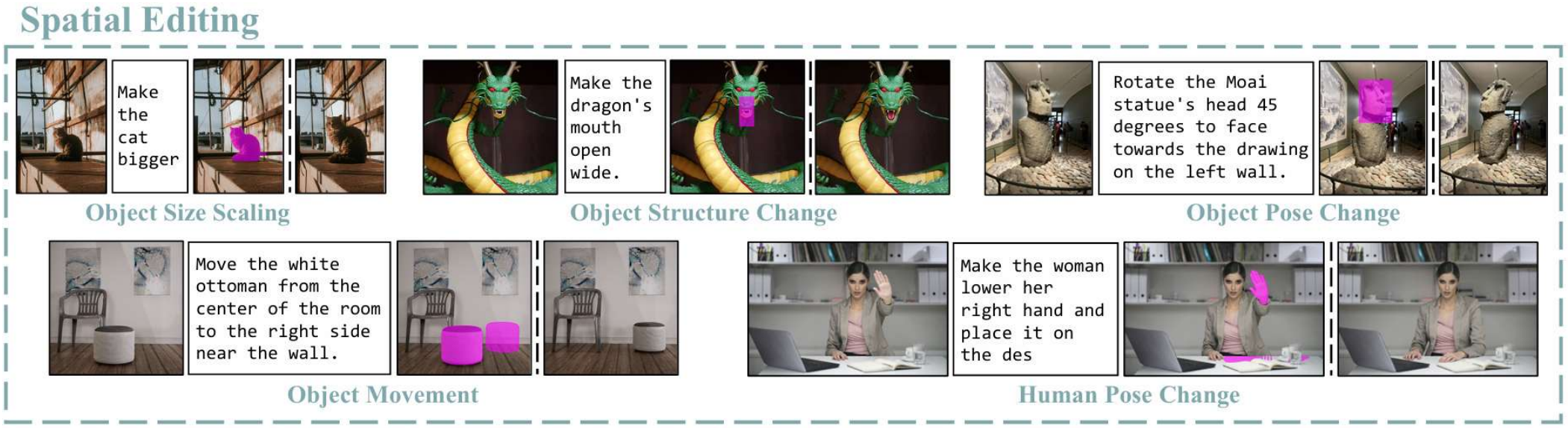
Qualitative Results

Experiments

Qualitative Results

Experiments

Qualitative Results



Experiments

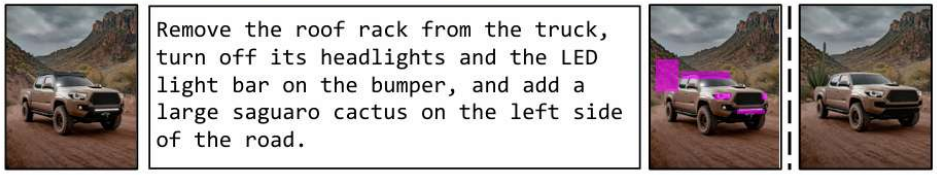
Qualitative Results

Restore

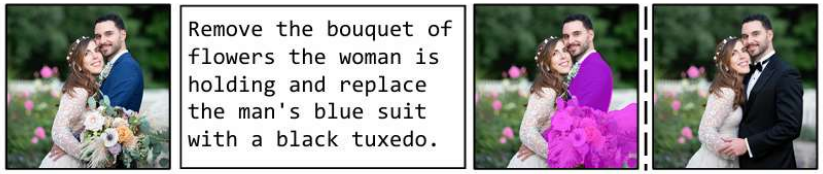
Defact Repair

Watermark Removal

Complex Editing



Multi-Attributes Editing



Multi-Subjects Editing

Outline

- Authors
- Background
- Methods
- Experiments
- Conclusion

Conclusions

Achievements

- Propose GVCoT — native visual reasoning for image editing
- Build GVCoT-Edit-Instruct (1.8M) and SREdit-Bench (590)
- Two-phase SFT + Flow-GRPO RL recipe with multi-dim rewards, SOTA among open-source on SREdit-Bench, ImgEdit, GEdit-Bench

Shortcoming

- Constrained on global style transfer
- Heavy compute

Thank you for Listening!

Methods

Limitation: Global Style Transfer

- "Modify the image style into line art"

- Model treats edit region as local; struggles with whole-image style
- Stylistic transformation may be less pronounced than baseline

- Future direction:
- Combine spatial reasoning with global style mechanism

Methods

Limitation: Global Style Transfer