WORLD LAUREATES FORUM
世界顶尖科学家论坛
国际工程智能大会
2026 International Conference on AI for Engineering
World Laureates Foundation
上海世界顶尖科学家发展基金会
2026
工程智能研究院
INSTITUTE OF AI FOR ENGINEERING

UniCustom: Unified Visual Conditioning for Multi-Reference Image Generation

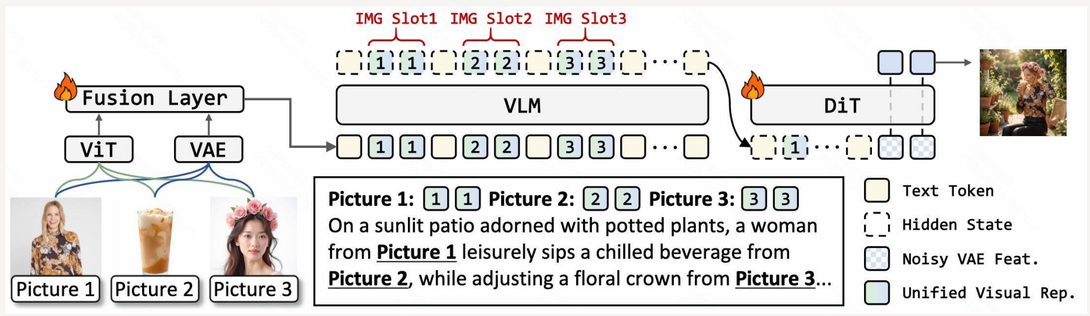
Yiyan Xu[1*], Qiulin Wang[2†‡], Wenjie Wang[1†], Yunyao Mao[2],
Xintao Wang[2], Pengfei Wan[2], Kun Gai[2], Fuli Feng[1]
[1] University of Science and Technology of China [2] KlingAI Research
† Corresponding authors ‡ Project lead * Work done during internship at KlingAI Research

01 ABSTRACT

Multi-reference image generation must identify the intended subjects and preserve their visual attributes. UniCustom unifies semantic grounding and visual binding in a shared representation, improving overall scores on OmniContext and MICo-Bench among the reported open-source models.

02 METHOD

ViT and VAE features are fused into unified visual tokens. A VLM grounds the instruction in these tokens, and a diffusion transformer (DiT) uses the resulting representation for image generation.

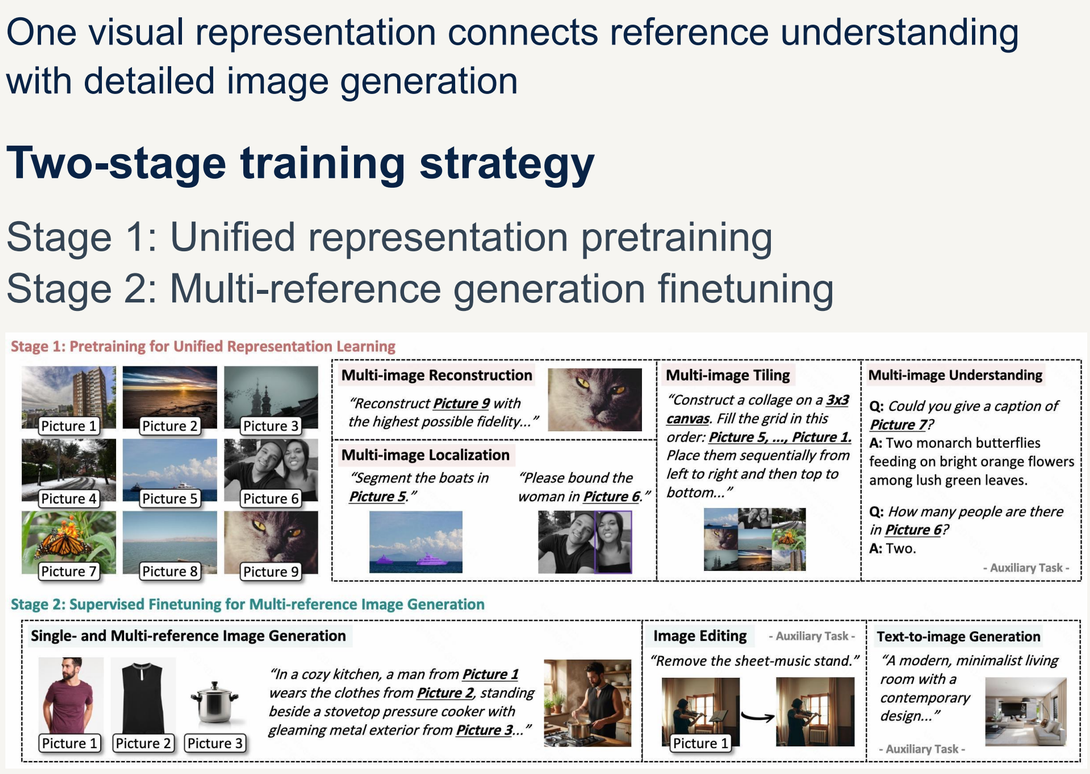


One visual representation connects reference understanding with detailed image generation

Two-stage training strategy

Stage 1: Unified representation pretraining

Stage 2: Multi-reference generation finetuning



03 RESULTS

Best overall scores among the open-source models reported on both benchmarks

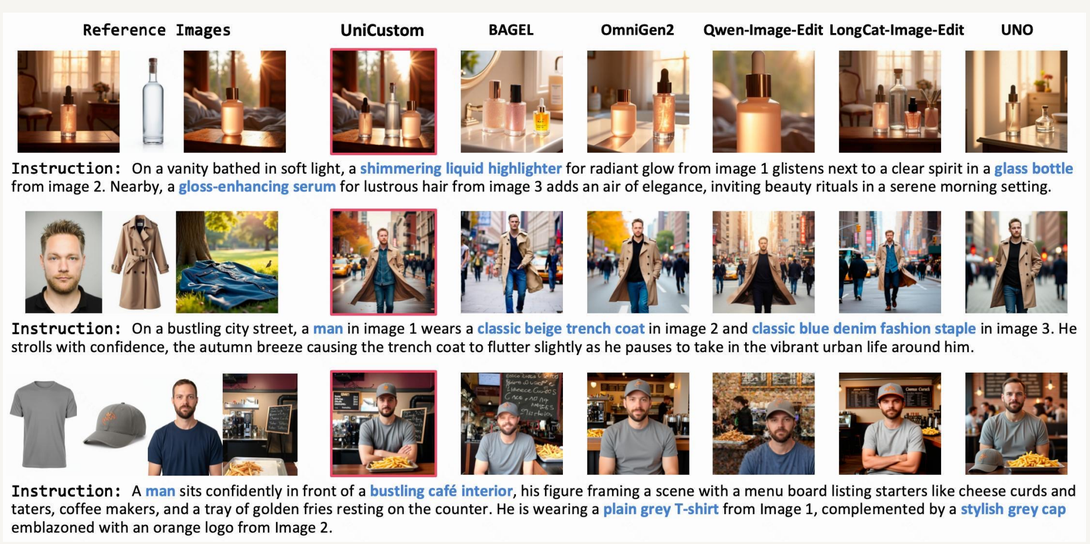
Overall benchmark scores (higher is better)

Method	OmniContext Average	MICo-Bench Overall
Closed-source models		
GPT-Image-2	9.24	61.78
Nano Banana 2	8.92	67.42
Open-source models		
FLUX-Kontext [dev]	4.62	12.51
UNO	6.19	32.43
USO	5.95	27.37
BAGEL	6.45	31.62
OmniGen2	7.60	—
Qwen-Image-Edit-2509	7.69	21.67
LongCat-Image-Edit	7.26	22.78
UniCustom (ours)	**7.84**	**41.71**

7.84 vs 7.69 on OmniContext
41.71 vs 32.43 on MICo-Bench

Compared with the strongest other open-source overall score in each benchmark

Qualitative comparison



04 DISCUSSION

Grounding identifies the intended reference. Binding preserves its appearance during generation. Despite stronger open-source overall scores, a gap to the reported closed-source models remains.

05 CONCLUSIONS

Unified visual conditioning links reference understanding and generation. Two-stage training supports multi-reference composition while retaining subject identity and visual attributes.

PAPER & PROJECT

Paper and project page

arXiv Project page

REFERENCE

Yiyan Xu et al. UniCustom: Unified Visual Conditioning for Multi-Reference Image Generation