CoLoGen: Progressive Learning of Concept–Localization Duality for Unified Image Generation

Supplementary Material

A. Related work

A.1. Unified Multi-modal to Image Generation

Recent advancements in multi-modal image generation strive to consolidate diverse generation and editing tasks into unified frameworks, moving beyond task-specific pipelines. Early approaches often relied on distinct encoders or adapters for different conditions. For instance, **ControlNet** [26] and **T2I-Adapter** [13] introduced extensive external modules to guide pre-trained diffusion models. While effective, these methods face scalability issues when expanding to new tasks due to the linear growth of parameters.

To address this, recent works have focused on unified architectures. **Unified-IO 2** [12] and **Janus** [17] demonstrate the power of autoregressive transformers in handling multi-modal inputs and outputs, though often at the cost of inference speed compared to diffusion models. In the diffusion domain, **OmniControl** [16] and **DreamOmni** [22] integrate visual conditions directly into Diffusion Transformers (DiT), achieving spatial alignment with minimal parameter overhead. **OmniGen** [23] and **PixWizard** [10] further push the boundary by treating image generation and editing as a unified sequence generation problem, removing the reliance on external condition encoders entirely. Similarly, **UniReal** [2] treats image generation tasks as discontinuous video frames to capture real-world dynamics.

More recently, **Qwen-Image** [18] presents a large-scale diffusion foundation model emphasizing strong text rendering, multi-task training, and improved semantic–visual consistency for unified generation and editing. **Query-Kontext** [15] decouples multimodal reasoning from high-fidelity synthesis by leveraging a vision-language model to produce contextual query tokens that guide diffusion-based image generation and editing. **Z-Image** [1] proposes an efficient single-stream diffusion transformer that unifies image generation and editing with scalable training, distillation, and accelerated inference.

However, these unified frameworks often struggle with what we identify as the *Concept–Localization Duality*. Tasks like subject-driven generation require rich semantic concept encoding, whereas tasks like layout-to-image generation demand precise spatial structure. Naively training a single unified model often leads to representational conflict, where optimizing for semantic fidelity degrades spatial precision [3]. Unlike these approaches, **CoLoGen** explicitly decouples and progressively weaves these representations, ensuring high performance across both concept-heavy and localization-heavy tasks.

A.2. Parameter-Efficient Composition and LoRA-MoE

Low-Rank Adaptation (LoRA) [8] has become the standard for parameter-efficient fine-tuning. To handle multi-task learning without catastrophic forgetting, recent research has explored Mixture-of-Experts (MoE) architectures combined with LoRA.

In the realm of Large Language Models (LLMs), **Octavius** [3] and **LoRAHub** [8] propose routing mechanisms to dynamically select or compose LoRA modules for unseen tasks. In visual generation, **Mix-of-Show** [7] addresses the challenge of multi-concept personalization by fusing multiple LoRAs, while **ZipLoRA** [14] attempts to merge content and style LoRAs by optimizing their orthogonality. **MoLE** [21] applies a mixture of LoRA experts to select layer-wise adapters dynamically. **ICEdit** [28] enables instruction-based image editing via in-context generation, combining with LoRA-MoE. While relevant, these methods typically employ static merging strategies or route based solely on input domains. They do not account for the evolving nature of representational needs during the diffusion process itself. **CoLoGen** advances this paradigm via our **Progressive Representation Weaving (PRW)**. Instead of static composition, we employ a time-step dependent "Veteran Gate" routing that dynamically balances expert usage. Crucially, our curriculum creates experts specialized specifically for *Concept* versus *Localization*, rather than just arbitrary data subsets, directly addressing the internal duality of generative tasks.

B. More Results

B.1. Controllable Image Generation

We expand our evaluation to recent state-of-the-art models built on stronger backbones (e.g., FLUX and SD3). While prior works typically report results under limited settings, we conduct a comprehensive comparison across both *Canny* and *Depth* conditions. As shown in Tab. 1, our method consistently achieves the best overall performance across different metrics.

B.2. Customized Image Generation

We further compare with recent large-scale customized generation methods, including Bagel and OmniGen2, on

Add a pair of wire-rimmed glasses to the man.

Turn the color of golden ring to be white.

Remove the person.

Change the setting to spring with blooming flowers.

Make Mario Laugh.

Change this into a Van Gogh painting.

Figure 1. Instructional editing results of our CoLoGen. Our method can adapt to various types of instructions, faithfully follow instructions while preserving the visual consistency of the input images, ensuring high-quality and coherent results.

Method	Base	Canny		Depth	
		C-S↑	FID↓	SSIM↑	FID↓
UNIC-Adapter [6]	SD3	–	23.47	31.10	–
RealGen [5]	CogV	–	**17.50**	35.0	23.40
OmniControl [24]	FLUX	30.60	20.63	39.0	27.26
EasyControl [27]	FLUX	28.60	–	35.9	20.39
CoLoGen(Ours)	FLUX	**33.31**	18.20	**40.1**	**19.56**

Table 1. **Controllable image generation comparison** on recent backbone models.

Method	Data	DINO	C-I	C-T
OmniControl [24]	200k	0.684	0.799	0.312
FLUX-IP-Adapter [25]	200k	0.582	0.820	0.288
CoLoGen(Ours)	200k	**0.714**	0.825	**0.315**
UNO-FLUX [20]	1M–5M	0.760	0.835	0.308
OmniGen2 [19]	10M+	0.749	0.830	0.314
BAGEL [4]	10M+	0.797	**0.859**	0.307

Table 2. **Customized image generation comparison** under different training data scales.

Subject-200k. Notably, these approaches are trained on substantially larger datasets (10M+ samples), whereas our method uses fewer than 1M samples. As reported in Tab. 2, CoLoGen achieves competitive or superior performance despite the significantly smaller training scale.

B.3. Image Editing Benchmark

We additionally evaluate on the recent **GEdit-Bench** full set. As shown in Tab. 3, CoLoGen achieves the best G_SC score and remains competitive across other editing quality metrics, demonstrating strong generalization ability in image editing tasks.

Method	**GEdit-Bench (Full Set)** ↑		
	G_SC	G_PQ	G_O
Step1X-Edit [11]	7.66	7.35	6.97
BAGEL [4]	7.36	6.83	6.52
FLUX.1 Kontext [9]	7.02	7.60	6.56
Qwen-Image [18]	8.00	**7.86**	**7.56**
CoLoGen (Ours)	**8.03**	7.15	7.31

Table 3. **Results on GEdit-Bench (Full Set)**.

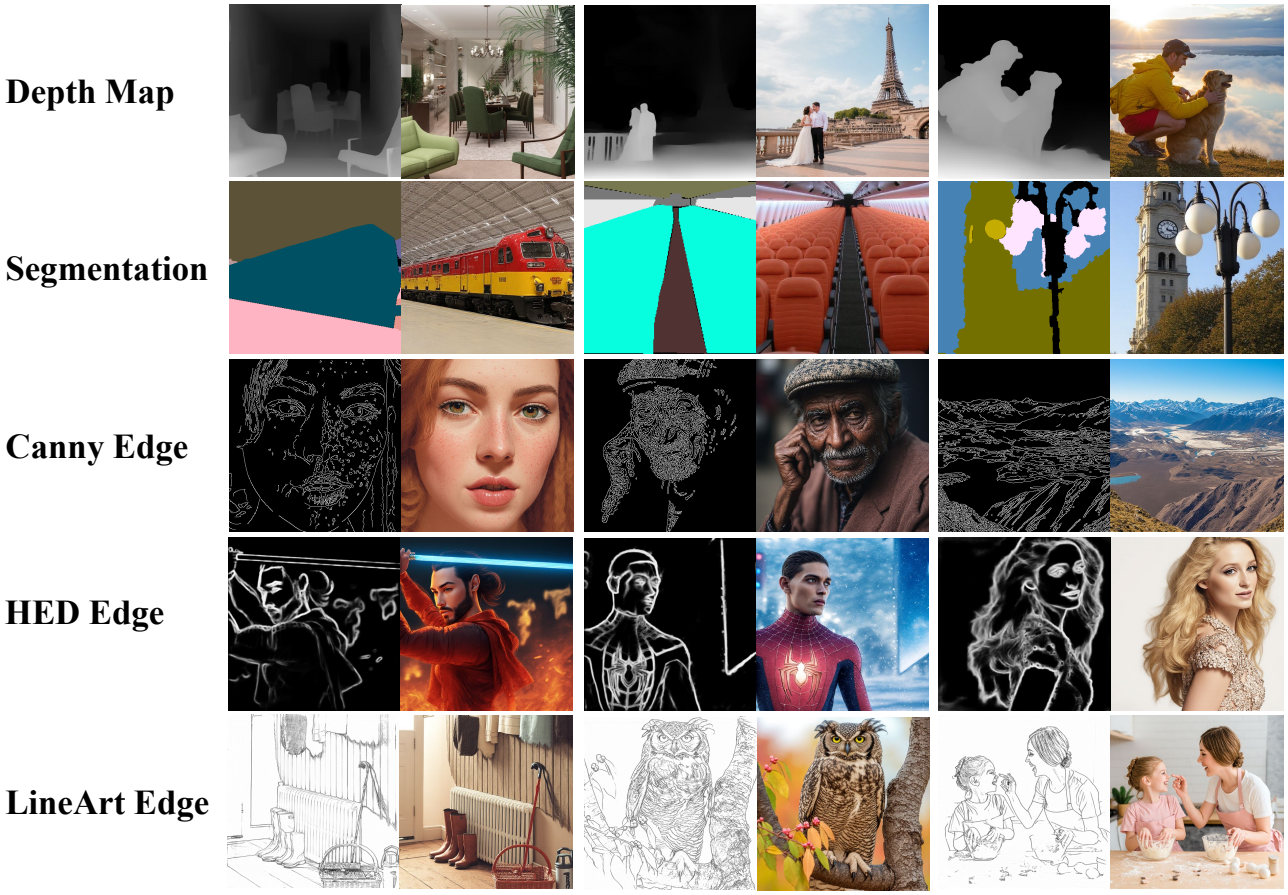

Figure 2. Controllable generation results of our CoLoGen.

Figure 3. Visual examples from the image grounding task demonstrate that CoLoGen, after undergoing endogenous pre-training, exhibits highly accurate visual localization capabilities.

C. Visualization

C.1. Instruction Editing

We provide expanded visual examples on the Instruction Editing benchmark in Figure 1. The results demonstrate CoLoGen's versatility in handling diverse editing instructions, ranging from localized object manipulation to global stylistic changes. These results validate that our Instruction-Image Alignment stage effectively fine-tunes the synergy between concept and localization representations.

C.2. Controllable Image Generation

Figure 2 showcases CoLoGen's performance on the Controllable Image Generation benchmark under various spatial conditions, including Depth maps, Segmentation masks, Canny edges, HED edges, and LineArt. The visualization highlights the effectiveness of the *Localization Representation* (R_l) acquired during the endogenous pre-training.

Image Grounding. CoLoGen acquires precise intent localization capabilities for the Image Grounding task during endogenous pre-training. The visualization in Fig. 3 demonstrates that the model possesses robust object perception

abilities and can accurately detect the referring instance, significantly enhancing its stability on complex tasks (e.g., instruction-based editing and customized generation).

References

[1] Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. arXiv preprint arXiv:2511.22699, 2025. 1

[2] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. arXiv preprint arXiv:2412.07774, 2024. 1

[3] Zeren Chen, Ziqin Wang, Zhen Wang, Huayang Liu, Zhenfei Yin, Si Liu, Lu Sheng, Wanli Ouyang, Yu Qiao, and Jing Shao. Octavius: Mitigating task interference in mllms via lora-moe. arXiv preprint arXiv:2311.02684, 2023. 1

[4] Anne de Jong, Sacha AFT van Hijum, Jetta JE Bijlsma, Jan Kok, and Oscar P Kuipers. Bagel: a web-based bacteriocin genome mining tool. Nucleic acids research, 34(suppl_2): W273–W279, 2006. 2

[5] Wenhao Ding, Yulong Cao, Ding Zhao, Chaowei Xiao, and Marco Pavone. Realgen: Retrieval augmented generation for controllable traffic scenarios. In European Conference on Computer Vision, pages 93–110. Springer, 2024. 2

[6] Lunhao Duan, Shanshan Zhao, Wenjun Yan, Yinglun Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, Mingming Gong, and Gui-Song Xia. Unic-adapter: Unified image-instruction adapter with multi-modal transformer for image generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 7963–7973, 2025. 2

[7] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wengang Xiao, Rui Zhao, and Ying Shan. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. NeurIPS, 2023. 1

[8] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. arXiv preprint arXiv:2307.13269, 2023. 1

[9] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742, 2025. 2

[10] Weifeng Lin, Xinyu Wei, Renrui Zhang, Le Zhuo, Shitian Zhao, Siyuan Huang, Huan Teng, Junlin Xie, Yu Qiao, Peng Gao, et al. Pixwizard: Versatile image-to-image visual assistant with open-language instructions. arXiv preprint arXiv:2409.15278, 2024. 1

[11] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761, 2025. 2

[12] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. arXiv preprint arXiv:2312.17172, 2023. 1

[13] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In AAAI, pages 4296–4304, 2024. 1

[14] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, and Varun Li, Yuanzhen andJampani. Ziplora: Any subject in any style by effectively merging loras. arXiv preprint arXiv:2311.13600, 2023. 1

[15] Yuxin Song, Wenkai Dong, Shizun Wang, Qi Zhang, Song Xue, Tao Yuan, Hu Yang, Haocheng Feng, Hang Zhou, Xinyan Xiao, et al. Query-kontext: An unified multimodal model for image generation and editing. arXiv preprint arXiv:2509.26641, 2025. 1

[16] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. arXiv preprint arXiv:2411.15098, 3, 2024. 1

[17] Chengyue Wu and et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. arXiv preprint arXiv:2410.13848, 2024. 1

[18] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. arXiv preprint arXiv:2508.02324, 2025. 1, 2

[19] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871, 2025. 2

[20] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 18682–18692, 2025. 2

[21] Xun Wu and et al. Mole: Mixture of lora experts. ICLR, 2024. 1

[22] Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamomni: Unified image generation and editing. arXiv preprint arXiv:2412.17098, 2024. 1

[23] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. arXiv:2409.11340, 2024. 1

[24] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. arXiv preprint arXiv:2310.08580, 2023. 2

[25] XLabs-AI. Flux-ip-adapter model card. https://huggingface.co/XLabs-AI/flux-ip-adapter, 2024. 2

[26] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. ICCV, 2023. 1

[27] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 19513–19524, 2025. 2

[28] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. Enabling instructional image editing with in-context generation in large scale diffusion transformer. In The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025. 1