VisualCloze: A Universal Image Generation Framework via Visual In-Context Learning

Supplementary Material

Task	Link
semantic-invariant style transfer	Fig. 3
semantic-variant style transfer	Fig. 4
different types of dense prediction	Fig. 5
semantic-invariant image editing	Fig. 6
semantic-variant image editing	Fig. 7
image resolution under various degradations	Fig. 8

Table 1. Generation results on more tasks.

1. Instruction Format

In our unified framework, the instruction consists of three parts: (1) layout instruction, which describes the layout of the grid image, (2) task instruction, which specifies the task type, and (3) content instruction, which describes the content of the target image. Fig. 1 illustrates the instructions for concept fusion of style, subject, and layout (Fig. 1 upper) and image editing with reference (Fig. 1 bottom). The content instruction is omitted for some tasks that provide strong visual cues, like style transfer.

2. More Visualizations

In the manuscript, we have provided results on conditional image generation and subject-driven generation. Here, we further show results for the other meta-tasks, as summarized in Tab. 1. In style transfer and dense prediction, we also show the impacts of in-context learning and further confirm its effectiveness.

3. Fine-tuning FLUX.1-dev Model

Apart from FLUX.1-Fill-dev, we also adapt our method to FLUX.1-dev [3], a common text-to-image generative model. Unlike the infilling model that shares a consistent objective with universal image generation, FLUX.1-dev requires customized modifications to process clean condition images and noise target images. Specifically, after concatenating images in a grid layout like the infilling model, we always keep the region corresponding to the conditions as clean latent embeddings throughout the sampling process. This strategy requires modifications in image sampling because FLUX.1-Fill-dev takes noise latent embeddings as input. Moreover, for the adaLN-Zero block [6], it is critical to calculate the separate mean and shift parameters for the regions of clean conditions and noise target by feeding $T = 0$ and $T = t$ into the adaLN-Zero, respectively. t indicates the timestep in each sampling step and gradually increases from 0 to 1 along the sampling process. This strategy aligns with the pre-training domain of FLUX.1-dev, where different noise levels correspond to different mean and shift. As shown in Fig. 2, this strategy ensures the visual fidelity.

4. Evaluation Metrics

4.1. Conditioning Generation

We assess the models from controllability, quality, and text consistency to evaluate image generation quality in conditioning generation and image restoration tasks.

Controllability. For conditional image generation, we measure the difference between the input conditions and those extracted from generated images. Specifically, we calculate the F1 Score for the cany-to-image task and RMSE for the depth-to-image task. Additionally, for debluring, we measure the RMSE between original and restored images.

Quality. We measure the Generation quality using FID [1], SSIM, MAN-IQA [10], and MAN-IQA [10]. FID [1] measures the similarity between generated and real image feature distributions. SSIM evaluates perceptual quality by comparing luminance, contrast, and structural patterns between images. It calculates local patch statistics and combines them into a composite score ranging from -1 to 1, with higher values indicating better structural preservation. MANIQA [10] and MUSIQ [2] leverage neural networks to predict image quality scores.

Text consistency. Leveraging the powerful multi-modal capability of CLIP [7], we also measure the semantic alignment between generated images and text prompts, which reflects how the model accurately follows instructions.

4.2. Subject Driven Generation

Following DreamBooth [8] and BLIP-Diffusion [4], we measure DINOv2 [5], CLIP-I [7], and CLIP-T scores for the comparison of subject-driven image generation. DINOv2 [5] and CLIP-I scores measure the alignment between the reference subject and generated images through cosine similarity and CLIP score, respectively. CLIP-T measures the alignment between the generated image and the corresponding text prompt.

4.3. Style Transfer

Following StyleDrop [9], we assess the performance of style transfer according to text consistency and style align-

ment. For text alignment, we measure the cosine similarity between embeddings of generated images and text prompts, where the embeddings are extracted by CLIP [7]. Regarding style consistency, we measure the cosine similarity between embeddings of generated images and style reference. Note that these two metrics should be considered together because the style consistency will reach 1.0 if the model collapses, where the model completely copies style reference as a composite image and ignores text instructions.

(a) Concatenated images

Layout instruction:

12 images are organized into a grid of 3 rows and 4 columns, evenly spaced.

Task instruction:

Each row describes a process that begins with [IMAGE1] white edge lines on black from canny detection, [IMAGE2] Photo with a strong artistic theme, [IMAGE3] a reference image showcasing the dominant object and results in [IMAGE4] High-quality visual with distinct artistic touch.

Content instruction:

∅

Layout instruction:

A 3x3 grid containing 9 images, aligned in a clean and structured layout

Task instruction:

Every row provides a step-by-step guide to evolve [IMAGE1] a reference image with the main subject included, [IMAGE2] an image with flawless clarity into [IMAGE3] a high-quality image.

Content instruction:

The bottom-right corner image presents: A glossy gel nail polish bottle. At the edge of a bustling city park, this item rests on vibrant green grass, captured with a subtle bokeh effect as joggers and pets move in the background.

(b) Language instructions

Figure 1. Examples of language instructions that contain prompts about the layout of the concatenated image, task intent, and content of the target image.

(a) separate mean and shift (b) unified mean and shift

Figure 2. Effects of separate mean and shift in fine-tuning FLUX.1-dev.

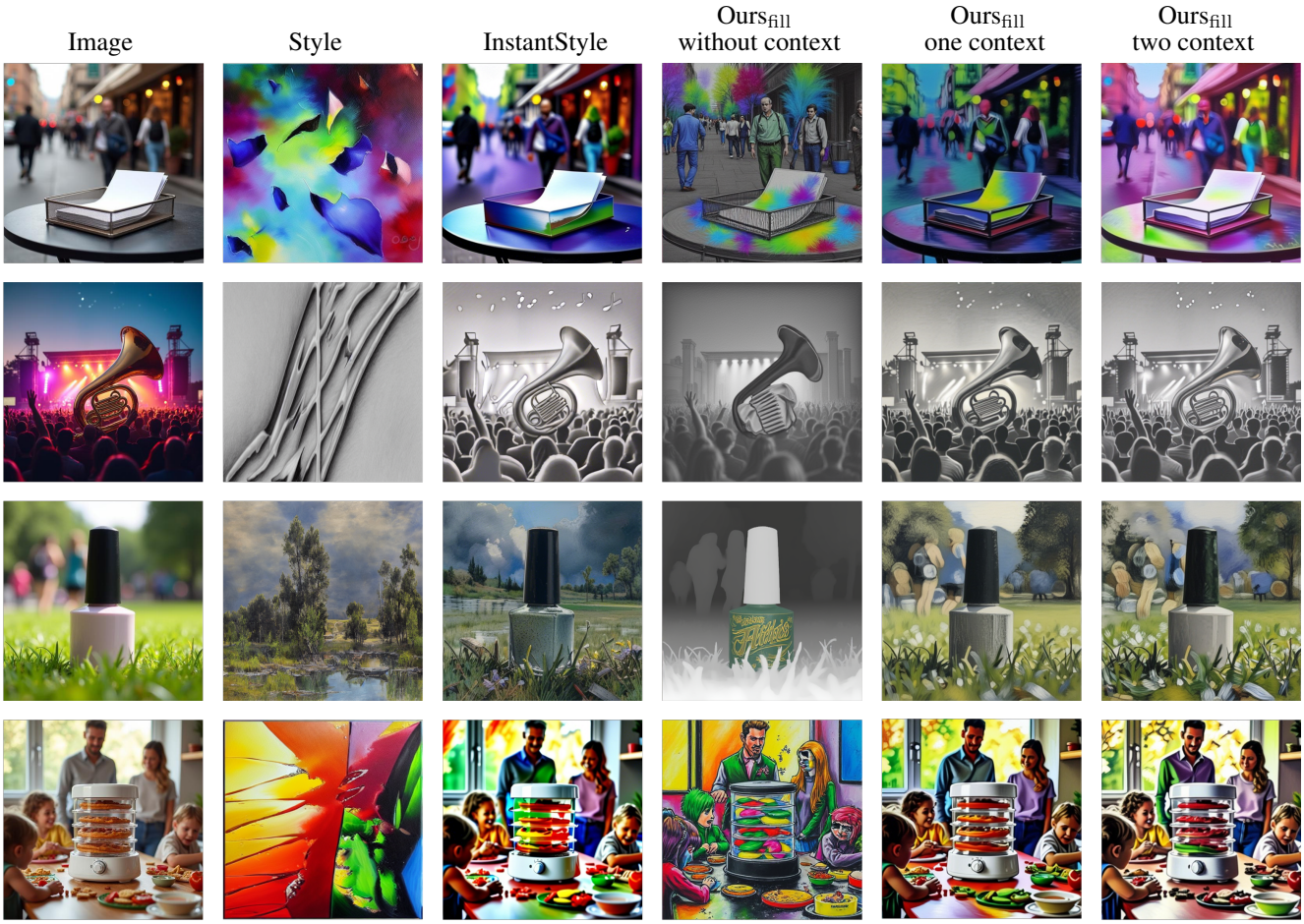

Figure 3. Semantic-invariant style transfer.

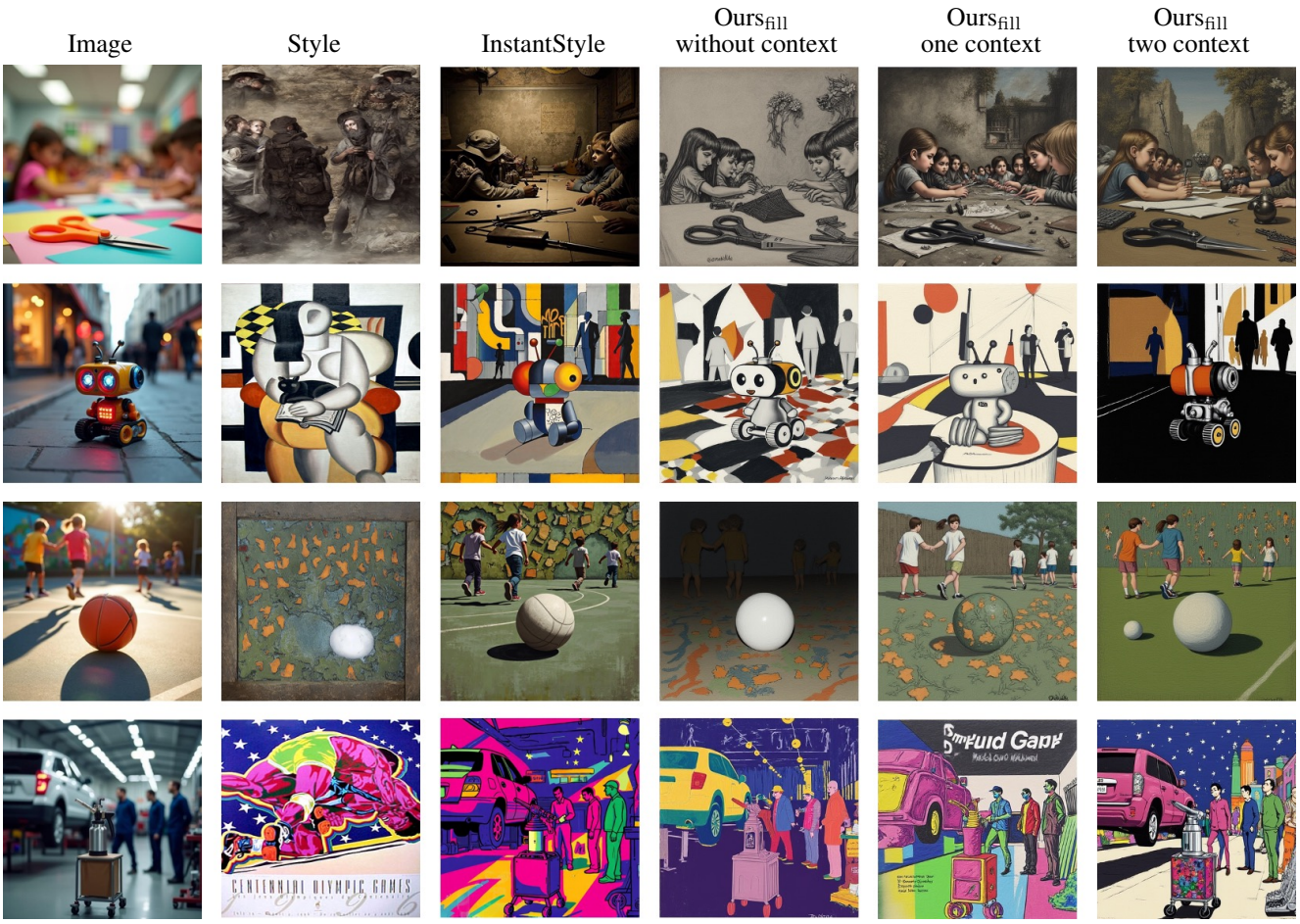

Figure 4. Semantic-variant style transfer.

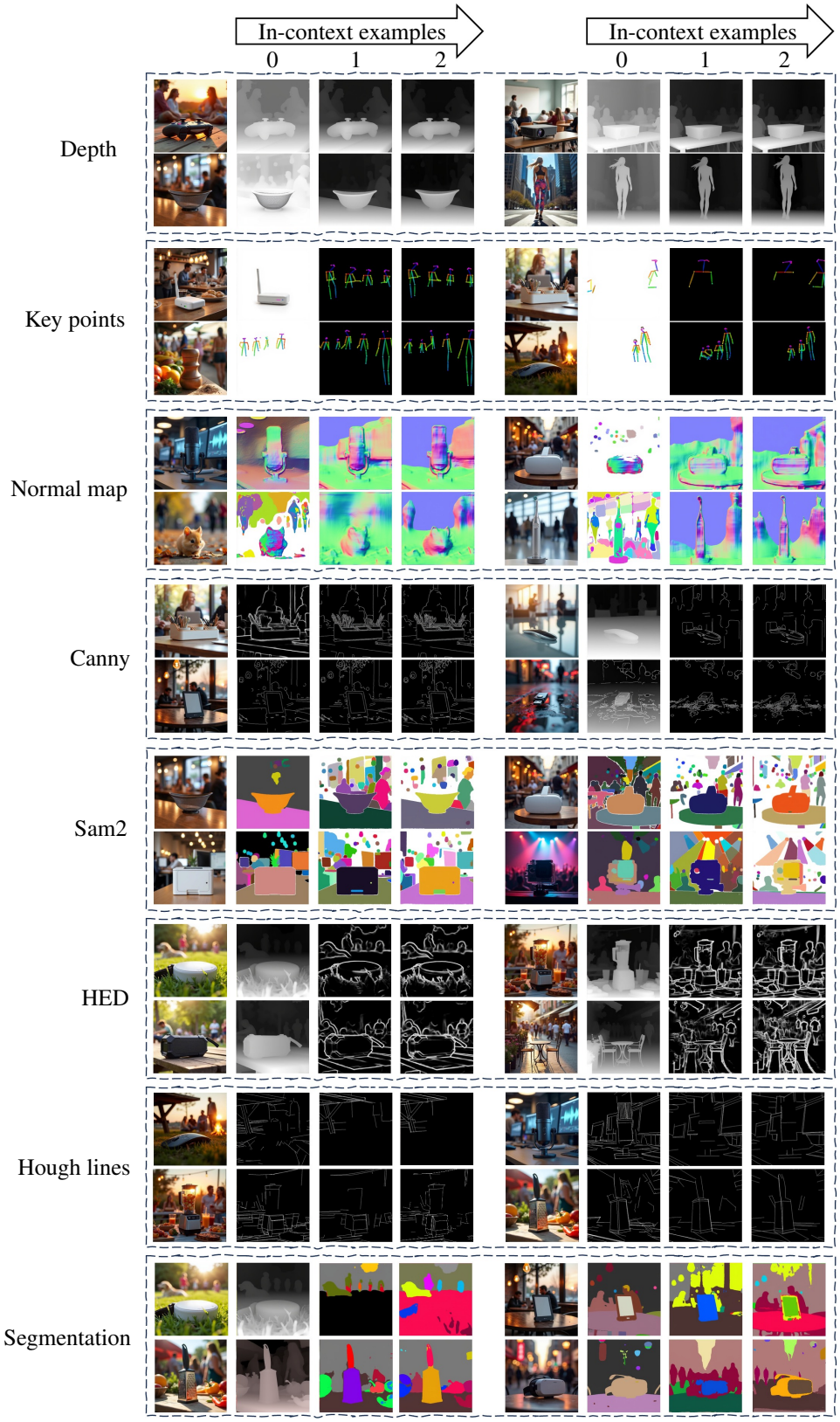

Figure 5. Dense prediction.

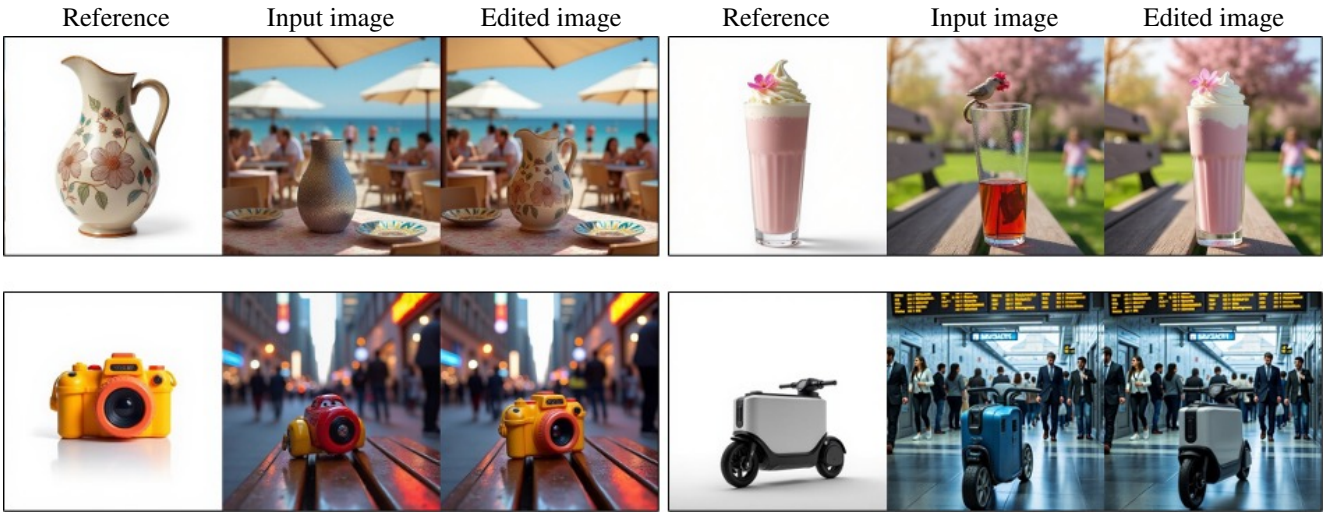

Figure 6. Semantic-invariant image editing with reference. We do not show the results with in-context examples because the model can achieve correct editing without examples. The three images for each item are reference, input, and output, respectively. We replace the subject in the input image with the reference.

Figure 7. Semantic-variant image editing with reference. We do not show the results with in-context examples because the model can achieve correct editing without examples. The three images for each item are reference, input, and output, respectively. We replace the subject in the input image with the reference.

Figure 8. Image restoration. We show the degraded image on the left and the restored image on the right for each degradation. We do not show the results with in-context examples because the model can achieve correct restoration without examples.

References

[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[2] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 1

[3] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1

[4] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023. 1

[5] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[6] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 1

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2

[8] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1

[9] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 1

[10] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022. 1