ProcessMaker: A Generalized Process Visualization Framework with Adaptive Sequence Steps on Diffusion Transformers

Supplementary Material

1. More Detailed Implementations

1.1. The Technical Details

In the Sparse Masks of stage 1, the mask ratio is 0.7. For domains within dataset, we select the corresponding masks during inference and remove any masks for unseen domains. During stage 1, we separately initialize a mask matrix and a LoRA matrix [3], and train them simultaneously, and the values in mask matrix are continuous. Afterwards, we mask the parameters corresponding to the LoRA matrix that rank in the bottom 70%, and unmask those in the top 30% that are most related to the domain, getting a binarized matrix. Finally, we multiply it by the LoRA matrix.

In the sliding window strategy of stage 2, we set the stride=1 and and the number of frames inserted between each pair does not exceed 3 totally.

1.2. The GPT4-o Evaluations

In the GPT-4o [1] evaluation process, we followed the setup of [5]. Scores are output in JSON format, for example:

{"Alignment": 4, "Coherence": 5}

1.3. The Human Evaluations

We conduct human evaluations on Painting, Icon, Cook, and other 5 domains. To be specific, the other 5 domains are excluded in datasets, including Metalworking, Origami, Embroidery, Flower Arranging, and Latte Art. We select 5 sequences from each domain and compute the average scores as our final results.

2. Supplemental Experiments

2.1. Cases of Conflicted Domains.

When removing masks, generated cases mismatch the style of "oil painting" or "digital icon", reflecting the interference from sketch and other domains.

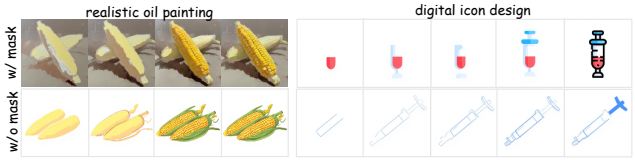


Figure 1. Cases of ablating masks on highly conflicting domains.

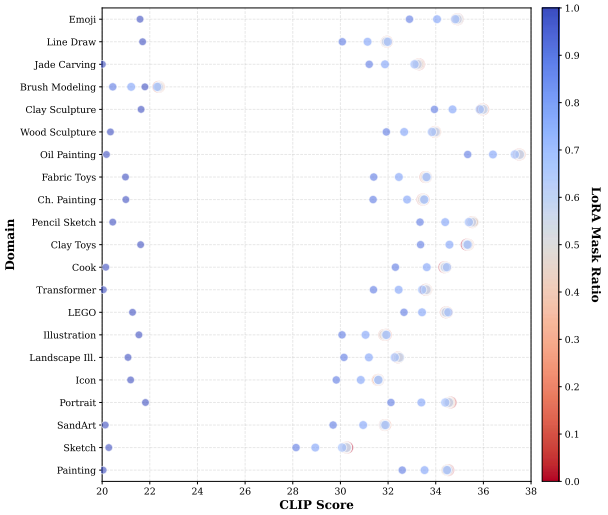


Figure 2. The performance with different mask ratio for LoRA.

2.2. Mask Ratio of LoRA

To explore the influence of different mask ratio of LoRA, we provide the CLIPScore [2] of 21 domains under different mask ratio. As shown in Figure 2, we find that when the mask ratio reaches 70% or less, the change in CLIPScore is not significant. The remaining 30% of parameters in LoRA play an important role in the generated results, while the 70% parameters were redundant. Therefore, based on the above observations, we set a mask rate of 70% for each domain to efficiently utilize the trainable parameters.

Figure 3. Generated procedural sequences of different mask ratio.

In addition, we also compare the generated procedural

Table 1. Ablation studies of other 4 frame domains.

Stage	Settings	Clay Toys		Ink Painting		Jade Carving		Pencil Sketch		Emoji	
		Align	**Cohere**	**Align**	**Cohere**	**Align**	**Cohere**	**Align**	**Cohere**	**Align**	**Cohere**
Stage 1	Base Model	35.25	4.50	33.46	4.90	32.93	4.85	34.44	4.50	34.20	3.60
	w/ LoRA Masks	35.31	4.55	33.52	4.90	33.08	4.85	35.20	4.65	34.75	3.85
	w/ RA	35.27	4.50	33.50	4.90	33.01	4.85	35.14	4.60	34.63	3.75
Stage 2	ProcessMaker	35.34	4.55	33.52	4.90	33.11	4.85	35.39	4.70	34.82	3.90
	w/o Δ_{glob}	35.30	4.55	33.49	4.90	33.10	4.85	35.15	4.65	34.77	3.85
	w/o Δ_{loc}	35.33	4.55	33.50	4.90	32.99	4.85	35.27	4.65	34.61	3.80
	w/o Δ_{sem}	35.27	4.55	33.47	4.90	32.95	4.85	35.01	4.70	34.50	3.80

Table 2. Ablation studies of other 9 frame domains.

Stage	Settings	LEGO		Cook		Portrait		Transformer		Sketch	
		Align	**Cohere**	**Align**	**Cohere**	**Align**	**Cohere**	**Align**	**Cohere**	**Align**	**Cohere**
Stage 1	Base Model	34.40	4.90	34.41	4.25	33.84	5.00	33.03	4.90	29.35	4.70
	w/ LoRA Masks	34.50	5.00	34.46	4.45	34.25	5.00	33.40	4.95	30.01	4.80
	w/ RA	34.46	4.95	34.43	4.40	34.09	5.00	33.37	4.95	29.97	4.75
Stage 2	ProcessMaker	34.53	5.00	34.47	4.50	34.40	5.00	33.43	4.95	30.07	4.85
	w/o Δ_{glob}	34.50	4.95	34.45	4.40	34.26	5.00	33.38	4.90	30.02	4.80
	w/o Δ_{loc}	34.48	4.95	34.45	4.45	34.15	5.00	33.35	4.95	29.98	4.80
	w/o Δ_{sem}	34.48	4.95	34.43	4.50	34.02	5.00	33.24	4.95	29.65	4.85

Table 3. The selection of different DiT layers when introducing the Representation Alignment (RA).

Stage	Alignment	Coherence
ProcessMaker	33.61	4.95
w/ RA (4, 8, 14)	32.80	4.60
w/ RA (12, 29, 34)	32.91	4.65
w/ RA (4, 8, 12, 19)	32.82	4.65
w/ RA (2, 9, 19, 30)	33.26	4.80
w/ RA (0, 7, 14, 23, 34)	32.85	4.75
w/ RA (2, 6, 12, 22, 30)	33.37	4.80
w/ RA (4, 9, 14, 19, 29)	**33.61**	**4.95**

Table 4. The ablation experiments of different α and β values, we evaluate the Alignment by CLIP.

$\alpha \backslash \beta$	0	0.2	0.4	0.6	0.8	1.0
0	32.35	33.02	33.10	32.93	32.87	32.85
0.2	32.54	33.25	33.14	33.05	32.96	–
0.4	32.60	33.42	33.54	33.28	–	–
0.6	32.71	**33.61**	33.59	–	–	–
0.8	32.58	33.50	–	–	–	–
1.0	32.50	–	–	–	–	–

sequences of different mask ratio in Figure 3. The results demonstrate that mask ratio plays a crucial role in balancing frame coherence and visual details. When the mask ratio is below 0.7, it effectively preserves the global structure while enhancing local texture and smooth transition of frames. However, when the ratio exceeds 0.7, the model will be overly constrained, resulting in a surface that is too smooth and a decrease in structural diversity. Therefore, a mask ratio of 0.7 achieves the tradeoff between visual details and parameter efficiency, and is used as default setting.

2.3. Ablation Studies

We present the supplementary ablation studies of key components on more domains in Table 1 and Table 2. In addition, we provide the results of applying the Representation Alignment (RA) to different MM-DiT Blocks and Single-DiT Blocks in Table 3. Furthermore, we evaluate the Alignment of different α and β values in Table 4.

Tables 1 and Table 2 evaluate the performance across 9 frame and 4 frame domains respectively. In Stage 1, introducing LoRA Masks improves both Alignment and Coherence compared to the Base Model, while adding Representation Alignment (RA) further can further improve performance, especially in complex fields such as “ink painting”

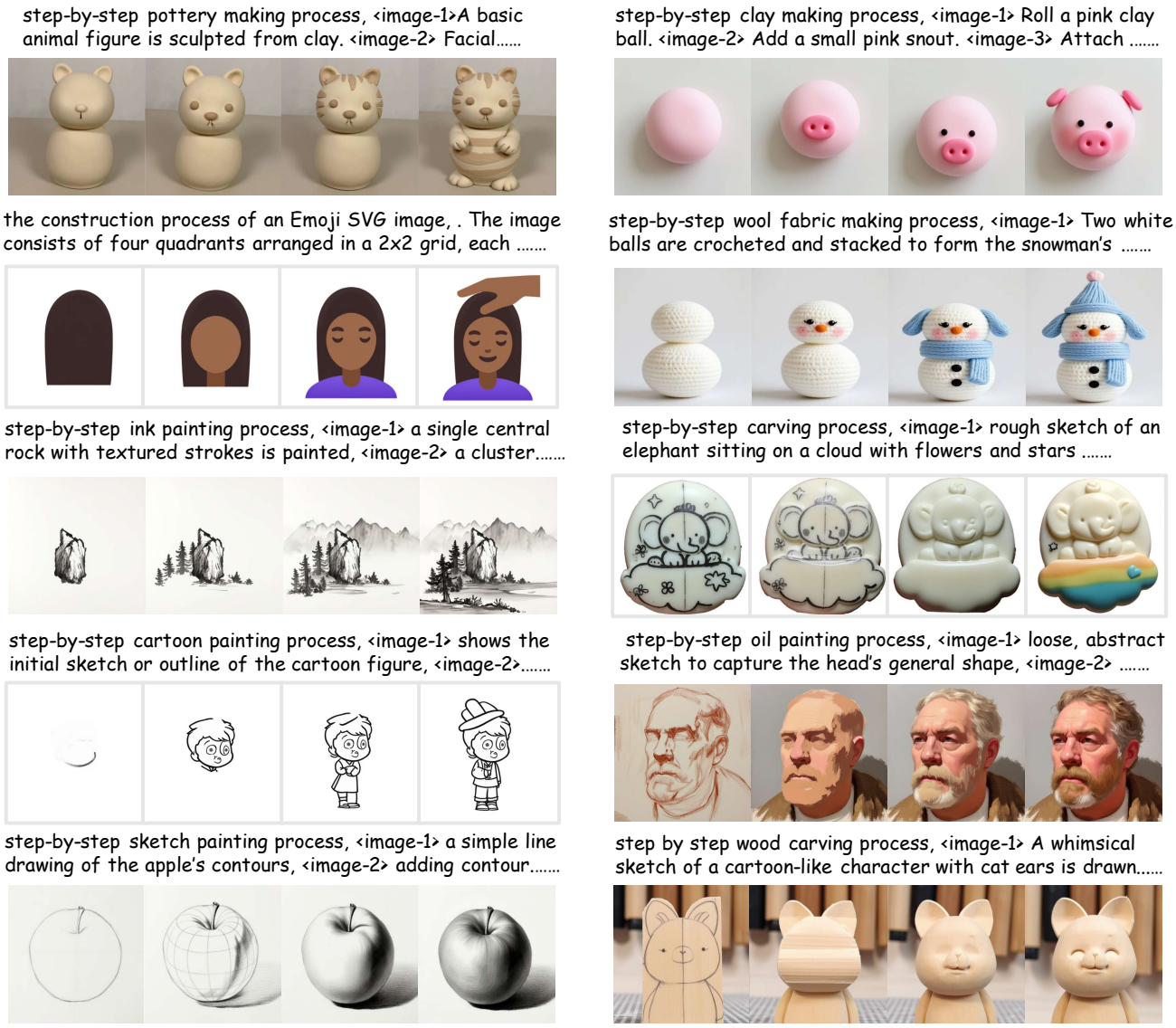


Figure 4. Generation results of our proposed ProcessMaker on the 4 frame dataset.

and "jade carving", where cross domain consistency is crucial. For simpler fields such as "Emoji" and "LEGO", the benefits of Representation Alignment are not so obvious. In Stage 2, the ProcessMaker achieves the best results across all domains, demonstrating its effectiveness in refining feature representations. Ablating global alignment (Δ_{glob}), local alignment (Δ_{loc}), or semantic alignment (Δ_{sem}) reveals their importance for specific domains. For example, removing Δ_{loc} significantly impacts Coherence in texture-heavy domains like "Sketch", while removing Δ_{sem} affects Alignment in semantically complex domains like "Cook".

The Flux.1 [4] consists of MM-DiT Blocks (layers 0–18) for multimodal fusion and Single-DiT Blocks (layers 19–37) for unimodal image refinement. In Table 3, we can find that extracting representations from both MM-DiT and Single-DiT Blocks achieves the best overall balance. This indicates that aligning features across both the multimodal interaction and visual refinement stages helps maintain semantic consistency and stable representation learning. Furthermore, Table 4 demonstrates that when $\alpha = 0.6, \beta = 0.2$, the generated sequences have better alignment.

2.4. More Qualitative Results

In this section, we provide more visualization results both on the domains within dataset and unseen domains. Figure 4 and Figure 6 present the generated 4 frame and 9frame results of the dataset domains. Figure 5 demonstrates the generated procedural sequences of unseen domains.

Folding origami jumping frog. Rectangular paper prepared and positioned. Diagonal folds create preliminary shape. Legs section formed with precise creases ……

Making a glass drinking goblet, Gather molten glass mass from crucible. Initial bubble blown into glowing glass. Stem pulled and shaped with tools. Bowl section …

Cross-stitch pattern creation. Aida fabric stretched in embroidery hoop. Pattern chart and floss colors prepared. First row of crosses stitched carefully. Color ……

Plant growth process from seed to bloom, A small seed is planted in dark soil. First green sprout emerges from the earth. Tiny leaves begin to unfold and expand ……

Creating sourdough bread, Active sourdough starter bubbly and ready. Starter mixed with flour and water. Dough undergoes stretch and fold technique. Long …

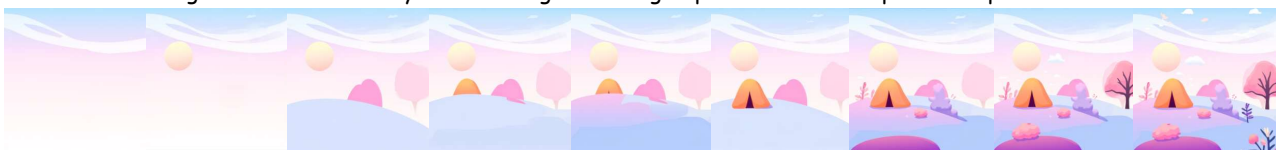
Leather bag, Pattern pieces cut from quality leather. Edges prepared and holes punched. Bottom and sides stitched together. Gussets added for bag structure…

Figure 5. The generalized procedural sequences of unseen domains without the guidance from additional LoRAs, including Origami, Glass firing, Embroidery, Plant growth, Bread making, and Leather crafting.

step-by-step cooking process of certain meal, <image-1> Adding garlic, ginger, and scallions into a heated wok. <image-2> Adding a cube of chili paste or fermented sauce. <image-3> Stirring the sauce into boiling water. <image-4> Adding ……

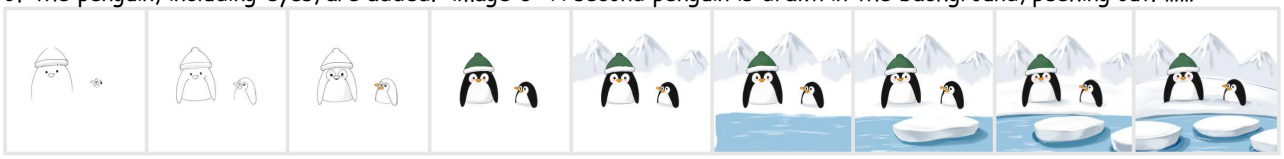
step by step SVG image construction process of daily life, <image-1> A rectangular block appears in the center, resembling a fireplace outline. <image-2> The fireplace is detailed with books and a fire at its base. <image-3> A rocking chair ……

Each panel depicts a different stage in the creation of a vibrant, whimsical digital illustration. The color palette primarily consists The background features a sky transitioning from a light pink hue at the top to a deeper blue at the bottom……



step-by-step construction process of LEGO model, <image-1> The base structure is created with two rods holding a small white brick assembly on a brown platform. <image-2> Additional white elements are added to the brick, forming small ……

step-by-step painting process, <image-1> A penguin's hat and head are drawn as the first step. <image-2> The face details of the penguin, including eyes, are added. <image-3> A second penguin is drawn in the background, peeking out. ……



Illustrate an older man with a bald head, a thick gray mustache, and a firm, serious expression. Start with bold shapes on a dark blue background. Gradually refine his sharp features, deep crow's feet, and upright posture.

step-by-step description of sand art creation, <image-1>: A simple curved line is drawn, forming the beginning of character. <image-2>: A rounded head and body take shape, resembling a cartoon-like figure. <image-3>: The body is shaded to ……

Figure 6. Generation results of our proposed ProcessMaker on the 9 frame dataset.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, pages 7514–7528, 2021. 1

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, pages 12513–12525, 2022. 1

[4] Black Forest Labs. Flux: Official inference repository for flux.1 models, 2024. 3

[5] Yiren Song, Cheng Liu, and Mike Zheng Shou. Makeanything: Harnessing diffusion transformers for multi-domain procedural sequence generation. *arXiv preprint arXiv:2502.01572*, 2025. 1