Towards Fine-Grained Attribution: Instance-Aware Preference Optimization for Aligning Diffusion Models

Supplementary Material

A. Additional Implementation details

To ensure the fairness and reproducibility of all experiments, this section provides additional key implementation details. Regarding model initialization, for Stable Diffusion 1.5 (SD1.5) [7], both IAPO and InPO [3] utilize the same high-quality training starting point: the officially provided supervised fine-tuned (SFT) model weights available in their Hugging Face repository. All other baseline models (DPO [6], KTO [2]) directly employ the official pre-trained checkpoints released by their respective authors on Hugging Face, ensuring all comparisons are conducted on an authoritative and consistent baseline. For Stable Diffusion XL (SDXL) [4], to enable the most rigorous comparison and highlight the advantage of our proposed instance-level dataset, we fully reproduce the official InPO codebase and train both IAPO and InPO using identical hyperparameters. During evaluation, the inference configuration remains unified across all text-to-image generation: the classifier-free guidance scale is set to 7.5 for SD1.5 and 5.0 for SDXL, with sampling steps fixed at 50. The training random seed is fixed at 42 for all training models, while the test-time random seed is fixed at 0. We provide the complete evaluation results for SDXL with seed 66 Tab. I to demonstrate the generalization capability of IAPO. Notably, the performance variance of IAPO remains minimal, demonstrating exceptional training stability. This consistent margin across multiple seeds confirms that our method's advantages are intrinsic to its design rather than artifacts of stochastic variation, thereby reinforcing the reliability of our main conclusions.

B. Training Efficiency Analysis on SDXL

In Fig. I, we present comparative results of IAPO and baseline methods in terms of image quality and training efficiency based on SDXL. As shown in Fig. I, IAPO reaches training speeds that are $6.25\times$ faster than Diffusion-DPO, and $1.31\times$ faster than InPO, while also generating higher-quality images.

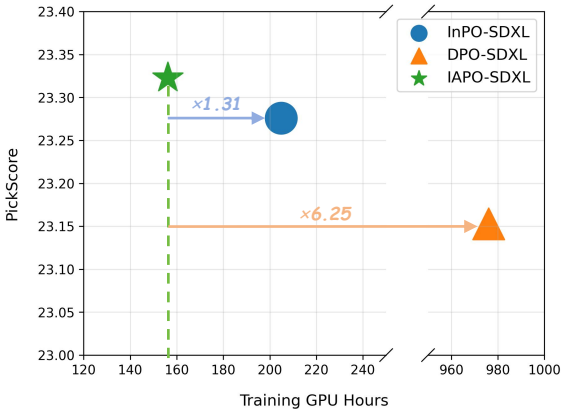

Figure I. A comparative evaluation of image quality and training efficiency between IAPO and baselines of SDXL on HPD v2.

C. Introduction to the Dataset

Pick-a-Pic v2 dataset is a large-scale collection of text-to-image pairs annotated with human preferences, sourced from the Pick-a-Pic web application. Each data point is structured as a tuple, `(caption, jpg_0, jpg_1, label)`, containing a text prompt, two corresponding generated images, and a binary label indicating the user's choice. The visual corpus was synthesized using a variety of text-to-image models—including Stable Diffusion 2.1, Dreamlike Photoreal, and several variants of Stable Diffusion XL (SDXL)—across a wide spectrum of Classifier-Free Guidance (CFG) scales.

HPD v2 dataset is curated by collecting human preferences on the "Dreambot" channel within the Stable Foundation Discord server. It contains 98,807 images generated from 25,205 prompts. The data structure links each prompt to a variable number of images, which are organized into pairs with labels denoting the user's preferred choice. For our analysis, we use the 3,200 prompts from the official test split.

Parti-Prompts dataset includes 1,632 text prompts created to evaluate text-to-image generation models. These prompts are divided into multiple categories to provide a wide range of challenges, which helps to assess model performance thoroughly across different areas.

DrawBench serves as a benchmark for assessing text-to-image generation models across 11 specific categories, such as object composition, spatial relationships, style consistency, and text accuracy. Through carefully crafted prompts, it evaluates a model's ability to translate semantic and stylistic cues into visuals, focusing on fidelity, creativity, and alignment with descriptions. By using side-by-side comparisons, DrawBench offers a qualitative approach to identifying strengths and weaknesses, fostering advancements in AI-driven generative models and computer vision.

Table I. We assess the stability of IAPO under multiple random seeds by fine-tuning SDXL and evaluating on Parti-Prompts, HPD v2 and Pick-a-Pic v2. Results are reported as mean and median scores; the best value for each metric is **bolded**.

Datasets	Models	Aesthetic [8]		PickScore [1]		HPS [9]		CLIP [5]	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
Parti-Prompts [10]	SDXL	5.7693	5.7571	22.6274	22.6496	28.4241	28.4723	35.5172	35.5675
	InPO (Seed=42)	5.8332	5.8150	22.9952	**22.9889**	28.8443	**28.8621**	**35.6453**	**35.6660**
	IAPO (Seed=42)	**5.9503**	**5.9439**	**23.0246**	22.9863	**28.8745**	28.8230	35.3439	35.3803
	InPO (Seed=66)	5.7592	5.7784	22.9152	22.9061	28.8096	28.8425	**35.8458**	**36.0351**
	IAPO (Seed=66)	**5.8767**	**5.8740**	**22.9360**	**22.9367**	**28.9871**	**28.9007**	35.4458	35.4845
HPD v2 [9]	SDXL	6.1338	6.1192	22.7835	22.7657	28.6278	28.6167	38.1623	38.3686
	InPO (Seed=42)	6.1797	6.1666	23.2761	23.2338	29.2413	29.2557	38.2722	38.5370
	IAPO (Seed=42)	**6.2580**	**6.2622**	**23.3230**	**23.2793**	**29.2848**	**29.2988**	**38.3220**	**38.5907**
	InPO (Seed=66)	6.1003	6.0803	23.1599	23.1429	29.1276	29.2130	**38.4731**	**38.6416**
	IAPO (Seed=66)	**6.1903**	**6.1789**	**23.2344**	**23.2543**	**29.2665**	**29.3127**	38.1847	38.3365
Pick-a-Pic v2 [1]	SDXL	6.0040	5.9788	22.1655	22.2426	27.9776	28.0080	36.1164	36.5889
	InPO (Seed=42)	6.0416	6.0167	22.5820	22.5573	28.5171	28.6445	36.5457	**36.7526**
	IAPO (Seed=42)	**6.1226**	**6.1282**	**22.6806**	**22.7006**	**28.6004**	**28.6707**	**36.7268**	36.7268
	InPO (Seed=66)	6.0551	6.0551	22.6064	22.6037	28.5605	28.2803	**37.9502**	39.0377
	IAPO (Seed=66)	**6.1918**	**6.2661**	**22.8349**	**23.0578**	**28.8182**	**28.7883**	37.3843	**37.3160**

D. Additional Quantitative Results

We further conduct extended quantitative studies to complement the results already reported. All baseline models are evaluated with the official released weights, without any re-training or fine-tuning. For reproducibility, the training random seed is fixed at 42 during generation; and the random seed for test is fixed at 0.

SDXL experiments are summarized in Table II. We compare our approach with two recent preference-alignment methods, Diffusion-DPO and MAPO. Evaluations are performed on Pick-a-Pic v2, HPD v2 and Parti-Prompts. Our model obtains the highest average score, outperforming all the baselines.

Table III reports DrawBench results for both SD 1.5 and SDXL, using the same baselines and metrics as above. Our method achieves the best overall accuracy on both model scales. Results substantiate that IAPO's improvements extend to scenarios demanding rigorous text-image correspondence, validating its effectiveness for applications where semantic accuracy is critical.

E. Additional Ablation Studies

We present additional ablation studies to dissect the contributions of key components in our Instance-Aware Preference Optimization (IAPO) framework. All experiments are conducted on SD1.5 using the Pick-a-Pic v2 test set unless otherwise specified.

E.1. Ablation on w_{pos} and w_{neg}

The weight assignment for w_{pos} follows the same establishment method as w_{neg}. When the preference for the n-th instance aligns with the global label—meaning we prefer the instance in $\boldsymbol{x}_0^w$ (denoted by $\rho_n = 0$)—we assign a weight of w_{pos} to the corresponding bounding boxes to reinforce the learning of these consistent instances:

$$M_n^*(i,j) = \begin{cases} w_{\text{pos}} & \text{if } (i,j) \in b_n^* \text{ and } \rho_n = 0 \\ w_{\text{neg}} & \text{if } (i,j) \in b_n^* \text{ and } \rho_n = 1 \\ 1 & \text{otherwise} \end{cases}. \quad (1)$$

As shown in Table IV, we vary w_{neg} while fixing $w_{pos} = 1$. Performance consistently improves as w_{neg} decreases, with the best results achieved at w_{neg}=0, i.e., completely suppressing gradients from conflicting instances in the losing sample. This confirms that ignoring these distractor regions is more effective than down-weighting them, as it prevents the model from reinforcing undesired artifacts. Notably, adjusting w_{pos} yields diminishing returns, suggesting that penalizing bad regions in positive samples is more critical than amplifying good regions in negative ones. We therefore adopt $w_{neg} = 0$ and $w_{pos} = 1$ as our default configuration.

We further set w_{neg} to -1 to investigate the model's behavior. Experimental results show that the training process becomes highly unstable under this setting. We argue that this instability stems from an inherent conflict between pixel-level weight assignment and the optimization objective: when the loss weights for certain pixels are nega-

Table II. We compare different preference optimization methods by fine-tuning SDXL and evaluating on Parti-Prompts, HPD v2 and Pick-a-Pic v2. Results are reported as mean and median scores; the best value for each metric is **bolded** and the second-best is underlined.

Datasets	Models	Aesthetic [8]		PickScore [1]		HPS [9]		CLIP [5]	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
Parti-Prompts	SDXL	5.7693	5.7571	22.6274	22.6496	28.4241	28.4723	35.5172	35.5675
	DPO	5.7946	5.8097	22.9307	22.9264	**28.9085**	**28.8873**	**36.4665**	**36.4017**
	MaPO	5.8979	5.9017	22.5946	22.5412	28.5551	28.4649	35.2734	35.3595
	IAPO	**5.9503**	**5.9439**	**23.0246**	**22.9863**	28.8745	28.8230	35.3439	35.3803
HPD v2	SDXL	6.1338	6.1192	22.7835	22.7657	28.6278	28.6167	38.1623	38.3686
	DPO	6.1124	6.1310	23.1330	23.1520	29.1650	29.1740	**38.8650**	**38.7110**
	MaPO	6.2416	6.2453	22.8488	22.8169	28.9939	29.0085	38.1587	38.5515
	IAPO	**6.2580**	**6.2622**	**23.3230**	**23.2793**	**29.2848**	**29.2988**	38.3220	38.5907
Pick-a-Pic v2	SDXL	6.0040	5.9788	22.1655	22.2426	27.9776	28.0080	36.1164	36.5889
	DPO	6.0127	6.0168	22.6397	22.6137	28.5933	28.5017	37.3825	**37.4003**
	MaPO	**6.2037**	**6.2294**	22.3179	22.3404	28.5381	28.6262	**37.4294**	37.3952
	IAPO	6.1226	6.1282	**22.6806**	**22.7006**	**28.6004**	**28.6707**	36.7268	36.7268

Table III. We compare different preference optimization methods by fine-tuning SD1.5 and SDXL and evaluating on DrawBench. Results are reported as mean and median scores; the best value for each metric is **bolded** and the second-best is underlined.

Base Model	Models	Aesthetic [8]		PickScore [1]		HPS [9]		CLIP [5]	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
SD1.5	SD1.5	5.2020	5.2677	21.3496	21.4011	27.4474	27.4640	32.5676	33.0921
	DPO	5.3099	5.3774	21.5932	21.7097	27.7069	27.8015	33.2266	33.7624
	KTO	5.3819	5.4458	21.6549	21.8707	28.2696	28.4872	33.4800	34.2055
	InPO	5.3927	5.4927	21.7661	21.8812	28.4733	28.6451	**34.2419**	34.8507
	IAPO	**5.5822**	**5.6711**	**21.8486**	**21.9922**	**28.5449**	**28.7431**	34.1385	**34.8584**
SDXL	SDXL	5.5860	5.6435	22.4924	22.6595	28.5996	28.8884	35.3466	36.5069
	DPO	5.6388	5.7231	22.8259	22.9755	29.0795	29.3985	**36.6062**	**37.8796**
	MaPO	5.7401	5.7255	22.5230	22.6339	28.8760	29.0263	34.6284	35.2852
	InPO	5.6779	5.6841	22.8186	23.0278	28.9748	29.2258	35.0993	36.0295
	IAPO	**5.7658**	**5.7566**	**22.9322**	**23.1392**	**29.2072**	**29.4069**	36.0807	36.5051

tive, the model receives opposing gradient signals within the same image. This signal conflict leads to inconsistent parameter update directions, disrupts the stability of the training process, and ultimately causes convergence failure.

E.2. Ablation on β Coefficient

The coefficient β balances preference alignment against deviation from the reference model. We test $\beta \in \{2000, 3000, 4000, 5000\}$ with all other settings fixed. Table V shows that while IAPO remains relatively stable across this range, $\beta = 2000$ achieves the highest scores on PickScore and HPS. A smaller β allows stronger optimization toward human preferences, which benefits fine-grained instance alignment; however, excessively low values risk overfitting to noisy annotations. The robustness across β values demonstrates that our instance-level credit assignment mechanism is the primary driver of performance gains, rather than precise regularization tuning.Inversion vs. Direct Sampling for SD1.5.

E.3. Ablation on Inversion Strategy for SD1.5

In our main experiments, we note that DDIM inversion produces semantically less faithful latents for SD1.5 compared to SDXL, leading us to use direct noise sampling for SD1.5 training. To systematically validate this design choice, we conduct an ablation where SD1.5 is trained with 10-step DDIM inversion. The results (Table VI) reveal a noticeable drop in generation quality and diversity compared to direct sampling, confirming that inversion-induced latent distortion hampers optimization for smaller models.

F. Additional Qualitative Results

Fig. III and Fig. IV displays additional outcomes for additional baselines and our IAPO on SD1.5 and SDXL. These

Table IV. Ablation study on weight assignment for positive and negative regions on Pick-a-Pic v2 dataset. Results are reported as mean and median scores; the best value for each metric is **bolded** and the second-best is underlined.

Dataset	w_{pos}	w_{neg}	Aesthetic [8]		PickScore [1]		HPS [9]		CLIP [5]	
			Mean	Median	Mean	Median	Mean	Median	Mean	Median
Pick-a-Pic v2	1	1	5.7352	5.7802	21.5235	21.6018	27.7578	27.9123	34.6614	35.0903
	2	1	5.7397	5.7660	21.5434	21.5860	27.7716	27.9111	34.6235	34.7461
	1	0.8	5.7384	5.7678	21.5264	21.5968	27.7672	27.9382	34.7159	34.9041
	1	0.5	5.7626	5.8009	21.5294	21.5695	27.7548	27.9478	34.7548	34.7695
	1	0.2	5.7731	5.8126	21.5502	21.5545	27.7894	27.9762	34.6510	34.7344
	2	0	5.7641	5.7865	21.5674	21.5735	27.8429	28.0401	**36.9766**	**36.4228**
	1	0	**5.7842**	**5.8168**	**21.5828**	**21.6021**	**27.8492**	**28.0742**	34.6979	35.0903

Table V. Ablation study on the β parameter on Pick-a-Pic v2 dataset. Results are reported as mean and median scores; the best value for each metric is **bolded** and the second-best is underlined.

Dataset	β	Aesthetic [8]		PickScore [1]		HPS [9]		CLIP [5]	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
Pick-a-Pic v2	5000	5.7025	5.7333	21.4195	21.4536	27.7656	27.8892	34.3700	34.8469
	4000	5.7199	5.7559	21.4356	21.4611	27.7901	27.9196	34.4997	34.8425
	3000	5.7497	5.7724	21.5032	21.5245	27.8175	28.0183	34.6176	34.9643
	2000	**5.7842**	**5.8168**	**21.5828**	**21.6021**	**27.8492**	**28.0742**	**34.6979**	**35.0903**

figures offer a comprehensive visual comparison that elucidates the performance disparities between state-of-the-art models and IAPO. Our method shows significant improvements over baseline models across multiple dimensions including aesthetic quality and text-image alignment, with particularly notable progress in instance-level generation performance.

G. Limitations

Our approach is constrained by the quality of the original dataset, which contains a non-negligible proportion of low-quality image-text pairs, and by the Vision-Language Model's tendency to hallucinate, producing occasional misalignments between visual content and textual descriptions; furthermore, our preference-learning framework is restricted to the instance level and may be too coarse for tasks that require fine-grained understanding such as distinguishing specific parts (e.g., an eagle's talons) or pixel-level attributes. The misuse of advanced image generation technologies presents ongoing threats to societal trust and information ecosystems. We consistently advocate for comprehensive ethical governance of IAPO throughout its lifecycle to ensure its development remains aligned with digital ethics standards.

H. Future work

We plan to construct a new high-quality dataset by leveraging state-of-the-art generative models to produce visually consistent and semantically accurate image-text pairs. To mitigate VLM hallucination, we will incorporate a more robust annotation pipeline that includes self-reflection mechanisms and cross-verification across multiple models. Furthermore, we aim to extend our preference modeling from the instance level to a more granular hierarchy, enabling part-level and even pixel-level preference learning. These improvements are expected to enhance both the reliability and expressiveness of the learned representations.

References

[1] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In *NeurIPS*, 2023. 2, 3, 4, 5

[2] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning Diffusion Models by Optimizing Human Utility. In *NeurIPS*, 2024. 1

[3] Yunhong Lu, Qichao Wang, Hengyuan Cao, Xierui Wang, Xiaoyin Xu, and Min Zhang. InPO: Inversion Preference Optimization with Reparametrized DDIM for Efficient Diffusion Model Alignment. In *CVPR*, 2025. 1

[4] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *ICLR*, 2023. 1

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual

Table VI. Ablation study on the effect of inversion in SD1.5 across different datasets. Results are reported as mean and median scores; the best value for each metric is **bolded** and the second-best is underlined.

Dataset	SD1.5	Aesthetic [8]		PickScore [1]		HPS [9]		CLIP [5]	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
DrawBench	w inversion	5.4230	5.4598	21.7540	21.9768	28.4336	28.6954	33.7922	34.5243
	w/o inversion	**5.5822**	**5.6711**	**21.8486**	**21.9922**	**28.5449**	**28.7431**	**34.1385**	**34.8584**
PartiPrompts	w inversion	5.5664	5.5970	21.8761	21.8947	28.2564	28.3230	34.5021	**34.5227**
	w/o inversion	**5.7270**	**5.7612**	**22.0082**	**21.9777**	**28.3587**	**28.3920**	**34.5909**	34.3746
HPD v2	w inversion	5.7916	5.7593	21.8716	21.8406	28.4611	28.4914	**36.6221**	**36.9317**
	w/o inversion	**5.9261**	**5.9227**	**22.0227**	**21.9931**	**28.6379**	**28.6878**	36.4029	36.7241
Pick-a-Pic v2	w inversion	5.6295	5.6464	21.4598	21.4744	27.6499	27.7286	**34.7737**	35.0402
	w/o inversion	**5.7842**	**5.8168**	**21.5828**	**21.6021**	**27.8492**	**28.0742**	34.6979	**35.0903**

Models From Natural Language Supervision. In *ICML*, 2021. 2, 3, 4, 5

[6] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *NeurIPS*, 2023. 1

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 1

[8] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 2, 3, 4, 5

[9] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv.2306.09341*, 2023. 2, 3, 4, 5

[10] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. In *TMLR*, 2022. 2

User

You are an object detection system. Your task is as follows:
You will be provided with two images simultaneously. Your goals are:
Detect significant, identical object instances that are present in both images. (e.g., man, woman, dog, moon, landmark building)
You must follow these instructions:
Instance Matching: Carefully analyze both images to identify prominent and recognizable similar object instances to the greatest extent possible. The matching should follow these principles:
Priority for Identical Instances: First, seek to identify instances confirmed to be the same entity (e.g., the same person, the same specific pet, the same landmark building).
Relaxed Similarity Standards: If truly identical instances cannot be found, identify highly similar object instances, allowing for:
1. Minor variations in viewpoint, pose, or size.
2. Similar but not perfectly identical object categories (e.g., generic categories like "person", "car", "building").

Your output must be a clear list and strictly adhere to the following format:
Matched Instance: [Instance Category 1]
Matched Instance: [Instance Category 2]
Matched Instance: [Instance Category 3]


Planner

Matched Instance: [Eagle]

User

You are an object comparison system. Your task is as follows:
You will be provided with two cropped images simultaneously. Your goals are:
For each instance, evaluate and select the better-performing instance from the two images based on three dimensions.

Quality Comparison: For each matched instance, evaluate the instance's performance in both images based on the following three dimensions and select the better one:
1. Anatomical/Structural Correctness: Are the object's proportions and structure correct and natural?
2. Naturalness & Artifacts: Does the object exhibit obvious AI-generated artifacts, blurriness, strange textures, or an unrealistic appearance?
3. Aesthetic Quality: Is the object visually appealing? Is the composition, color scheme, and overall concept harmonious and interesting?
Provide the final selection (Image A or Image B) without any detailed reason analysis.

Your output must be a clear list and strictly adhere to the following format:
Quality Comparison Result: {Image A / Image B} is better


Judge

Quality Comparison Result: Image B is better

Figure II. **Prompts** of Planner and Judge.

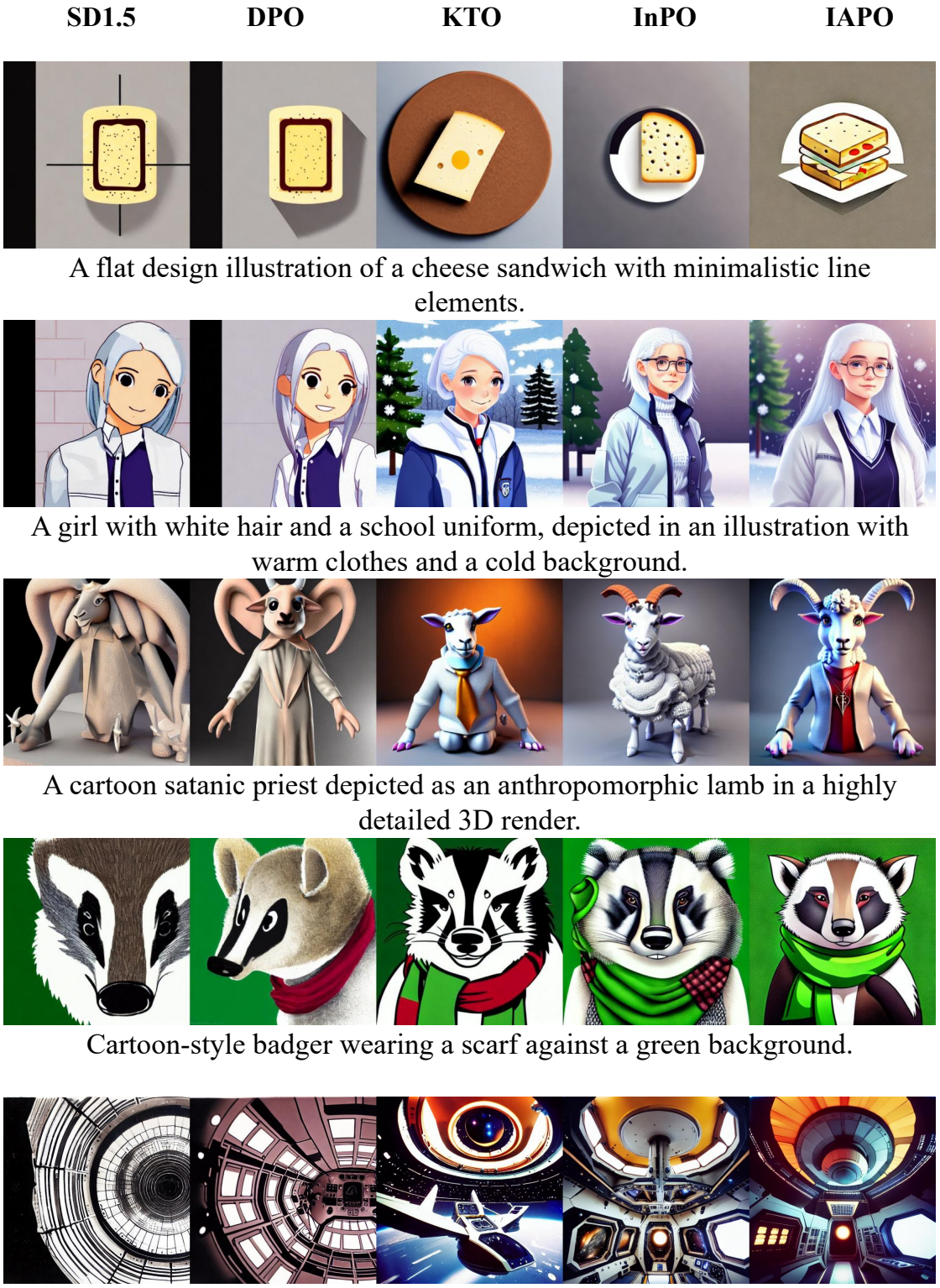

Figure III. Additional qualitative evaluation of IAPO-SD1.5 in comparison with Base-SD1.5, DPO-1.5, KTO-SD1.5 and InPO-SD1.5 on T2I generation tasks

Figure IV. Additional qualitative evaluation of IAPO-SDXL in comparison with Base-SDXL, DPO-SDXL, MaPO-SDXL and InPO-SDXL on T2I generation tasks