

1 Supplementary Material

2 The document provides supplementary information not elaborated on in our main paper due to space
3 constrains. Specifically, it includes details of Visual Relation Dataset (VisRel) (Section A), additional
4 study on the effect of visual prompts (Section B), additional comparisons on personalized test-time
5 tasks (Section C), and broader impact (Section D).

6 A Visual Relation Dataset (VisRel) Details

7 The Visual Relation Dataset (VisRel) is a diverse collection of 2D visual tasks reformulated as visual
8 transformations ($A \rightarrow A'$). It spans a wide range of task types and annotation formats, enabling the
9 modeling of a unified visual relation space. It aims to trigger cross-task generalization and test-time
10 adaptation via relation-space interpolation. VisRel integrates heterogeneous datasets, each of which
11 contributing different visual relations. For clarity, we categorize them into four groups based on their
12 underlying task semantics: (1) Image Restoration and Enhancement, (2) Physical and Geometric
13 Perception, (3) Semantic Perception, and (4) Generative Manipulation.

14 Table 1 provides a detailed overview of the datasets included in VisRel. For each dataset, we list
15 the task type, the visual transformation (input-output pair) that defines the task, and the annotation
16 source. This diverse and well-structured dataset provides the foundation for our visual in-context
17 learning framework, enabling PICO to generalize to novel user-personalized visual transformations
18 at test time.

Table 1: **Summary of datasets in VisRel.** Each dataset is represented by its task type, exemplar relation ($A \rightarrow A'$), and annotation source. **Ground Truth** denotes annotations provided by the original dataset, while **Human-labeled** indicates annotations created by us.

Dataset	Task Type	Visual Relation ($A \rightarrow A'$)	Annotation Source
Restoration / Enhancement			
DIV2K [1]	Deblurring	Blurry Image \rightarrow Clean Image	Ground Truth
DIV2K [1]	Denoising	Noisy Image \rightarrow Clean Image	Ground Truth
Synthetic Rain [2]	Deraining	Rainy Image \rightarrow Clean Image	Ground Truth
Dense-Haze [3]	Dehazing	Hazy Image \rightarrow Clean Image	Ground Truth
LOL [4]	Low-Light Enhancement	Low-Light Image \rightarrow Enhanced Image	Ground Truth
Physical / Geometric Perception			
Taskonomy [5]	Surface Normal Estimation	RGB Image \rightarrow Surface Normal Map	Ground Truth
Taskonomy [5]	Euclidean Distance Estimation	RGB Image \rightarrow Distance Map	Ground Truth
Taskonomy [5]	Z-buffer Depth Estimation	RGB Image \rightarrow Z-buffer Map	Ground Truth
Taskonomy [5]	Principal Curvature Estimation	RGB Image \rightarrow Curvature Map	Ground Truth
Taskonomy [5]	Reshading	RGB Image \rightarrow Re-rendered Image	Ground Truth
Taskonomy [5]	2D Keypoint Estimation	RGB Image \rightarrow 2D Keypoint Heatmap	Ground Truth
Taskonomy [5]	3D Keypoint Estimation	RGB Image \rightarrow 3D Keypoint Heatmap	Ground Truth
Taskonomy [5]	Occlusion Edge Detection	RGB Image \rightarrow Occlusion Edge Map	Ground Truth
Taskonomy [5]	Texture Edge Detection	RGB Image \rightarrow Texture Edge Map	Ground Truth
Semantic Perception			
MS-COCO [6]	Instance Segmentation	Image \rightarrow Instance Masks	Ground Truth
MS-COCO [6]	Panoptic Segmentation	Image \rightarrow Panoptic Masks	Ground Truth
MS-COCO [6]	Semantic Segmentation	Image \rightarrow Class Masks	Ground Truth
DIS5K [7]	Dichotomous Segmentation	Image \rightarrow Binary Mask	Ground Truth
CORE50 [8]	Object Detection	Image \rightarrow Bounding Boxes	Human-labeled
MS-COCO [6]	Human Pose Estimation	Image \rightarrow Keypoint Map	Ground Truth
OPRA [9]	Accordance Detection	Image \rightarrow Highlighted Accordance Part	Ground Truth
Generative Manipulation			
DIV2K [1]	Inpainting	Masked Image \rightarrow Completed Image	Ground Truth
MS-COCO [6]	Style Transfer	Image \rightarrow Stylized Image	StyleID [10]
PhotoDoodle [11]	Doodling	Image \rightarrow Image with Doodles	Ground Truth
Cat [12]	Sticker Addition	Image \rightarrow Image with Stickers	Human-labeled
PaintBucket [13]	Line Art Colorization	Line Art \rightarrow Colored Image	Ground Truth
ReNé [14]	Object Relighting	Image under Light A \rightarrow Light A'	Ground Truth

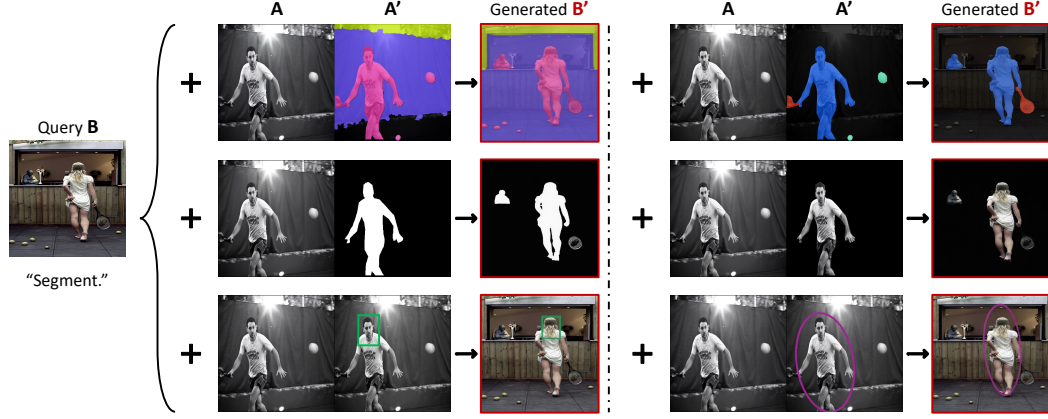


Figure 1: **Personalized segmentation with visual prompt control.** Given the same query image B and text prompt ("Segment"), PICO generates diverse outputs B' conditioned solely on the visual exemplar ($A \rightarrow A'$). The outputs vary in task type (e.g., stuff vs. semantic), style (e.g., binary vs. matting), granularity (e.g., boxes, circles), and spatial focus (e.g., head vs. full body).



Figure 2: **Context-aware sticker addition with visual prompt control.** Given the same query image B and text prompt ("Add the sticker"), PICO generates diverse outputs B' solely based on visual prompt ($A \rightarrow A'$). The model captures variation in object type, position, and scale, demonstrating precise spatial and semantic interpretation from visual prompts.

19 B Additional Study: Effect of Visual Prompts

20 In the main paper, we demonstrate how text prompts help resolve task ambiguity. Here, we focus
 21 on the role of visual prompts, *i.e.*, the in-context input-output exemplars ($A \rightarrow A'$), in achieving
 22 fine-grained control over output behavior. Given a fixed query image B and text prompt, PICO
 23 flexibly adapts to different task intents by interpreting the visual demonstration ($A \rightarrow A'$), producing
 24 diverse and context-appropriate outputs B' . We explore this behavior in three representative task
 25 categories:

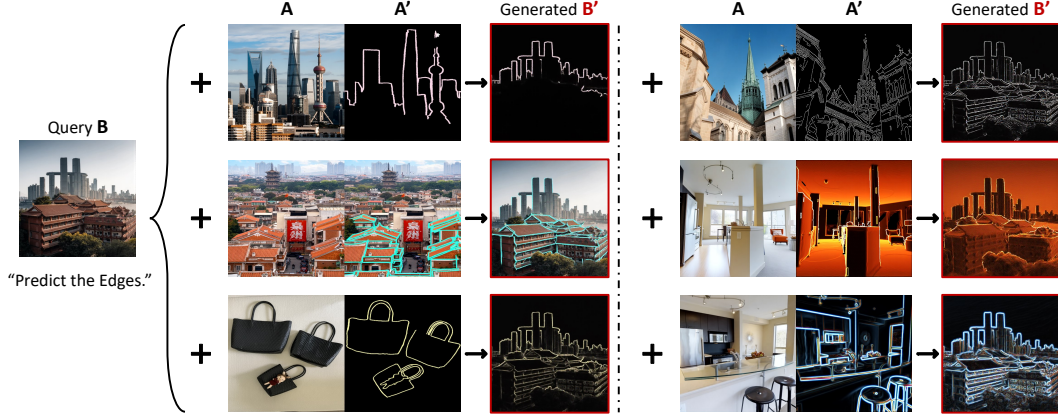


Figure 3: **Personalized edge detection with visual prompt control.** Given the same query image B and text prompt (“Predict the edges”), PICO generates diverse outputs B' solely based on visual exemplars ($A \rightarrow A'$). The model adapts spatial focus (e.g., top or bottom) and edge style (e.g., canny, euclidean, texture), guided entirely by visual cues.

(1) **Personalized Segmentation.** As shown in Figure 1, PICO supports diverse segmentation outputs defined by the contextual visual prompts, while the text prompts (e.g., “Segment”) remain unchanged. These outputs vary along several dimensions: (i) Task types: stuff segmentation vs. semantic segmentation (Row 1); (ii) Style: different mask representations, such as binary masks with white silhouette and matting-like masks that preserved original region (Row 2); (iii) Granularity: sparse annotations like bounding boxes and circles (Row 3); (iv) Spatial focus: segmenting the entire body vs. specific parts like the head (Row 3). The model consistently aligns its predictions with the intent, style, structure, and semantics expressed in the in-context visual prompts.

(2) **Context-Aware Sticker Addition.** Figure 2 shows how the visual prompt controls what customized object or doodling is added, where it is placed, and how it is scaled. For example, the size of a Christmas hat (Row 2) changes based on the visual exemplar, despite the same text prompt (“Add the sticker”). This task highlights the limitations of text-only instructions and the strength of visual exemplars for conveying spatial and compositional intent.

(3) **Personalized Edge Detection.** As shown in Figure 3, PICO handles edge detection tasks defined by spatial constraints and style cues in the visual prompts. The model is able to adaptively predict edges of specific regions (e.g., top vs. bottom) or emulate particular edge styles (e.g., Canny vs. Euclidean vs. texture-based) without making any changes to the text prompt (“Predict the edges”).

These results confirm that PICO effectively comprehends and infers the visual relation conveyed by the in-context input-output pairs, and successfully applies the visual logic to query images. Our quad-grid in-context format provides a strong structural prior for visual reasoning, enabling flexible, controllable, and freeform test-time personalization that extends beyond the expressiveness of text prompts alone.

C Additional Comparison: Personalized Test-time Tasks Generalization

C.1 Quantitative Evaluation

We conduct quantitative experiments to evaluate PICO’s ability to generalize to novel, compositional tasks. We focus on two representative task combinations that offer clear evaluation protocols:

(1) **Deraining with inpainting.** We evaluate 200 images corrupted by rain and occlusions. We use: (i) PSNR to assess pixel-level reconstruction fidelity, and (ii) SSIM to measure structural similarity between the predicted output B' and the clean reference image $\text{Cleaned}(B)$.

(2) **Inpainting with Stylization.** We evaluate 265 stylized images across 40 style different styles, each stylized using StyleID [10] and then corrupted by watermarks or inpainting masks. Evaluation metrics include: (i) Gram Matrix Distance between B' and the reference style image A' to measure

style fidelity, (ii) LPIPS between B' and the original $\text{Cleaned}(B)$, to evaluate content preservation and occlusion removal, and (iii) ArtFID [10], defined as $(\text{LPIPS} + 1) \cdot (\text{FID} + 1)$, which captures the overall trade-off between perceptual faithfulness and style fidelity. As a reference upper bound, we include the “ground truth” result: applying StyleID [10] directly to the clean image $\text{Cleaned}(B)$ using the same target style as A' .

Table 2 reports the quantitative results. PICO consistently achieves the best performance across both tasks and all evaluation metrics. In the deraining with inpainting task, PICO significantly outperforms all baselines in PSNR and SSIM, demonstrating superior restoration quality. In the inpainting with stylization task, PICO obtains the lowest Gram distance, LPIPS, and ArtFID, indicating strong style alignment with the in-context reference style while effectively preserving content and removing occlusions. Although GPT-4o produces visually appealing outputs, we observe noticeable spatial inconsistencies and content drift.

C.2 Additional Quantitative comparisons

Figures 4, 5, 6, 7, 8 present additional qualitative comparisons across diverse test-time personalized tasks: background-only stylization, edge detection with spatial constraints, joint deraining with inpainting, watermark removal with stylization, and context-aware sticker addition. PICO demonstrates consistent superiority in aligning with the task intent, as defined by in-context visual exemplar pair $(A \rightarrow A')$. GPT-4o shows strong semantic-level understanding but lacks precision in content fidelity and spatial alignment, especially in tasks that require geometric fidelity or pixel-aligned outputs.

Table 2: **Quantitative comparison on two test-time personalized tasks.** The best results are highlighted in **bold**, and the second-best are underlined. GPT-4o* results are based on 10 random samples due to API constraints.

	Ref	VP [15]	Analogist [16]	PromptDiff [17]	InstaManip [18]	GPT-4o* [19]	PICO (Ours)
<i>deraining with inpainting</i>							
PSNR (dB)↑	∞	<u>14.62</u>	12.35	9.64	10.94	12.29	22.24
SSIM ↑	1.0	<u>0.36</u>	0.35	0.10	0.33	0.26	0.67
<i>inpainting with stylization</i>							
Gram↓	17.29	28.96	26.53	61.61	44.39	<u>22.04</u>	21.27
FID↓	1.71	<u>1.86</u>	1.82	<u>1.86</u>	1.88	1.87	1.87
LPIPS↓	0.62	0.82	0.70	0.77	<u>0.60</u>	0.68	0.52
ArtFID↓	4.38	5.19	<u>4.79</u>	5.06	4.59	4.81	4.38

D Broader Impact

Our method, PICO, enables users to perform personalized vision tasks at test time from a single in-context input-output exemplar. Positively, PICO democratizes access to vision systems by allowing non-experts to personalize models for their own objects and tasks. It is highly data-efficient, reducing the need for large-scale annotations or fine-tuning. Methodologically, it sheds new light on the use of generative priors for in-context learning with diffusion transformer, raising important questions about how generative features can support deterministic perception tasks that require stable, task-specific representations. This may help bridge the gap between generation and understanding in vision. As with any personalization technology, misuse is possible. One-shot segmentation could potentially be exploited for surveillance or military targeting of individuals or assets. Responsible deployment in sensitive domains requires careful ethical consideration and appropriate safeguards.

References

- [1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017.
- [2] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *CVPR*, 2020.
- [3] Codruta O. Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense haze: A benchmark for image dehazing with dense-haze and haze-free images. In *ICIP*, 2019.

- 95 [4] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light
96 enhancement. In *BMVC*, 2018.
- 97 [5] Amir R Zamir, Alexander Sax, , William B Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese.
98 Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018.
- 99 [6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár,
100 and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.
- 101 [7] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate
102 dichotomous image segmentation. In *ECCV*, 2022.
- 103 [8] Vincenzo Lomonaco and Davide Maltoni. CORE50: a new dataset and benchmark for continuous object
104 recognition. In *CoRL*, 2017.
- 105 [9] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object
106 affordances from online videos. In *CVPR*, 2018.
- 107 [10] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for
108 adapting large-scale diffusion models for style transfer. In *CVPR*, 2024.
- 109 [11] Shijie Huang, Yiren Song, Yuxuan Zhang, Hailong Guo, Xueyin Wang, Mike Zheng Shou, and Ji-
110 aming Liu. PhotoDoodle: Learning artistic image editing from few-shot pairwise data. *arXiv preprint*
111 *arXiv:2502.14397*, 2025.
- 112 [12] Nick Crawford. Cat dataset. <https://www.kaggle.com/datasets/crawford/cat-dataset>, 2019.
- 113 [13] Yuekun Dai, Shangchen Zhou, Qinyue Li, Chongyi Li, and Chen Change Loy. Learning inclusion matching
114 for animation paint bucket colorization. *CVPR*, 2024.
- 115 [14] Marco Toschi, Riccardo De Matteo, Riccardo Spezialetti, Daniele De Gregorio, Luigi Di Stefano, and
116 Samuele Salti. ReLight My NeRF: A dataset for novel view synthesis and relighting of real world objects.
117 In *CVPR*, 2023.
- 118 [15] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via
119 image inpainting. *NeurIPS*, 2022.
- 120 [16] Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. Analogist: Out-of-the-box visual in-context
121 learning with image diffusion model. *TOG*, 2024.
- 122 [17] Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan
123 Zhou, et al. In-context learning unlocked for diffusion models. *NeurIPS*, 2023.
- 124 [18] Bolin Lai, Felix Juefei-Xu, Miao Liu, Xiaoliang Dai, Nikhil Mehta, Chenguang Zhu, Zeyi Huang, James M
125 Rehg, Sangmin Lee, Ning Zhang, et al. Unleashing in-context learning of autoregressive models for
126 few-shot image manipulation. In *CVPR*, 2025.
- 127 [19] OpenAI. Hello GPT-4o. <https://cdn.openai.com/gpt-4o-system-card.pdf>, 2024.

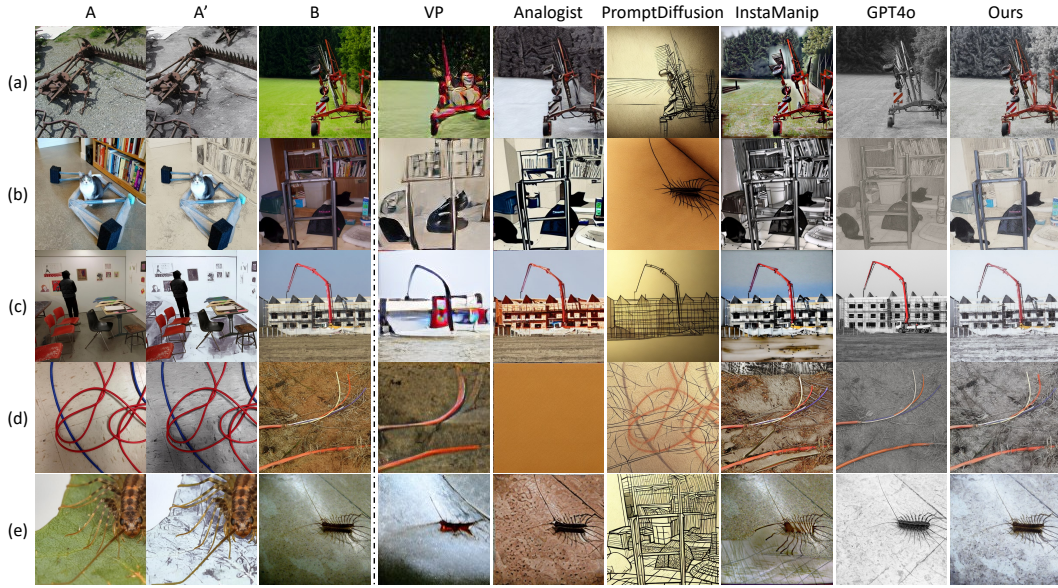


Figure 4: **Qualitative comparisons on background-only stylization.** PICO selectively stylizes the background while preserving the foreground.

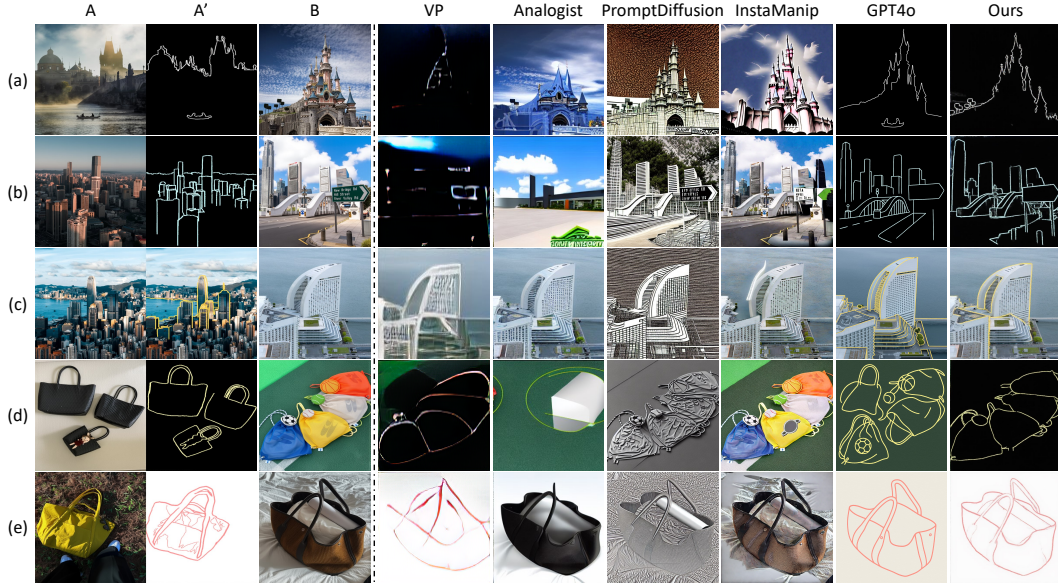


Figure 5: **Qualitative comparisons on edge detection with spatial constraints.** PICO accurately predicts personalized edge maps guided by the visual prompt.

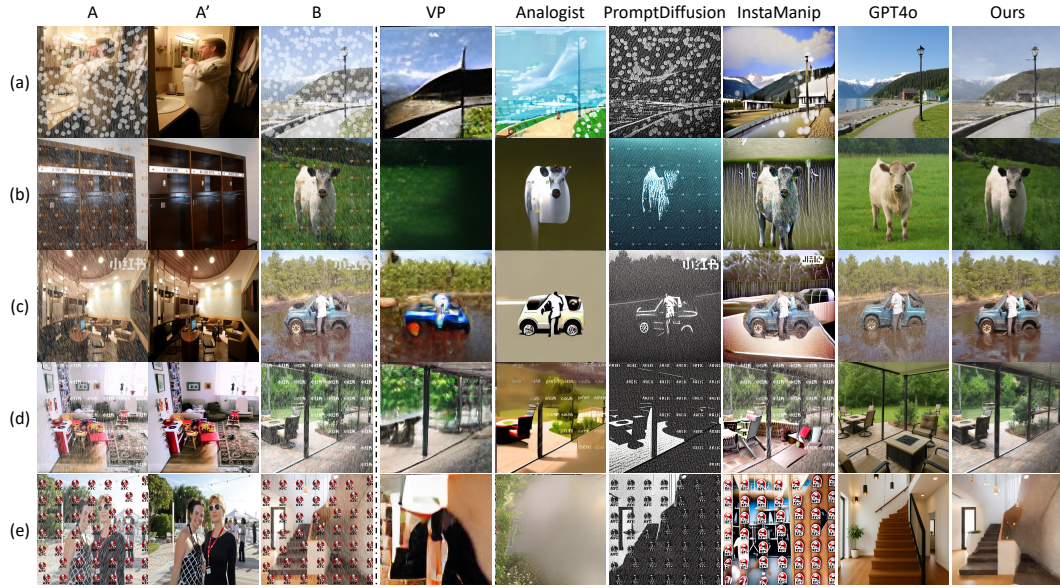


Figure 6: **Qualitative comparisons on joint deraining with inpainting.** PICO removes both rain and occlusions simultaneously.

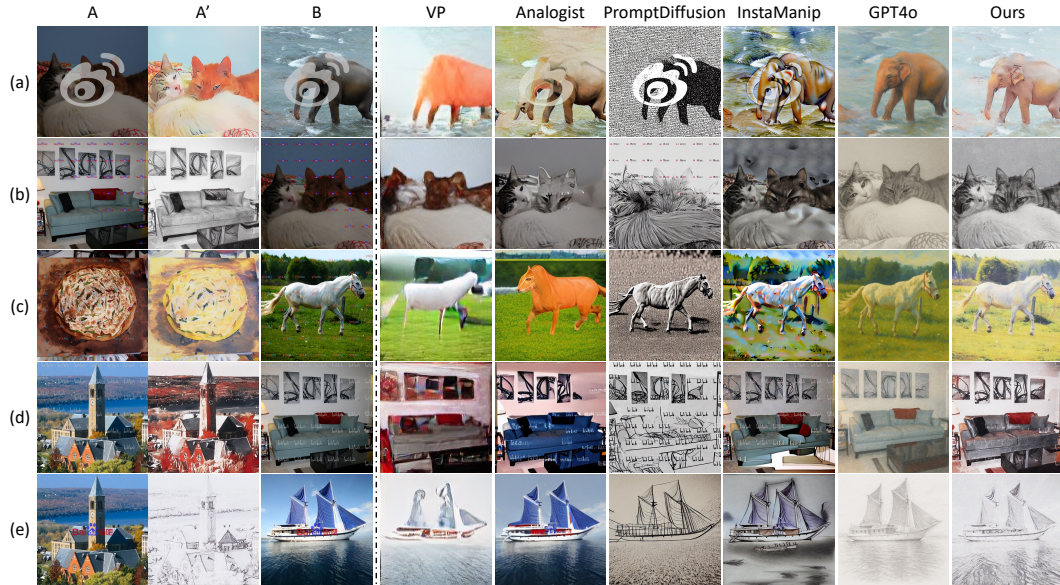


Figure 7: **Qualitative comparisons on watermark removal with stylization.** PICO removes occlusions while transferring target style.

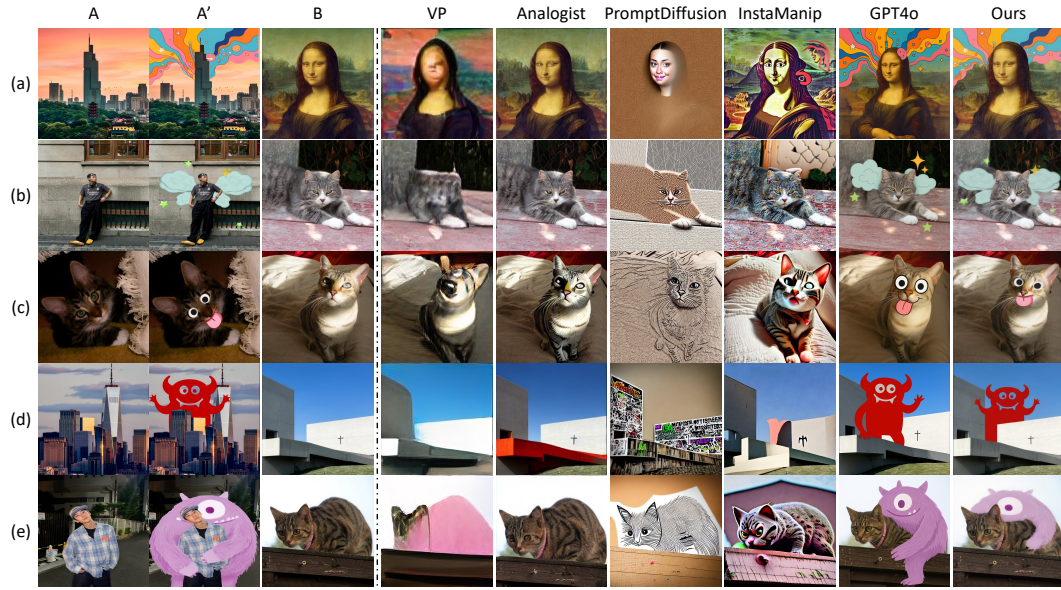


Figure 8: **Qualitative comparisons on context-aware sticker addition.** PICO learns from the visual exemplar where and how to place the sticker (*e.g.*, object type, size, position).