Stable Diffusion Models are Secretly Good at Visual In-Context Learning — Supplementary —

Trevine Oorloff[1,2*] Vishwanath Sindagi[1] Wele Gedara Chaminda Bandara[1]
Ali Shafahi[1] Amin Ghiasi[1] Charan Prakash[1] Reza Ardekani[1]

[1]Apple [2]University of Maryland - College Park

A. Overview

This document is structured as follows:

- Section B: Related work
- Section C: Implementation details
- Section D: Additional quantitative results
- Section E: Discussion on V-ICL models trained on task-related data
- Section F: Additional ablations
- Section G: Limitations and future work
- Section H: Additional qualitative results

B. Related Work

In-context learning (ICL) has garnered significant attention in the field of natural language processing (NLP) with the advent of large-scale language models like GPT-3 [5] and its successors [6, 19, 25, 26]. These models demonstrate the ability to perform tasks by conditioning on a small number of source-target examples, termed prompts, without any gradient updates or finetuning, effectively adapting to new tasks on-the-fly [12, 31]. The success of ICL in NLP has sparked interest in extending these capabilities to other domains, particularly in the realm of computer vision.

However, translating the concept of in-context learning from NLP to computer vision presents unique challenges due to the diversity in images and the inherent complexity of visual tasks. This has led to the emergence of two primary schools of thought in adapting ICL to computer vision, termed visual in-context learning (V-ICL).

The first approach adapts vision foundation models for in-context learning by training on uncurated datasets composed of random crops that potentially include examples of source images and corresponding targets (*e.g.* figures from computer vision papers). Research such as Visual Prompting [3] and IMProv [33] exemplify this approach, where they train a ViT-based MAE-VQGAN architecture [9, 13] on the task of masked inpainting. During inference, these methods involve creating composite images by stitching together a query image with prompt examples, forming a grid-like structure with a placeholder mask for the prediction, that the inpainting model can process. While these methods yield promising results, this approach often suffers from weaker inference of context between the query image and the prompt, lower resolution predictions, and overall weaker prediction quality.

The second school of thought aims to enhance prediction performance by training vision foundation models on curated/annotated task-related datasets. This method involves training/finetuning a model but uses paired source-target images of multiple tasks as training data. Notable examples of this method include Painter [27], Prompt Diffusion [30], SegGPT [28], Skeleton-In-Context [29], and Point-In-Context [10]. While models such as Painter and Prompt Diffusion target a relatively diverse set of tasks, the others focus on building generalist models to cater specific tasks such as segmentation, skeleton sequence modeling, or 3D point cloud estimation. Although these models achieve improved results and provide important insights for future research in visual in-context learning, they require updating model weights using datasets related to the out-of-domain tasks. This in turn implies the need for training data on related out-of-domain tasks that we are trying to adapt to. We believe that this ideology diverges from the core principles of ICL as they often fall short in generalizing to novel tasks that are unrelated to the training set and rely on large annotated datasets. This approach, therefore, somewhat undermines the fundamental idea of ICL, which emphasizes the ability to adapt to new tasks without retraining nor requiring a large annotated dataset.

C. Implementation Details

SD-VICL: We base our experiments on an off-the-shelf Stable Diffusion model [20], specifically the v1.5 checkpoint. Unless specified otherwise, we use the following hyperparameters for all our evaluations: denoising time steps (T) = 70, attention temperature (τ) = 0.4, contrast strength (β) = 1.67, and swap-guidance scale (γ) = 3.5. Further, we

*This work was completed during internship at Apple

Model	Foreground Segmentation (mIoU ↑)					Single Object Detection (mIoU ↑)				
	Split 0	Split 1	Split 2	Split 3	Avg.	Split 0	Split 1	Split 2	Split 3	Avg.
Number of Example Prompts: 1										
Visual Prompting [3]	34.85	38.55	34.51	32.24	35.04	48.82	48.52	45.11	42.72	46.29
IMProv (w/o text) [33]	41.46	43.60	39.70	33.22	39.50	46.10	47.26	41.97	39.96	43.82
IMProv (w/ text) [33]	41.31	44.64	40.86	35.93	40.69	44.69	48.10	44.53	40.34	44.42
SD-VICL (ours)	**44.05**	**45.17**	**44.36**	**42.11**	**43.92**	**54.45**	**52.92**	**51.56**	**47.27**	**51.55**
Number of Example Prompts: 5										
Visual Prompting [3]	36.70	40.02	36.18	32.56	36.37	51.59	49.30	46.80	44.66	48.09
SD-VICL (ours)	**55.55**	**56.08**	**55.84**	**54.49**	**55.49**	**58.99**	**56.31**	**57.09**	**56.01**	**57.10**

Table 1. Quantitative performance comparison of the proposed approach with recent approaches on foreground segmentation and single object detection for each split of the Pascal-5i dataset.

set the text condition of the Stable Diffusion pipeline to an empty string, and thus, no supplementary guidance is provided beyond the input prompts.

Comparison baselines: We use the publicly available repositories and checkpoints for Visual Prompting [3], IMProv [33], Painter [27], LVM [2], and Prompt Diffusion [30] to generate the results for all the experiments. For the text-guided variant of IMProv, as specified in their paper, we provide the model with a string comprising of the location and task information (*e.g.* "Left - input image, right - Black and white foreground/background segmentation"). To ensure a fair comparison, all methods, including ours, are evaluated using the same set of prompts, which we obtain using the unsupervised prompt retrieval method outlined by Zhang et al. [36].

Tasks and datasets: Below, we provide details on the tasks and datasets used for evaluations in our experiments:

- **Foreground segmentation:** This is a binary segmentation task, which predicts a binary mask of the object of interest (*i.e.* foreground) in an image. The prompt groundtruth is a black-and-white image with the foreground being white and the background being black. For evaluation, we use the Pascal-5i dataset [23], which comprises of 1864 images belonging to 20 object classes. The images are divided into four splits, where each split consists of five unique classes. We use the *mean intersection-over-union* (mIoU) as the evaluation metric.
- **Single object detection:** This task is similar to the foreground segmentation task, however, in this task, the bounding box of the object of interest is predicted instead of the mask with the exact boundary. For this task, the prompt groundtruth is a black-and-white image with the bounding box colored in white. We use the same dataset as foreground segmentation but include only images with single instances of objects following [3, 33]. The subset thus chosen consists of 1312 images and we report the mIoU scores.
- **Semantic segmentation:** This task predicts the per-pixel semantic label of a given image. We follow the method proposed by Wang et al. [27] to compose the prompt groundtruth, which assigns equally-spaced unique colors to each class. We use the Cityscapes dataset [7], which consists of 19 classes (excluding the void classes), and the COCOStuff dataset [16], which consists of 27 mid-level classes. We report the mIoU and pixel accuracy scores as evaluation metrics.
- **Keypoint detection:** The task of keypoint detection entails locating the critical points or landmarks of an object. In this study, we focus on human pose keypoint detection, which predicts the locations of the 17 keypoints defined in COCO [16]. Since the prompt groundtruth needs to be in the form of an image, we create an image that depicts the keypoints in the form of a heatmap as shown in Fig. 4. Each heatmap is created by superimposing Gaussian distributions centered on each keypoint. To accommodate the different spatial scales, we apply Gaussians with smaller variance for facial keypoints, which are relatively finer, and larger variance for body keypoints. These are visualized in two color channels: red for facial keypoints and green for body keypoints, facilitating easier decoding. For evaluation, following Hedlin et al. [14], we use the DeepFashion dataset [18] and report metrics: *mean sqaured error* (MSE) and the *percentage of correct keypoints* (PCK).
- **Edge detection:** The goal of this task is to predict the boundaries and edges within an image. For evaluation, we utilize the validation set of the NYUDv2 dataset [24] comprising 654 images. Since the validation set did not have the groundtruth, we used the soft edge maps generated using HED [32] as the pseudo-groundtruth. For evaluation we compute the MSE and the LPIPS loss [35] between the HED-predicted edge map and the V-ICL predictions.
- **Colorization:** In this task, the objective is to colorize a given grayscale image. Similar to [3, 33] we randomly sample 1000 images from the validation set of ImageNet

Model	Single Object Detection (mIoU ↑)			
	Split 0	Split 1	Split 2	Split 3
Number of Example Prompts: 1				
Visual Prompting	42.94	35.02	37.77	32.76
IMProv (w/o text)	42.32	36.52	36.32	31.83
IMProv (w/ text)	40.61	35.79	38.74	32.55
Ours	**47.74**	**39.86**	**44.93**	**37.92**
Number of Example Prompts: 5				
Visual Prompting	45.07	34.86	38.37	34.23
Ours	**51.74**	**43.15**	**50.23**	**47.20**

Table 2. Quantitative evaluation of single object detection on a subset of the Pascal-5i dataset, where larger objects with an area greater than 50% were excluded.

Model	Semantic Segmentation	
	mIoU ↑	Acc. ↑
Number of Example Prompts: 1		
Visual Prompting	15.31	39.07
IMProv (w/o text)	17.09	41.64
IMProv (w/ text)	17.19	42.35
Ours	**28.32**	**56.84**
Number of Example Prompts: 5		
Visual Prompting	13.01	36.12
Ours	**21.80**	**53.01**

Table 3. Quantitative evaluation of semantic segmentation on the COCOStuff dataset.

[21] for evaluation. We compute the LPIPS loss and the FID score [15] between the original colored image and the colorized prediction to evaluate the perceptual similarity.

D. Additional Quantitative Results

While in Tab. 1 we present the average performance for foreground segmentation and single object detection across all splits of the Pascal-5i dataset, in Tab. 1, we report the metrics for each split.

Additionally, in Tab. 1, we present the results of single object detection evaluated on the entire dataset for a more generalized assessment. However, in Tab. 2, we follow the approach of Bar et al. [3] and evaluate single object detection on a subset of the Pascal-5i dataset, where images with objects covering more than 50% of the image are excluded. While we observe an overall drop in absolute scores for all methods, the performance trends remain consistent with Tab. 1. This decline in performance can be attributed to the fact that larger objects are generally easier to detect than smaller ones, as noted by Bar et al. [3] as well.

Furthermore, we evaluated semantic segmentation on the COCOStuff dataset [16], where we report the results in Tab. 3. In contrast to the trend observed in Tab. 2, where performance improved with five example prompts compared to the single prompt, we could see a performance deterioration with five prompts in this case. Upon analysis, we identified that this performance decline was caused by the inconsistencies in labeling within the dataset, which creates confusion when inferring the context with multiple prompts, thereby negatively impacting the results.

E. Discussion on V-ICL Models Trained on Task-Related Data

As emphasized in the main paper, our work the first to propose a fully training-free paradigm that *uncovers* the V-ICL properties of a vision foundation model. For fairness, Sec. 3.1 of the main paper evaluates our approach against Visual Prompting [3] and IMProv [33], as they are the closest in methodology. While these models involve training, they do so on uncurated datasets, unlike models such as Painter [27], LVM [2], and Prompt Diffusion [30], which are trained on task-related annotated data.

To ensure completeness, we extend our evaluations to these V-ICL models trained on task-related data. Painter leverages a ViT-Large [8] backbone trained on multiple annotated datasets (*e.g.* COCO [16], ADE20K [37], and NYUv2 [24]). LVM is built on OpenLLaMA's 7B model [11] and trained on the UVD-V1 [2] dataset, a large-scale vision corpus comprising 50 datasets (*e.g.* LAION5B [22]) that span annotated, unannotated, and sequence images. Prompt Diffusion is a generative model based on Stable Diffusion, jointly finetuned on three forward tasks (*i.e.* image-to-depth, image-to-edge, image-to-segmentation) and their inverse variants, using vision-language prompts with paired images and text guidance. The training is conducted on a dataset adapted from Brooks et al. [4].

The quantitative and qualitative comparisons are presented in Tab. 2 and Fig. 1, respectively. Overall, we observe that our method outperforms all three baselines across multiple tasks.

We observe that all three modes often suffer from overfitting to training tasks, leading to poor generalization when exposed to novel tasks. Although visual in-context learning should ideally infer the task from the relationship between the prompt image and its groundtruth, these models demonstrate weakness in this regard.

Painter performs well on simple tasks like foreground segmentation and object detection when the query image contains a single foreground category (Fig. 1, row 1). However, in multi-class scenarios (Fig. 2a), Painter segments the entire foreground rather than focusing on the specific region of interest defined by the relationship between the prompt image and its groundtruth. Further, overfitting to

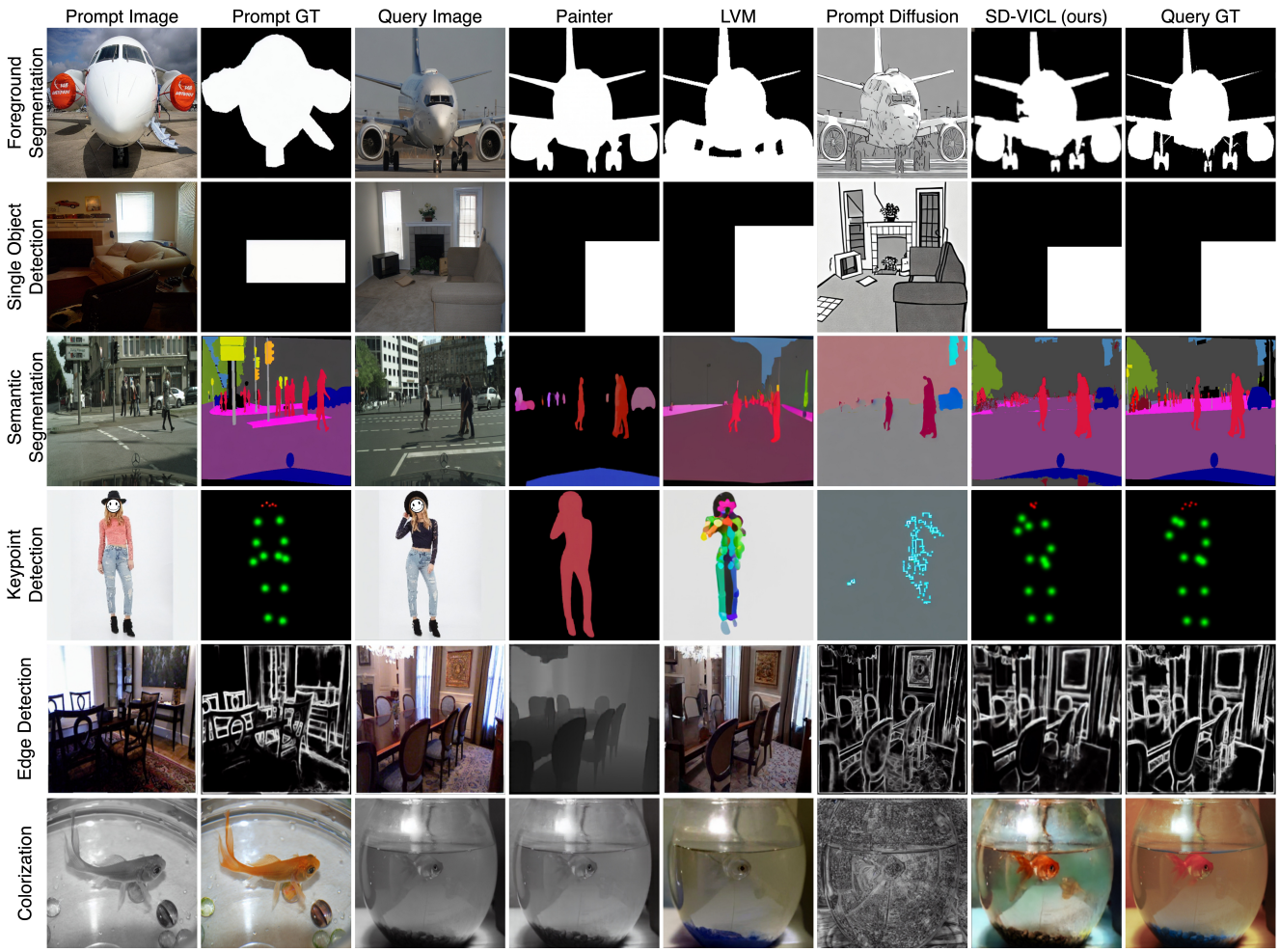


Figure 1. Additional qualitative comparisons illustrating the performance of training-based V-ICL models, Painter [27], LVM [2], and Prompt Diffusion [30] our proposed method on six different tasks. It can be seen that our method produces visually superior results as compared to the baselines.

training tasks is evident in rows 3 and 4 of Fig. 1, where Painter outputs a segmentation map in semantic segmentation with a different color scheme than defined in the prompt groundtruth. Similarly, for keypoint detection, Painter outputs a segmentation map instead of a heatmap for keypoints. Moreover, Painter struggles with colorization, often outputting the grayscale image itself. In edge detection, Painter outputs a depth map instead of the expected edge map (Fig. 1, row 2). This behavior suggests overfitting to the NYUv2 dataset, where the edge map query/prompt images overlap with those used for depth estimation during their training.

Similar limitations are observed for LVM, including poor performance on multi-class foreground segmentation (Fig. 2a), overfitting to training tasks (Fig. 1, row 4), and lack of generalization. Additionally, LVM exhibits inconsistencies in its outputs, as shown in Fig. 2b. Specifically, for a given task, despite the format/domain of the inputs remaining unchanged, we observe that the generated outputs belong to diverse domains. For example, in foreground segmentation, while some outputs align with foreground segmentation, others unexpectedly belong to unrelated domains such as keypoints, segmentation maps, or RGB images. This inconsistency highlights LVM's inability to produce coherent predictions despite the task and input format remaining unchanged.

Prompt Diffusion, while aiming to unlock in-context capabilities via vision-language prompts, remains constrained by the six tasks it is explicitly trained on. Although it exhibits relatively stronger performance in edge detection and segmentation — tasks included in its training, it struggles on tasks outside this scope. For instance, Prompt Diffusion produces structurally incoherent outputs for keypoint detection, object detection, and colorization, failing to align with the semantics illustrated by the prompt pair. It also occasionally hallucinates incorrect colors, textures, or layouts,

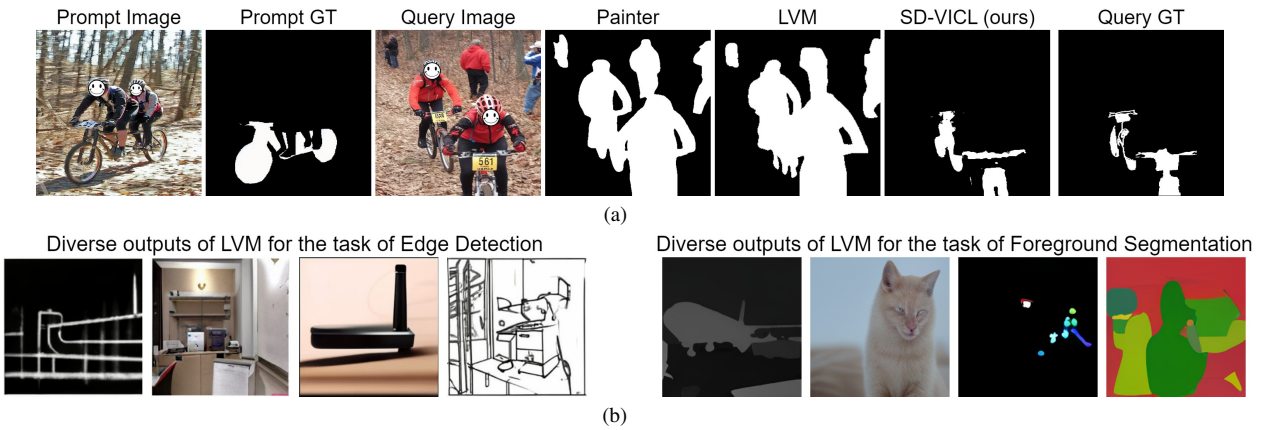


Figure 2. **Failure cases of V-ICL models trained on task-related data**, Painter [27] and LVM [2], implying poor task inference. In (a), both models fail in multi-class scenarios, segmenting the entire foreground instead of focusing on the region of interest defined by the prompt image and its corresponding ground truth. Examples in (b), depict inconsistent outputs generated by LVM for the same task (left: edge detection, right: foreground segmentation). The inputs for each of these outputs adhered to the same format as shown in Fig. 1, yet LVM produces outputs in diverse domains, deviating from the domain of the prompt groundtruth. These cases further emphasize the poor task inference capabilities of Painter and LVM.

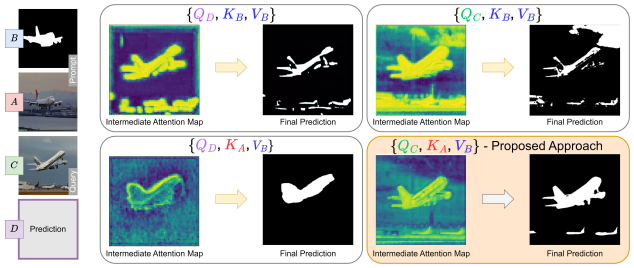


Figure 3. Qualitative examples of alternative attention formulations.

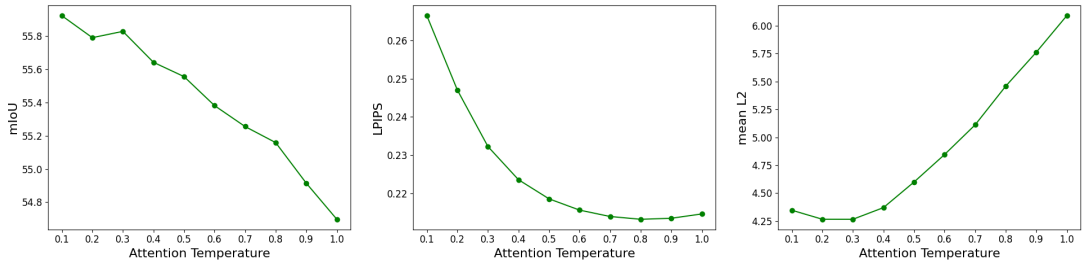
especially when confronted with novel task types or subtle prompt-query domain shifts.

These observations highlight shared limitations among Painter, LVM, and Prompt Diffusion in inferring tasks and context purely from input prompts. Their reliance on task-specific training data results in overfitting, leading to poor generalization on novel tasks. In contrast, our proposed training-free method demonstrates robust generalization and effective task inference, underscoring the benefits of uncovering V-ICL properties without additional training and the superiority of the proposed method to explicitly infer the context and task from the inputs, as intended by V-ICL.

F. Additional Ablations

In addition to the ablations discussed in the main paper, we also experimented with alternative attention formulations and the effects of several other factors such as temperature hyperparameter, resolution of the self-attention layers, contrastive strength parameter, swap-guidance scale, and AdaIN.

Alternative attention formulations: With regards to the attention formulation between query and the prompt, there are potentially multiple variants that could be used instead of the one described by Eq. (7). These candidate formulations can be derived by substituting the Q and K of Eq. (7) with the corresponding elements of each of these sets: $\{Q_D, K_B\}$, $\{Q_C, K_B\}$, and $\{Q_D, K_A\}$. Since the prediction needs to correspond to the features of the prompt groundtruth, the value vector, V, needs to come from B and cannot be substituted with other alternate options. Fig. 3 illustrates a subset of these variants along with the predictions obtained using these alternate formulations. The quantitative performance corresponding to these candidate formulations are presented in Tab. 4. However, since the prompt groundtruth lacks semantics, such formulations (*i.e.* $\{Q_D, K_B, V_B\}$, $\{Q_C, K_B, V_B\}$) tend to focus on color similarities rather than inferring the underlying semantic correlations. Alternatively, we can formulate the attention using the Query vector from the prediction (D) itself, similar to the approach followed by Alaluf et al. [1]. In this scenario, the intermediate predictions at early denoising stages closely resemble those produced by our formulation. However, in later denoising stages, the performance deteriorates as the prediction gradually shifts towards the prompt groundtruth, which lacks semantics, impairing the prediction performance. As seen in Fig. 3 and Tab. 4, the proposed formulation demonstrates superior performance which is achieved by ensuring that at each denoising step, the process is guided by the query and prompt latents at the corresponding denoising stages, thereby preserving the essential semantics needed for better context and task inference.

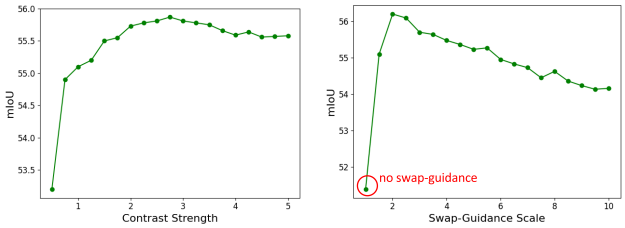


(a) (b) (c)

Figure 4. We illustrate the performance variation with respect to the attention temperature hyperparameter for the following tasks: (a) foreground segmentation, (b) colorization, and (c) keypoint detection.

Method	mIoU ↑
$\{Q_C, K_B, V_B\}$	12.54
$\{Q_D, K_B, V_B\}$	12.95
$\{Q_D, K_A, V_B\}$	23.68
Proposed: $\{Q_C, K_A, V_B\}$	**55.49**

Table 4. Ablation of attention formulations on foreground segmentation evaluated on Pascal-5i.



(a) (b)

Figure 5. Performance variation with respect to (a) contrast strength and (b) swap-guidance scale hyperparameters.

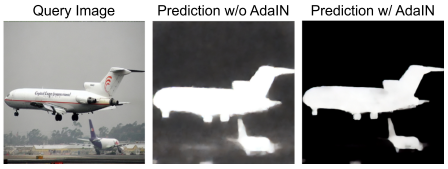


Figure 6. Example comparing the prediction with and without AdaIN.

Model	mIoU ↑
w/o AdaIN	51.55
w/ AdaIN	55.49

Table 5. Quantitative evaluation of with and without AdaIN evaluated using foreground segmentation.

Temperature hyperparameter, τ: As shown in Eq. (7), we introduce a temperature hyperparameter (τ) to the attention computation in order to control the sharpness of correspondence between the patches of the query image and the prompt image. While we use a constant temperature hyperparameter (*i.e.* $\tau = 0.4$) across all tasks to preserve generalization, we investigated the effect of τ on the performance of a few proxy tasks. We observe that the optimal temperature parameter varies notably with the task, which we depict in Fig. 4.

Contrast strength (β) and swap-guidance scale (γ) hyperparameters: We adapt the *attention map contrasting* (Eq. (8)) and *swap-guidance* (Eq. (9)) methods from Alaluf et al. [1] to address the domain gap introduced by using multiple images from different domains (*i.e.* source and target images belong to distinct domains). While we utilize the hyperparameter values proposed by [1] (*i.e.* $\beta = 1.67, \gamma = 3.5$), we investigate their impact on performance using foreground segmentation as a proxy task. We depict the variation of the performance with respect to the contrast strength and the swap-guidance scale in Fig. 5. A notable improvement in performance can be observed with a contrast strength greater than 1.0 and with swap-guidance enabled.

Adaptive instance normalization (AdaIN): As explained in Sec. 2.2, we utilize AdaIN to align the color distribution between the prediction (D), which is initialized using the noise space of the query image (C), and the expected groundtruth color space (*i.e.* color space of B). In Fig. 6 we present a comparison example with and without AdaIN, and in Tab. 5 we tabulate the overall performance on foreground segmentation. A clear performance improvement could be observed with the incorporation of AdaIN.

Resolution of attention layers: The denoising U-Net in the Stable Diffusion pipeline contains self-attention layers at multiple resolutions: 16×16, 32×32, and 64×64. Consequently, we can apply the proposed in-place attention reformulation to any combination of these layers. We evaluated different combinations of these resolutions, with the

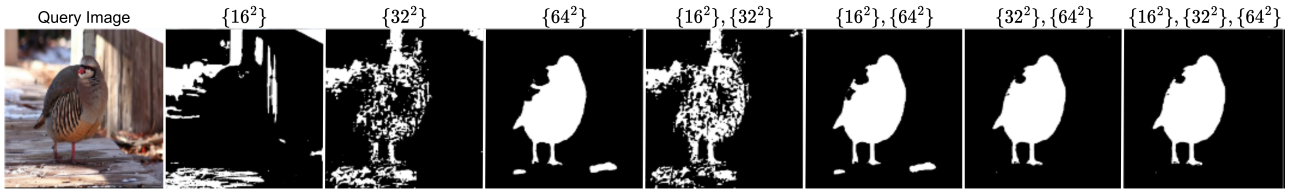


Figure 7. Qualitative examples of the output for each combination of self-attention layers modified using the proposed in-place attention reformulation.

Resolution			mIoU ↑
16×16	32×32	64×64	
✓	-	-	11.39
-	✓	-	32.50
-	-	✓	50.33
✓	✓	-	35.48
✓	-	✓	52.52
-	✓	✓	53.76
✓	✓	✓	**55.49**

Table 6. Quantitative evaluation on different combinations of resolutions which the self-attention layers could be modified using the proposed in-place attention reformulation.

results presented in Tab. 6. Additionally, we provide qualitative performance comparisons for each combination in Fig. 7. The best performance was achieved when modifying self-attention layers at all resolutions. This is intuitive, as it aggregates correspondences at multiple granularities, leading to a more comprehensive representation. In all our experiments, we use self-attention layers at all resolutions unless stated otherwise.

G. Limitations and Future Work

As with other diffusion-based methods, the primary limitation of our approach lies in its high inference time. In this work, our focus has been on exploring the V-ICL properties of Stable Diffusion, with less emphasis on computational efficiency. We believe it is crucial to first establish a robust and generalizable framework, with efficiency optimizations forming an important avenue for future work. In particular, integrating faster diffusion techniques [17, 34], which offer up to $100\times$ speedups without sacrificing output quality, could significantly reduce inference costs.

Another limitation, shared with other V-ICL methods, is sensitivity to noisy prompts. Since V-ICL methods rely on a small number of visual examples to infer both context and task, inaccuracies in prompt pairs can lead to degraded performance. While our implicitly-weighted prompt ensembling and attention temperature scaling partially mitigate this issue, further improvements in robustness to noisy or ambiguous prompts remain an open challenge.

Finally, extending our approach to the temporal domain, by adapting it to video generative models, presents a promising direction. Such an extension could enable training-free visual in-context learning for video-based tasks, further broadening the applicability of our framework.

Addressing these limitations could substantially enhance both the practicality and generality of visual in-context learning systems.

H. Additional Qualitative Results

We present additional qualitative examples for each task, foreground segmentation, single object detection, semantic segmentation, keypoint detection, edge detection, and colorization in Figs. 8 to 13 respectively.

Figure 8. Qualitative examples of foreground segmentation in comparison with Visual Prompting [3] and IMProv [33].

Figure 9. Qualitative examples of single object detection in comparison with Visual Prompting [3] and IMProv [33].

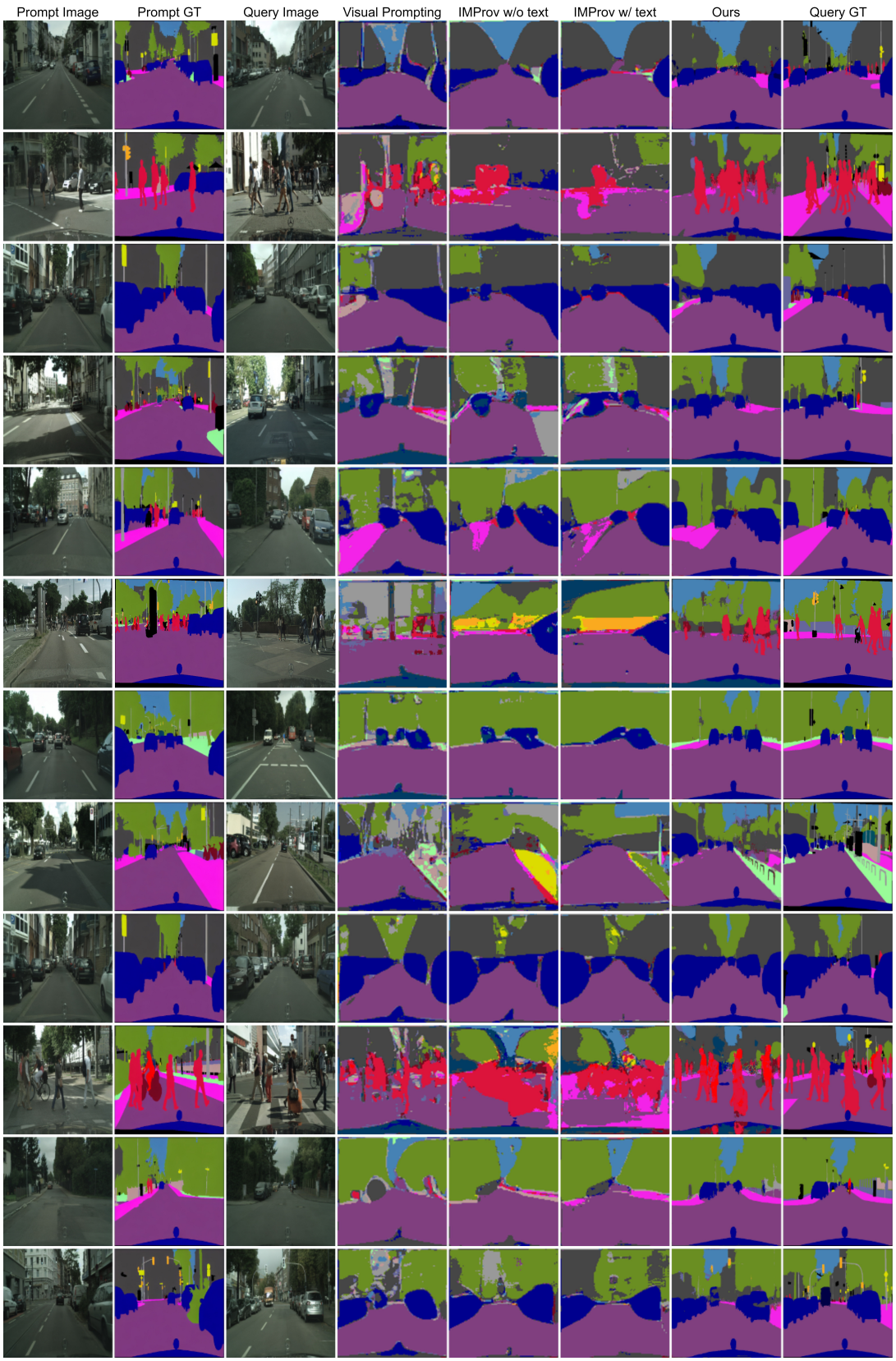


Figure 10. Qualitative examples of semantic segmentation in comparison with Visual Prompting [3] and IMProv [33].

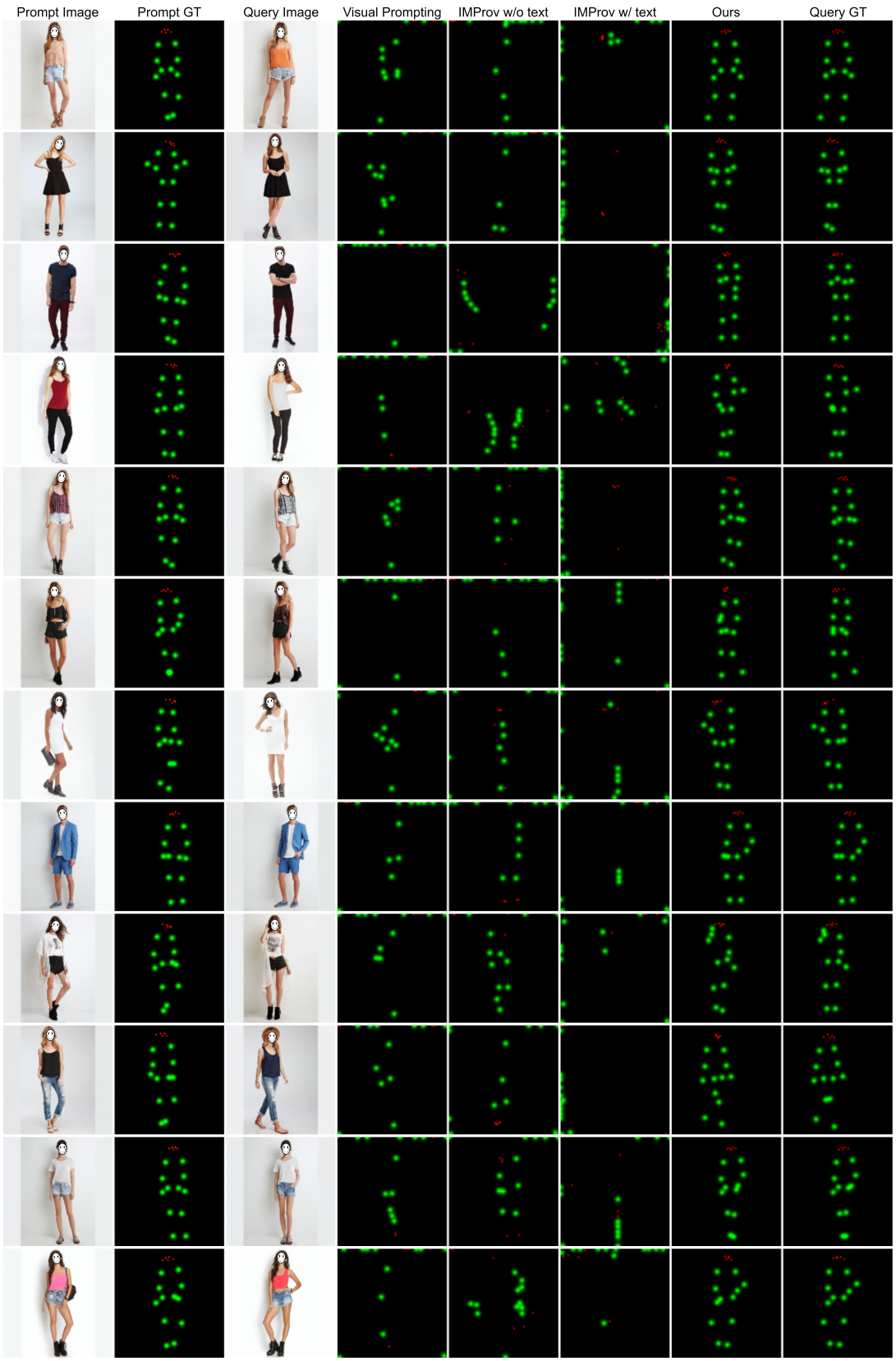


Figure 11. Qualitative examples of keypoint detection in comparison with Visual Prompting [3] and IMProv [33].

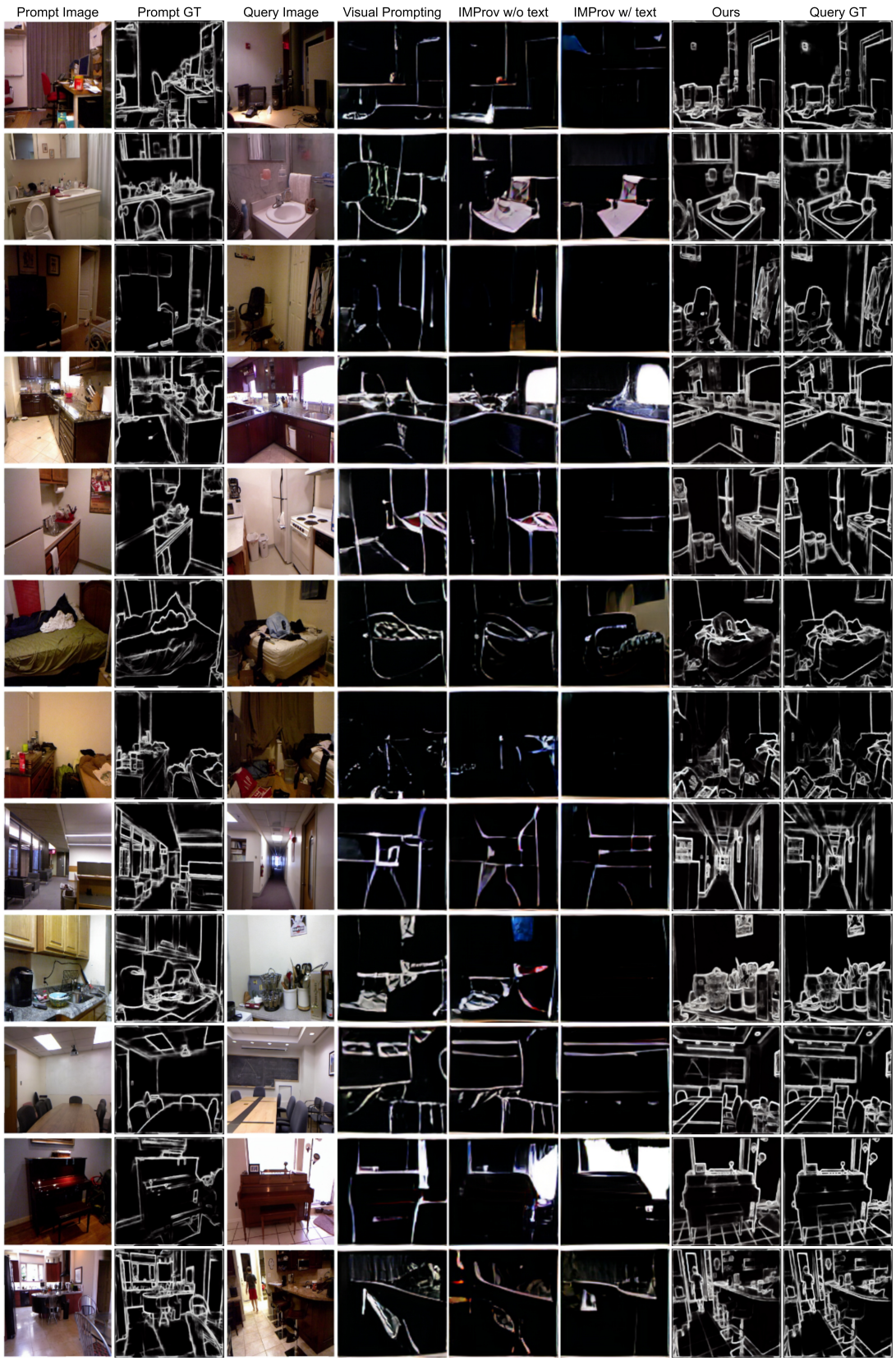


Figure 12. Qualitative examples of edge detection in comparison with Visual Prompting [3] and IMProv [33].

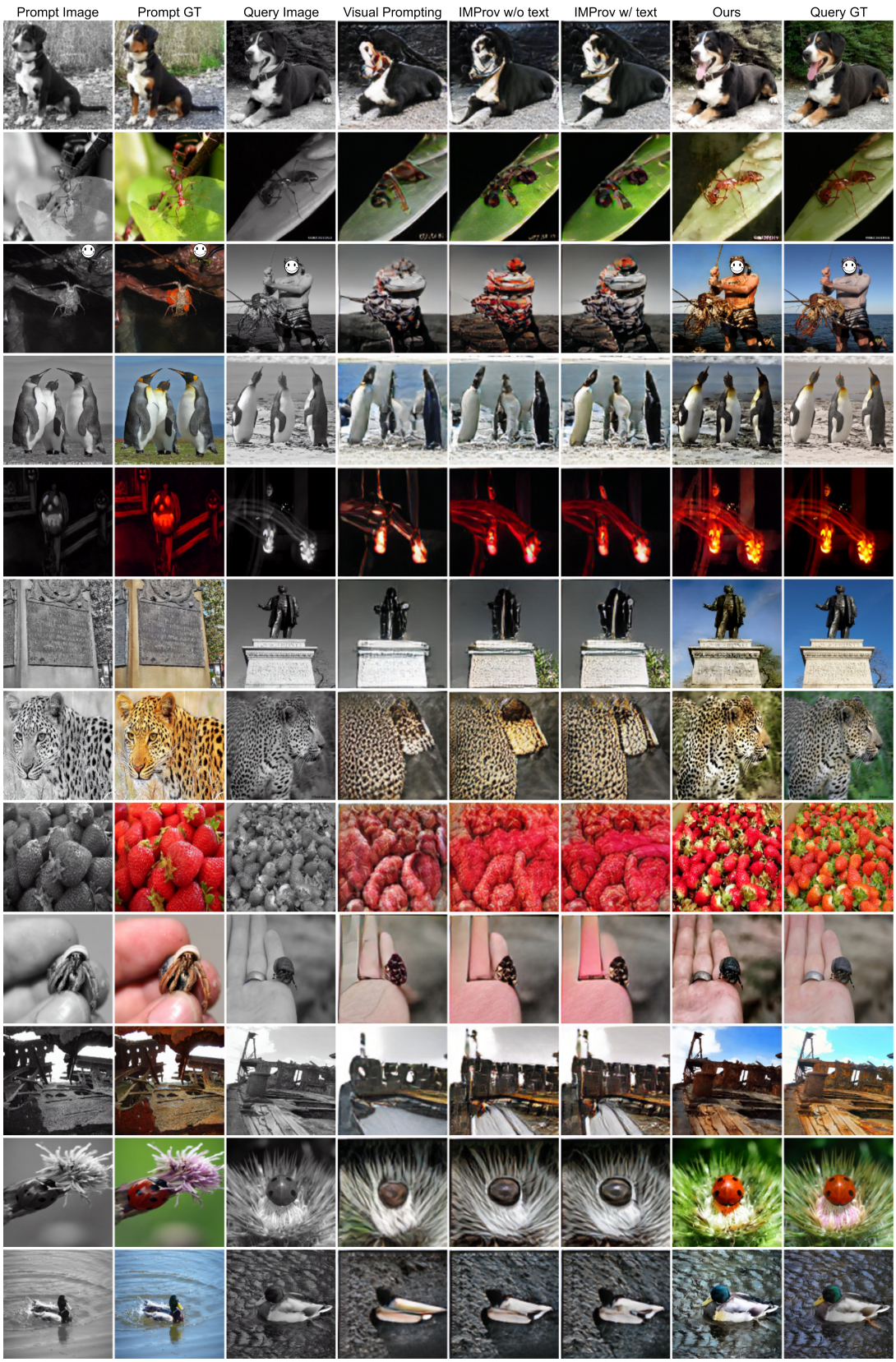


Figure 13. Qualitative examples of colorization in comparison with Visual Prompting [3] and IMProv [33].

References

[1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 5, 6

[2] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22861–22872, 2024. 2, 3, 4, 5

[3] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022. 1, 2, 3, 8, 9, 10, 11, 12, 13

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. 1

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1

[10] Zhongbin Fang, Xiangtai Li, Xia Li, Joachim M Buhmann, Chen Change Loy, and Mengyuan Liu. Explore in-context learning for 3d point cloud understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[11] Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama, 2023. 3

[12] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022. 1

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[14] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Xingzhe He, Hossam Isack, Abhishek Kar, Helge Rhodin, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised keypoints from pretrained diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22820–22830, 2024. 2

[15] 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. 3

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3

[17] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023. 7

[18] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 2

[19] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 1

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 3

[22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3

[23] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017. 2

[24] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from

rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 2, 3

[25] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022. 1

[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[27] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 1, 2, 3, 4, 5

[28] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 1

[29] Xinshun Wang, Zhongbin Fang, Xia Li, Xiangtai Li, Chen Chen, and Mengyuan Liu. Skeleton-in-context: Unified skeleton sequence modeling with in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2436–2446, 2024. 1

[30] Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. In-context learning unlocked for diffusion models. *Advances in Neural Information Processing Systems*, 36:8542–8562, 2023. 1, 2, 3, 4

[31] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. 1

[32] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 2

[33] Jiarui Xu, Yossi Gandelsman, Amir Bar, Jianwei Yang, Jianfeng Gao, Trevor Darrell, and Xiaolong Wang. Improv: Inpainting-based multimodal prompting for computer vision tasks. *arXiv preprint arXiv:2312.01771*, 2023. 1, 2, 3, 8, 9, 10, 11, 12, 13

[34] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024. 7

[35] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2

[36] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36:17773–17794, 2023. 2

[37] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 3