Learning A Low-Level Vision Generalist via Visual Task Prompt Supplementary Material

Anonymous Authors

1 INTRODUCTION

In this supplementary file, we present additional experiments and results to complement the main paper. Firstly, we explore the performance of various backbone networks to illustrate how we select the backbone network in our VPIP framework. Then, we present the quantitative comparison of using different information mechanism in the VPIP framework. Following this, we perform a more comprehensive comparison between GenLV and PromptGIP under several different settings. Afterwards, we show the computation cost of different parts of our model. Finally, we present more visual comparison results of our GenLV with the other methods.

2 EXPLORATION ON DIFFERENT IMAGE RESTORATION BACKBONE NETWORKS

The quantitative comparison of various backbone networks for different image restoration tasks is presented in Table 2. All the models are trained on the same multi-task restoration setting. We explore the different backbone networks based on image restoration because it has a clear quantitative evaluation scheme (i.e., PSNR/SSIM) and numerous low-level vision networks are designed based on it. As one can see that the overall performance of X-Restormer is the best, so it was selected as the backbone network in our method. It is noteworthy that the dehazing results show unusual performance, as both Restormer and X-Restormer perform much worse over the other comparison networks on this task. After inspecting the results, we find that these two models do not process some haze images. This suggests that these two networks may have fatal optimization difficulty in handling multi-task image restoration when dehazing is considered. Nevertheless, we can see that the introduction of task prompts effectively mitigates this problem, as depicted in Table 1 of the main paper.

3 EXPLORATION ON DIFFERENT PROMPT INTERACTION MECHANISMS

In this section, we explore the impact of different prompt interaction mechanisms in the proposed VPIP framework. A model without using prompt and two models using common modulation strategies (i.e., global feature modulation (GFM) [1] and spatial feature transform (SFT) [2]) for low-level vision tasks are compared. All models are trained on the same settings involving 30 tasks. In Table 3, we present the quantitative results of these models on restoration tasks. We can see that the model without using prompt performs much worse than other models. This is reasonable because the model cannot handle tasks with different target domains (e.g., edge detection), which greatly affects the optimization. Models with GFM and SFT can achieve much better performance than the model without prompt interaction, but their performance is still lower than our model. This suggests that the feature modulation schemes can also achieve task guidance to a certain extent, but their ability to learn the task representation is not as effective as prompt cross-attention.

4 COMPREHENSIVE COMPARISON BETWEEN PROMPTGIP AND OUR GENLV

We conduct comprehensive experiments and demonstrate the quantitative comparison of PromptGIP and GenLV under three training settings. Trained only for restoration tasks, we can see that our GenLV* can already outperforms PromptGIP*. This is mainly due to the powerful backbone that our VPIP framework can use. When the number of tasks increases to 15 (i.e., the PromptGIP setting), the performance of both PromptGIP$^\#$ and GenLV$^\#$ decreases slightly, while no more than 0.5dB. As the complexity of tasks continues to increase (i.e., the GenLV setting), we can find that the performance of both PromptGIP† and GenLV† drops significantly. However, the performance degradation of GenLV† on most tasks is within 1dB, while PromptGIP†'s performance degradation is around 2 to 4dB. This intuitively indicates that PromptGIP is more easily affected by the increase in the number and complexity of tasks. From the main paper, we illustrate that this is because PromptGIP is sensitive to the prompt content. When more tasks involving low-frequency processing are considered, its performance would be greatly affected. Many cases in the visual comparison (in the main paper Figure 5, Supp. Figure 1, 2 and 3) can more directly reflect this point.

5 COMPUTATION COST BREAKDOWN

In Table 1, we present the computation cost of different components of our GenLV model. The computational cost of the main network is similar to the original X-Restormer. The computational cost of prompt encoder comes from the residual blocks. The computational cost of the extra fusion part is from the prompt cross-attention modules in PCABs. Our prompt interaction scheme can bring considerable performance improvement at limited additional cost.

Table 1: Computation cost of different parts of our model.

Component	Params	MACs
Main Network	26.0M	192.3G
Prompt Encoder	2.8M	22.8G
Extra Fusion Part	3.9M	24.1G

6 MORE VISUAL RESULTS

In Figure 5 of the main paper, we show the visual results of 9 representative tasks. In Figure 1, Figure 2 and Figure 3, we present more visual results on the remaining 21 tasks.

REFERENCES

[1] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. 2020. Conditional sequential modulation for efficient global image retouching. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 679–695.

[2] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 606–615.

Table 2: Quantitative results (PSNR) of different image restoration backbone networks.

	GN	PN	S&P Noise	GB	JPEG	Ringing	R-L	Inpainting	SimpleRain	ComplexRain	Haze
RRDB	26.05	27.42	24.85	22.77	25.37	24.51	25.01	24.28	24.20	22.69	21.54
ViT	24.67	25.39	23.71	22.17	24.76	23.89	24.09	23.11	23.21	23.04	24.91
SwinIR	28.83	31.19	36.59	23.45	26.65	26.00	29.51	27.00	29.78	22.26	21.23
Restormer	28.56	31.21	35.42	24.16	26.65	27.00	29.83	27.77	29.38	24.16	14.83
X-Restormer	28.70	31.36	35.33	24.13	26.68	26.88	30.01	27.68	29.65	24.39	16.73

Table 3: Quantitative results of using different prompt interaction mechanisms.

	GN	PN	S&P Noise	GB	JPEG	Ringing	R-L	Inpainting	SimpleRain	ComplexRain	Haze
Without Prompt Interaction	24.30	25.85	26.54	20.63	19.26	16.88	17.87	22.57	19.56	21.55	14.22
Feature Modulation - GFM	27.76	30.04	32.42	22.52	25.68	24.55	25.65	26.12	25.66	24.56	28.55
Feature Modulation - SFT	28.03	30.58	33.20	22.82	26.04	24.72	26.46	26.42	26.48	24.46	28.17
Prompt Cross-Attention (ours)	28.28	30.80	33.47	23.14	26.06	25.50	27.51	26.66	27.68	25.13	28.65

Table 4: Comprehensive comparison between PromptGIP and GenLV under three different settings. ⋆: trained only for restoration tasks. #: trained on the PromptGIP setting (15 tasks). †: trained on the GenLV setting (30 tasks).

	GN	PN	S&P Noise	GB	JPEG	Ringing	R-L	Inpainting	SimpleRain	ComplexRain	Haze
PromptGIP*	26.48	27.76	28.08	22.88	25.86	25.69	27.05	25.28	25.79	24.33	24.55
GenLV*(ours)	28.99	31.69	36.63	24.58	26.91	27.74	31.50	28.11	31.10	24.71	28.91
PromptGIP$^{\#}$	26.22	27.29	27.49	22.77	25.38	25.45	26.79	25.02	25.46	24.08	24.32
GenLV$^{\#}$(ours)	28.92	31.58	36.32	24.33	26.55	27.55	31.11	27.86	30.35	24.47	28.73
PromptGIP†	23.63	23.98	25.05	20.84	22.21	23.86	24.94	22.11	23.16	21.79	21.90
GenLV†(ours)	28.28	30.80	33.47	23.14	26.06	25.50	27.51	26.66	27.68	25.13	28.65

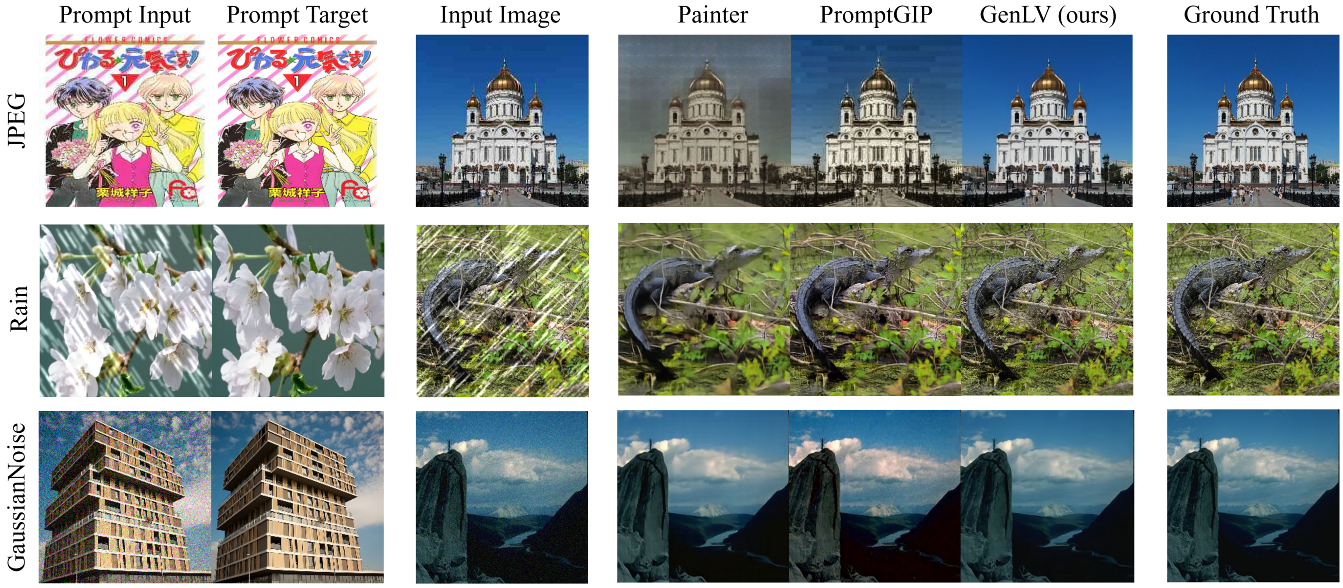

Figure 1: More visual results of different models on various low-level vision tasks.

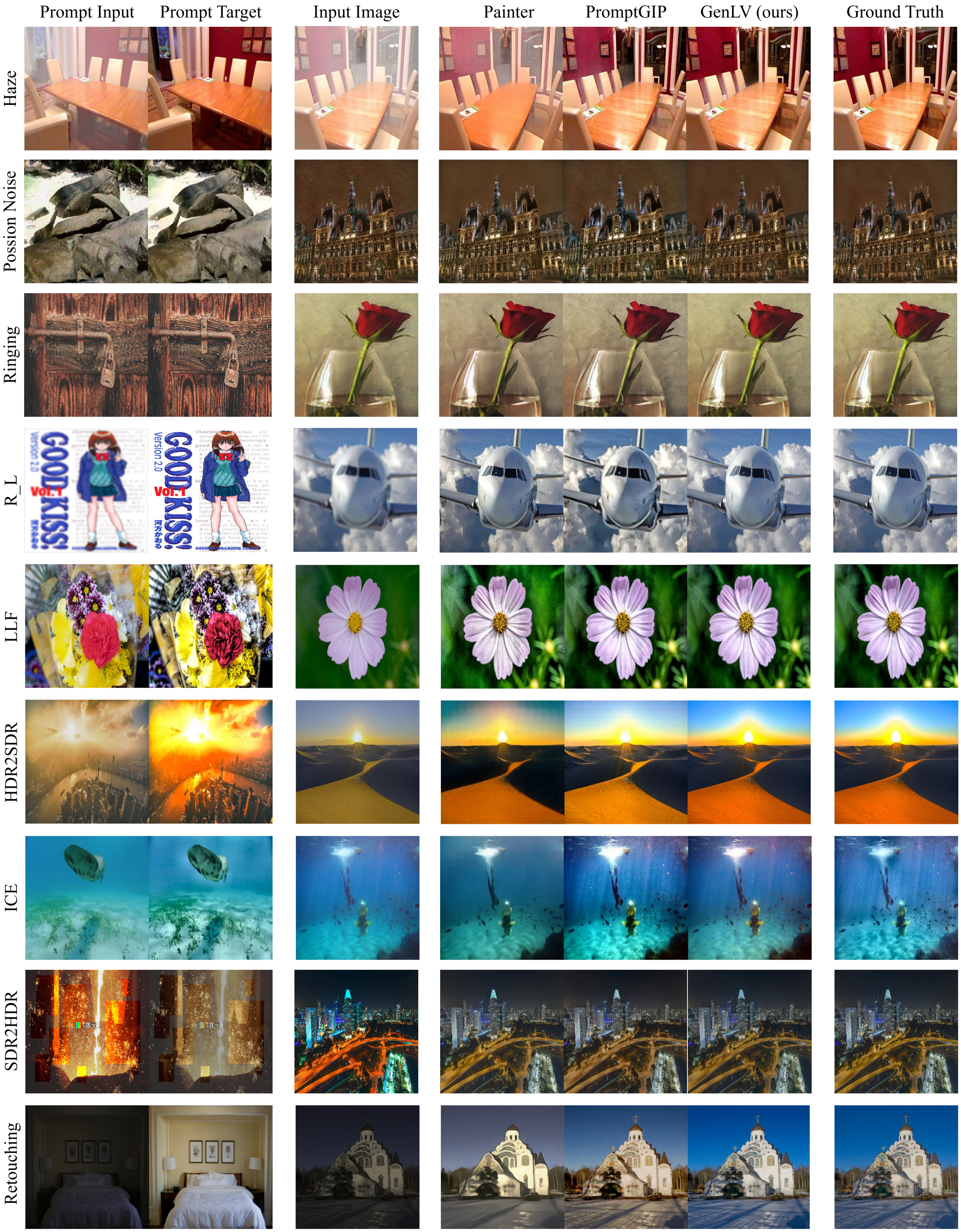

Figure 2: More visual results of different models on various low-level vision tasks.

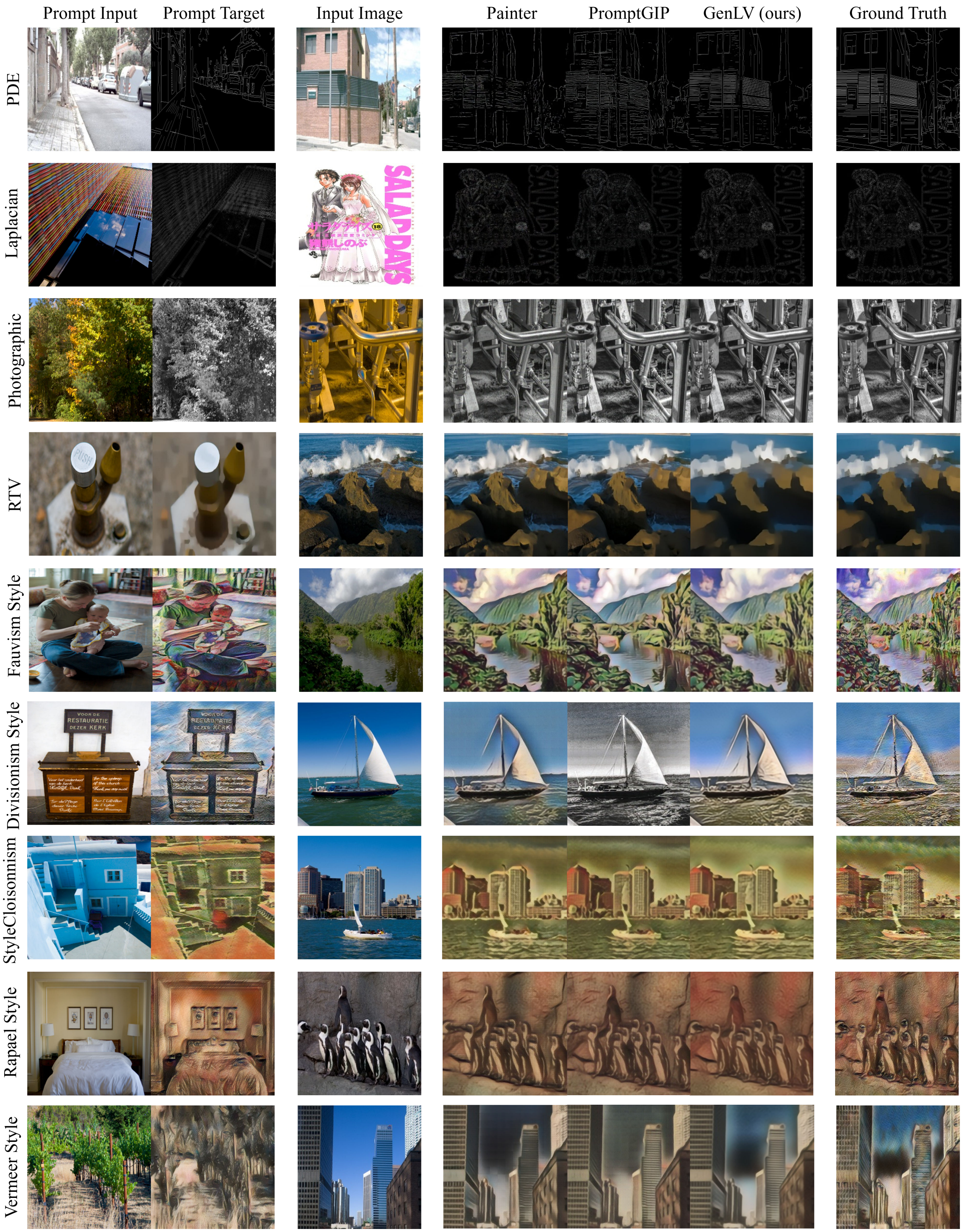

Figure 3: More visual results of different models on various low-level vision tasks.