Supplemental Material

Anonymous Author(s)
Affiliation
Address
email

Supplemental Material Outlines

A Detailed Related Work

A.1 UHD and All-in-One Image Restoration

Ultra-High Definition (UHD) image restoration focuses on recovering high-fidelity images from low-quality UHD observations. This task poses significant challenges due to the substantial computational overhead of processing vast data volumes and the stringent requirement to preserve fine high-frequency details. Direct application of deep learning models in pixel space is computationally prohibitive for UHD images. To address this, a prevalent approach is the downsample-enhance-upsample paradigm, where UHD images are downsampled, processed, and subsequently upsampled. For example, UHDFour [1] enables full-resolution inference on edge devices through 8x downsampling, UHDformer [2] leverages high-resolution features to guide low-resolution restoration, and UDR-Mixer [3] employs frequency feature modulation to enhance spatial feature recovery at lower resolutions. However, this downsampling strategy inevitably incurs information loss, particularly detrimental to UHD images with intricate textures, which subsequent processes struggle to fully recover, thus capping restoration quality.In parallel, an alternative efficiency-driven approach utilizes latent space models, notably Variational Autoencoders (VAEs), to shift the restoration process into a lower-dimensional latent space. For instance, DreamUHD [4] integrates a VAE framework with frequency augmentation and high-frequency injection to manage UHD details, while CD²-VAE [5] employs active feature decoupling and a reversible fusion network within a VAE structure to balance background consistency and degradation removal. These efforts highlight the pivotal role of meticulously designed latent space models in UHD restoration. Nevertheless, enhancing latent representation capacity to accommodate more complex degradations while mitigating inherent VAE limitations—such as the trade-off between generalization and reconstruction fidelity—remains a critical research frontier.

All-in-One image restoration aims to devise a unified model capable of addressing diverse, mixed, or unknown degradation types, necessitating exceptional generalization and adaptability [6–11]. Conventional methods typically design models tailored to specific degradations, a strategy impractical for real-world scenarios characterized by complex and variable degradation patterns. All-in-One models must contend with challenges such as managing degradation heterogeneity, mitigating conflicts among restoration sub-tasks, and achieving awareness of unknown degradations. Prevailing approaches predominantly adopt a strategy that combines a degradation-aware branch with an image restoration backbone. The degradation-aware branch is typically designed based on Mixture-of-Experts (MoE) [12, 11] or Prompting [13–15], while the image restoration backbone often employs established architectures such as Restormer or NAFNet. For instance, PromptIR [8] pioneered the introduction of a prompt-based degradation-aware branch into the all-in-one image restoration task, enhancing the model's adaptability to diverse degradations. Similarly, MoCE-IR [16], through its MoE design, enables specialized handling of different degradation inputs, further bolstering performance. Although these methods have demonstrated superior performance, their overall network efficiency remains relatively low (a challenge pertinent to the general 'Restoration Network' paradigms, as

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

conceptually outlined in the leftmost panel of main text Figure 1), making full-resolution inference on high-resolution images challenging on consumer-grade GPUs, which consequently limits their practical applicability.

UHD All-in-One image restoration merges the high-resolution demands of UHD with the multi-degradation capabilities of All-in-One tasks, presenting dual challenges: efficient processing of UHD data and robust multi-task optimization with strong generalization. UHD-Processor [17] established a benchmark for this intersection, achieving initial progress by combining a VAE framework with degradation-aware prompt learning. However, such prompt-based methods necessitate training specific prompts for known degradations, incurring additional parameter overhead and exhibiting limited generalization to novel degradation combinations or unseen types, as conceptually illustrated by the VAE framework incorporating a degradation-aware branch in the middle panel of Figure 1(main text). This limitation inspires our work to pursue a more streamlined solution: *harnessing the intrinsic strengths of VAEs to deliver efficient and generalizable UHD All-in-One restoration without relying on supplementary degradation-aware structures or extensive prompt parameters, a solution which can improve efficiency while ensuring the model's generalization* ability, as shown in the rightmost panel of Figure 1 in the main text.

A.2 Variational Autoencoders and Latent Space Optimization

Variational Autoencoders (VAEs) are extensively utilized in image restoration owing to their encoder-decoder architecture, which maps images into a low-dimensional latent space for reconstruction. The VAE's learning objective, the Evidence Lower Bound (ELBO), balances reconstruction fidelity against latent space regularity through the KL divergence term, constraining the latent distribution to a prior. This inherent trade-off significantly influences restoration quality: excessive regularization may yield insufficient latent information, resulting in blurry reconstructions, whereas prioritizing reconstruction can compromise latent space structure and generalization. VAEs have proven effective in tasks such as denoising, deblurring, and super-resolution, and they underpin Latent Diffusion Models (LDMs), where latent space quality dictates performance ceilings. However, the information compression inherent to VAEs often leads to the loss of high-frequency details, a pressing concern for UHD restoration. Efforts like FA-VAE [18] incorporate frequency-complementary modules and dynamic spectral losses to bolster high-frequency reconstruction, while Wavelet-VAE [19] and LiteVAE [20] leverage wavelet transforms to improve capture and recovery of high-frequency components. These developments underscore the necessity of tailoring VAEs to preserve high-frequency information for superior restoration outcomes.

latent space regularization To address the shortcomings of standard VAEs and enhance the quality and generalization of latent representations, researchers have explored diverse latent space regularization strategies. Beyond tuning the KL divergence weight as in β-VAE [21], techniques include contrastive learning to boost discriminability (e.g., Hi-CDL in CD²-VAE [5]), geometric regularization [22, 23] to shape the latent manifold, and diffusion-based decoders to elevate generation quality (e.g., ϵ-VAE [24]). These approaches aim to render latent representations more resilient to transformations such as degradations or geometric shifts. For instance, aligning VAE latent variables with features from robust pre-trained vision models like DINOv2 [25]—as seen in VAVAE [26]—injects valuable semantic priors, enhancing robustness and generalization. Furthermore, works such as REPA [27] and REPA-E [28] investigate feature alignment or alignment losses to optimize training, including end-to-end joint training of VAEs and LDMs, which also refines latent space structure. These insights suggest that integrating multiple regularization strategies, particularly by leveraging external priors and internal structural constraints, offers a promising avenue for crafting a latent space optimized for All-in-One UHD image restoration—a foundational principle of our Latent Harmony framework's initial stage.

B Experimental Details

B.1 Experimental Setups

UHD All-in-One Restoration Task. We follow the benchmark proposed in [17] for evaluating All-in-One methods in UHD scenarios. This benchmark includes UHD-LL [1], UHD-blur [29], UHD-haze [30], UHD-rain [31], UHD-haze [32], and UHD-noise [33]. The overall distribution of these datasets is presented in Table 1. We use PSNR [34], SSIM [34], and LPIPS [35] for evaluation,

Table 1: Dataset details and corresponding tasks.

Dataset	Training samples	Testing samples	Task
UHD-Snow	2,000	200	Desnowing
UHD-Blur	1,964	300	Deblurring
UHD-Rain	2,000	500	Deraining
UHD-LL	2,000	115	LLIE
UHD-Haze	2,290	231	Dehazing
UHD-Noise	2,000	500	Denoising

where **FS** denotes the method's capability for full-image inference on 4K resolution images using an NVIDIA RTX 3090 GPU.

Generalization Experiment Setup. To validate the comprehensive generalization capability of our model, we conducted verification in two primary aspects: First, generalization to unseen degradations: For the four-degradation UHD all-in-one setting, where the model was trained on tasks of low-light enhancement, image dehazing, image deblurring, and image denoising, we evaluated its performance on UHD-rain and UHD-snow, along with the additionally included UHD-moire [36] dataset. Second, generalization to composite degradations: We synthesized novel composite degradation scenarios by combining degradations derived from UHD-LL, UHD-haze, and UHD-noise datasets to evaluate the model's performance on these more complex, mixed-degradation inputs.

Standard-Resolution All-in-One Task. We utilized the original multi-degradation dataset curated from GenDeg [37], evaluating degradation types such as haze, rain, snow, motion blur, raindrops, and low-light conditions. This task primarily aims to validate the versatility of our method for standard-resolution all-in-one tasks. To this end, we applied our proposed LH-VAE to representative discriminative-based, LDM-based, and VAE-based methods in standard-resolution scenarios. Our evaluation primarily focused on perceptual metrics, employing LPIPS [35] and FID [38] for assessment.

B.2 Implementation Details

Baseline Construction. We first construct two distinct VAE baselines. For the first VAE_1, the objective is to achieve more generalizable representations, with an emphasis on encoding the semantic information of the image. In this configuration, the KL loss weight is set to 110^{-4}, the downsampling factor is 32, and the number of latent channels is 8. For the second VAE_2, the goal is to attain stronger reconstruction capabilities, enabling a more complete reconstruction of the input image. For this setup, the KL loss weight is significantly reduced to 110^{-8}, the downsampling factor is 8, and the number of latent channels is increased to 32.

Training Process. Our framework, which builds upon a foundation similar to VAE_2(the baseline VAE variant with enhanced reconstruction capabilities), is trained using 4 NVIDIA A8000 GPUs. The first stage focuses on training the VAE component (LH-VAE) through an image reconstruction task on clean images, while concurrently incorporating the progressive degradation perturbation strategy (PDPS). As the training step t increases, $sev(t)$ and $\beta(t)$ are employed to control the severity of synthetic degradations and the linear interpolation strength with paired real degraded images, respectively. In the second stage, training is performed using paired degraded and clean (ground truth) images for the restoration task. Specifically, the first stage utilizes a batch size of 16 and is trained for 64,000 steps; for the second stage, the batch size is 8, and training proceeds for 32,000 steps. Furthermore, FHF-LORA and PHF-LORA undergo alternating updates every 10 steps.

C Additional Experimental Results

C.1 Additional Visual Results for the UHD All-in-One Task.

We provide further visual comparative results for both the four-degradation and six-degradation settings, as presented in Figure 1 and Figure 2, respectively. These results demonstrate that our method achieves superior visual quality compared to other approaches.

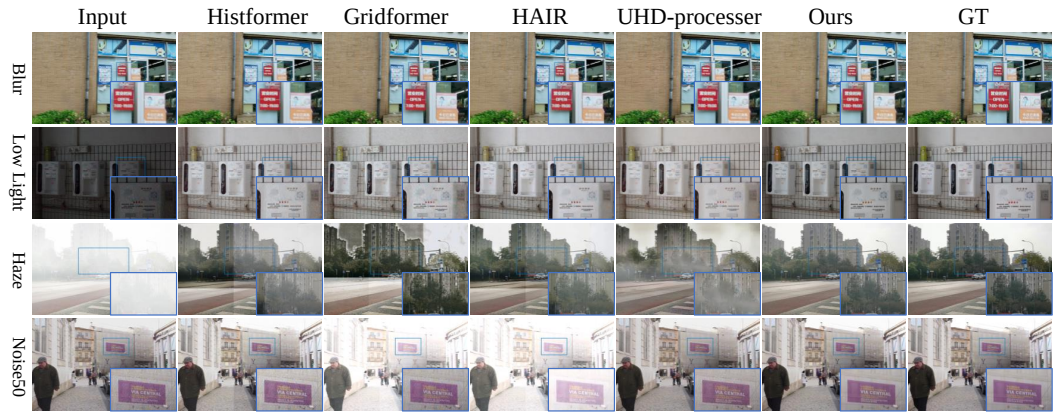

Figure 1: Additional Visual Results for the UHD All-in-One Task under the Four-Degradation Setting.

Figure 2: Additional Visual Results for the UHD All-in-One Task under the Six-Degradation Setting.

C.2 Single-Task Restoration on UHD Scenes

Quantitative results for single-task restoration in UHD scenarios are presented in Table 2. The results indicate that our design does not compromise the model's single-task performance and similarly achieves optimal quantitative results.

C.3 More Detailed Results on Adaptability in Standard-Resolution Scenarios

In Table 3, we present more comprehensive experimental results for standard-resolution tasks. Our proposed LH-VAE achieves performance improvements across various frameworks, demonstrating our method's versatility for standard-resolution all-in-one tasks.

C.4 More Detailed Verification of Generalization Capabilities

In Table 4, we present our complete generalization experiment results. The findings indicate that the generalization capability of our method is significantly superior to that of other approaches. This superiority is attributed to our method's design, which eschews degradation-specific branches,

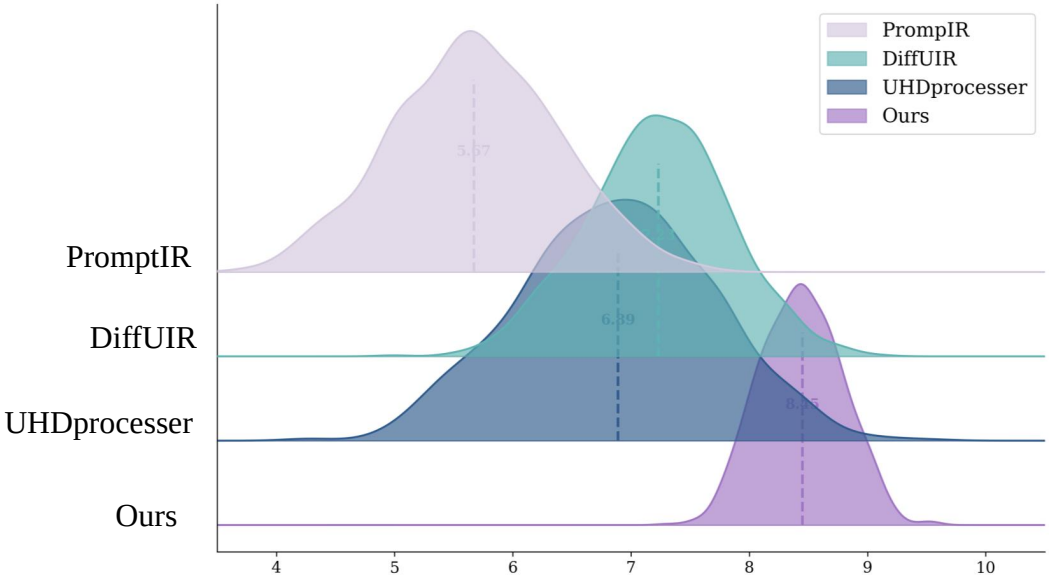

Figure 3: User Study Results Comparing Different Methods: These findings indicate that our method yields human subjective ratings significantly superior to those of other approaches.

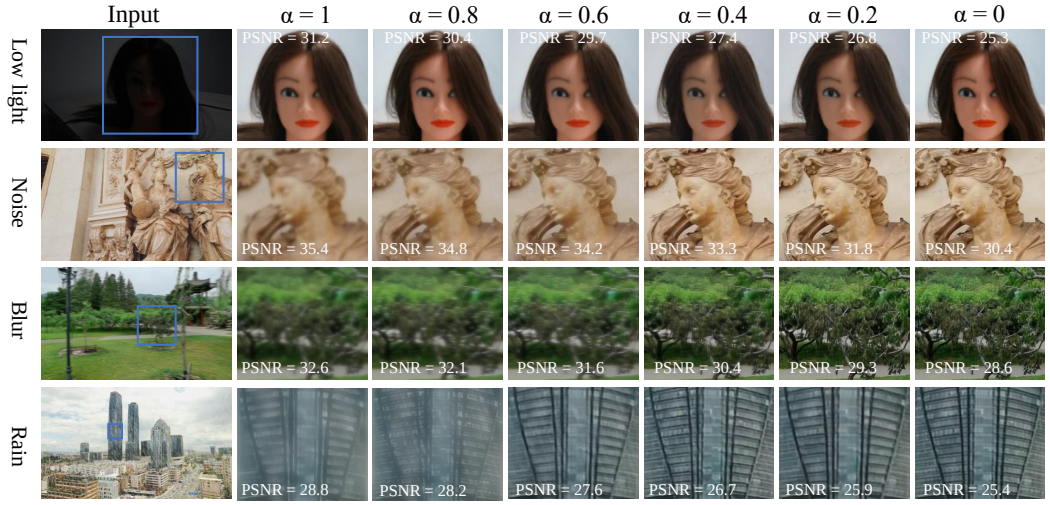

Figure 4: Additional Comparative Visual Results for Different Values of α: Higher α values yield enhanced fidelity, while lower α values lead to superior perceptual quality.

consequently enabling enhanced generalization performance. Our method achieves optimal results in scenarios involving both unseen degradations and composite degradations (.

C.5 User study

We further engaged 20 volunteers to evaluate the visual results of our method in comparison with other approaches. A 1-10 point rating scale was employed to assess the perceptual quality and discern differences between the methods. The evaluation results, as presented in Figure 3, indicate that our method aligns more closely with human aesthetic preferences, achieving the highest average scores.

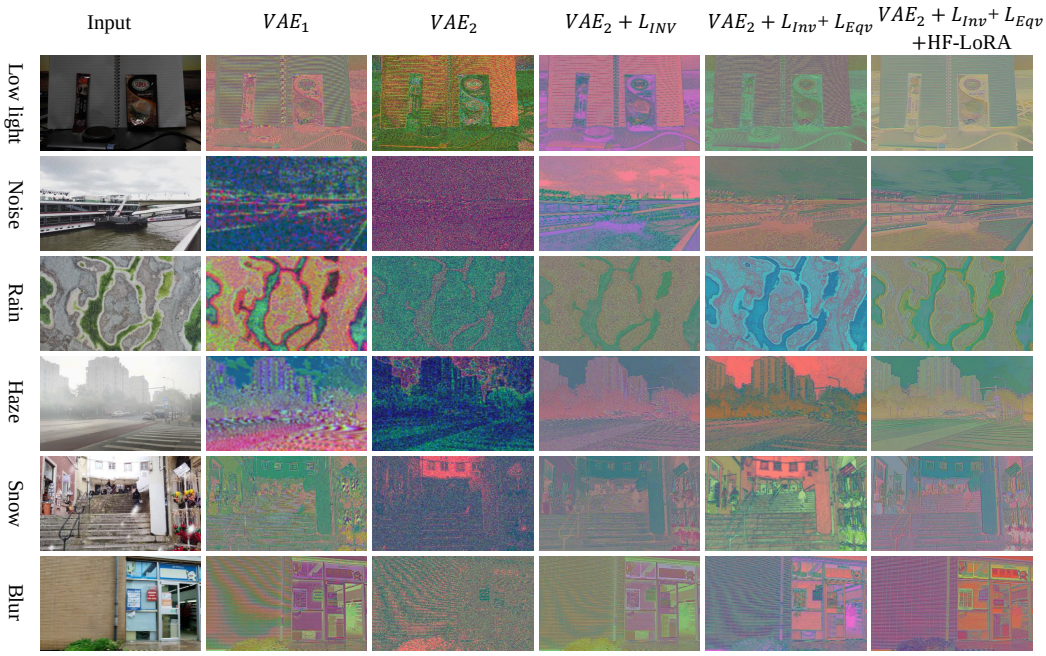

Figure 5: Comparative Visualization of Latent Space under Different Conditions.This analysis visually compares latent space characteristics under various conditions, where VAE_1 denotes a baseline with stronger generalization capabilities, and VAE_2 represents a baseline with enhanced reconstruction fidelity. Upon integrating the degradation-invariant visual semantic loss L_{Eqv}, the latent representations exhibit reduced interference from high-frequency noise and an increased focus on the semantic information of the input image. Following the incorporation of the latent space equivariance loss L_{Eqv}, its equivariant decoding regularization lessens the model's reliance on high-frequency components, thereby facilitating the reconstruction of richer textural details. Finally, when High-Frequency Low-Rank Adaptation (HF-LoRA) is combined and jointly optimized with the downstream task, the model's high-frequency information extraction is significantly enhanced, enabling clear representation of image details while effectively suppressing noise.

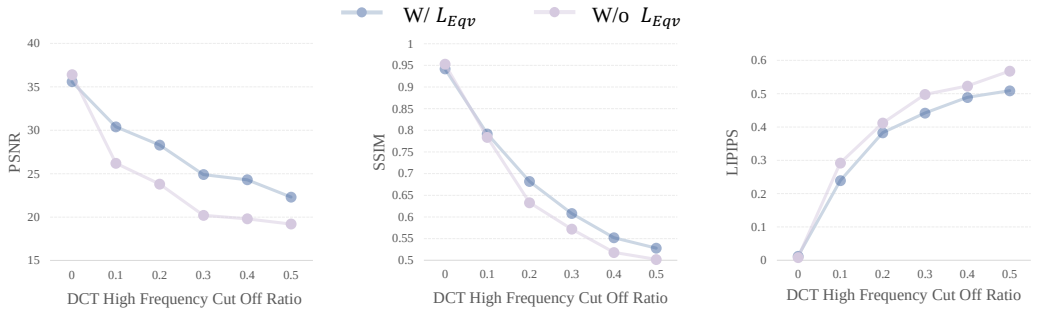

Figure 6: Ablation Study on High-Frequency Filtering with Equivariance Loss L_{Eqv}: Observations indicate that as the DCT High-Frequency Cut-off Ratio increases (implying more aggressive filtering of high frequencies), models trained with the equivariance loss L_{Eqv} exhibit stronger robustness and a reduced dependency on these high-frequency components.

D More Detailed Ablation Studies

D.1 Controllable Inference Parameter α.

Table 5 in the main text presents the impact of different α values on fidelity and perceptual metrics. We provide corresponding visual comparisons in Figure 4. These visualizations further illustrate that

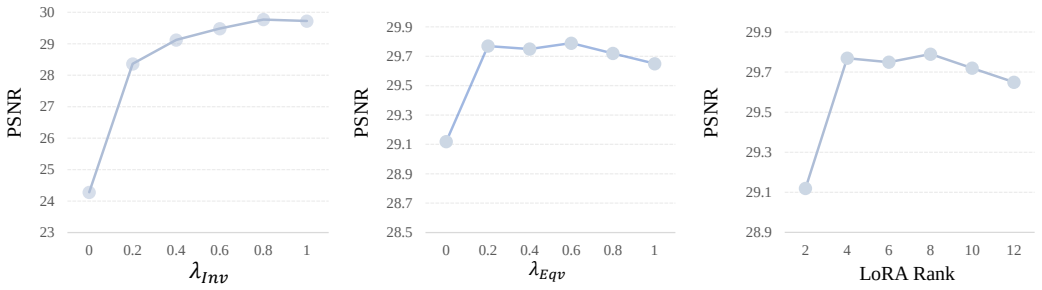

Figure 7: Ablation Studies on the Loss Weights $\lambda_{Inv}$and λ_{Eqv}, and the LoRA Rank.

Table 2: *Comparison to state-of-the-art for single degradations.* PSNR (dB, ↑),SSIM (↑) and LPIPS (↓). **Best** and **second best** performances are highlighted.

(a) *Low light*

Method	FS	UHD-LL		
Restormer [39]	✗	21.54	.843	.3608
LLformer [40]	✗	24.06	.858	.3516
UHDFour [1]	✓	26.22	.900	.2390
UHDformer [2]	✓	27.11	.927	.2240
LMAR [41]	✓	26.27	.919	.2248
AirNet [6]	✗	24.61	.827	.3841
PromptIR [8]	✗	24.81	.912	.3116
Histoformer [42]	✗	25.46	.894	.3213
HAIR [43]	✗	26.08	.901	.2964
UHDprocesser [17]	✓	**27.22**	**.929**	**.2042**
Ours	✓	**27.47**	**.932**	**.2028**

(b) *Deblurring*

Method	FS	UHD-blur		
Restormer [39]	✗	25.21	.752	.3695
Uformer [44]	✗	25.26	.751	.3851
Stripformer [45]	✗	25.05	.750	.3740
FFTformer [46]	✗	25.41	.757	.3708
UHDformer [2]	✓	28.82	.844	.2350
AirNet	✗	23.48	.824	.3942
PromptIR	✗	23.24	.829	.3754
Histoformer	✗	27.12	.848	.2986
HAIR	✗	26.74	.847	.3867
UHDprocesser [17]	✓	**28.91**	**.851**	**.2129**
Ours	✓	**28.96**	**.856**	**.2102**

(c) *Dehazing*

Method	FS	UHD-haze		
Restormer [39]	✗	12.72	.693	.4560
Uformer [44]	✗	19.83	.921	.4220
DehazeFormer [47]	✗	15.37	.737	.3998
TaylorFormer [48]	✗	20.99	.919	.3124
UHDformer [2]	✓	22.58	.942	.1188
AirNet	✗	18.44	.848	.4018
PromptIR	✗	20.21	.927	.3642
Histoformer	✗	21.72	{.948	.2942
HAIR	✗	20.38	.912	.2412
UHDprocesser [17]	✓	**23.24**	**.953**	**.1086**
Ours	✓	**23.62**	**.958**	**.1078**

higher values of α lead to enhanced fidelity, whereas lower values of α result in superior perceptual quality.

D.2 Comparative Visualization of Latent Space under Different Conditions.

In Figure 5, we present visualizations of the latent space under different conditions. This analysis visually compares latent space characteristics under various conditions, where VAE_1denotes a baseline with stronger generalization capabilities, and VAE_2 represents a baseline with enhanced reconstruction fidelity. VAE_1 tends to lose excessive high-frequency information, resulting in an overly smooth latent space, while VAE_2 encodes more high-frequency information, which can lead to the presence of substantial noise or degradation-related variations within its latent space.

Upon integrating the degradation-invariant visual semantic loss L_{Eqv}, which leverages features from a pre-trained DINOv2 model to enforce semantic consistency with clean reference images ,the latent representations exhibit reduced interference from high-frequency noise and an increased focus on the semantic information of the input image.

Following the incorporation of the latent space equivariance loss L_{Eqv}, the model's intrinsic reliance on specific high-frequency components within the latent representation is lessened. This regularization promotes scale robustness and a more balanced frequency characteristic in the latent space, thereby facilitating the decoder's ability to reconstruct richer and more consistent textural details from these refined encodings, particularly across different scales or views.

Finally, when High-Frequency Low-Rank Adaptation (HF-LoRA) is introduced in Stage Two and jointly optimized with the downstream restoration task. While the foundational latent structure established in Stage One is largely preserved through selective gradient propagation, this targeted fine-tuning ensures that the latent representations are more effectively utilized for the precise extraction of high-frequency details. This ultimately enables a clearer representation of image details and effective noise suppression in the final restored output, demonstrating the synergistic effect of the regularized latent space and the controllable refinement stage.

Table 3: Adaptability in Standard-Resolution Scenarios.Comparisons use LPIPS and FID scores, with lower values indicating superior performance.

Type	Method	Haze	Rain	Snow	Motion Blur	Raindrop	Low-light
Discriminative-based	NAFNet [49]	0.190/118.22	0.074/21.84	0.067/8.20	0.136/28.72	0.085/39.91	0.349/172.36
	NAFNet /w Ours	0.178/112.12	0.072/21.72	0.062/8.11	0.133/28.42	0.072/32.89	0.342/162.72
	PromptIR [8]	0.309/141.05	0.097/32.61	0.100/18.34	0.163/35.79	0.189/84.48	0.421/189.87
	PromptIR /w Ours	0.224/121.12	0.086/28.68	0.092/17.12	0.161/35.12	0.182/76.84	0.378/172.59
LDM-based	ResShift [50]	0.284/129.43	0.072/23.97	0.129/18.03	0.132/24.78	0.142/52.33	0.393/164.82
	ResShift/w Ours	0.262/125.73	0.069/23.57	0.121/17.87	0.124/23.98	0.138/50.43	0.386/160.54
	Diff-Plugin [51]	0.340/143.66	0.165/39.71	0.178/18.08	0.147/37.68	0.185/60.64	0.466/167.63
	Diff-Plugin/w Ours	0.321/131.12	0.162/39.43	0.174/18.02	0.138/35.42	0.146/44.26	0.432/152.28
VAE-based	OPR [52]	0.384/149.45	0.152/36.84	0.152/46.68	0.223/54.56	0.132/44.76	0.576/212.33
	OPR/w Ours	0.253/112.14	0.106/31.20	0.118/28.66	0.189/42.64	0.098/33.24	0.393/163.26
	CosAE [53]	0.328/148.78	0.146/38.27	0.162/16.78	0.186/41.28	0.182/49.27	0.482/182.24
	CosAE/w Ours	0.224/128.12	0.098/28.79	0.121/11.56	0.168/36.22	0.119/40.62	0.382/159.83

Table 4: Generalization Verification. PSNR (dB, ↑), SSIM (↑), and LPIPS (↓) are reported.

Method	*Unseen*									*Composite Degradation*								
	UHD-rain			UHD-snow			UHD-moire			LLIE+Noise			Haze+LLIE			Noise+Blur		
AirNet	23.32	0.842	0.378	24.19	0.912	0.282	14.82	0.782	0.468	19.02	0.825	0.443	15.82	0.832	0.397	18.12	0.822	0.368
PromptIR	24.53	0.843	0.362	22.78	0.898	0.263	17.42	0.782	0.463	17.86	0.818	0.433	16.32	0.844	0.412	19.24	0.883	0.483
DiffUIR-L	26.92	0.878	0.284	26.32	0.913	0.209	18.07	0.763	0.368	19.18	0.803	0.392	18.44	0.892	0.364	22.83	0.872	0.338
Histoformer	24.12	0.834	0.402	25.82	0.891	0.292	16.87	0.698	0.492	17.68	0.823	0.473	15.88	0.812	0.398	17.67	0.823	0.423
Gridformer	23.94	0.802	0.423	21.68	0.884	0.283	17.42	0.768	0.436	18.34	0.792	0.454	17.43	0.883	0.498	19.89	0.872	0.483
adaIR	24.86	0.881	0.352	25.98	0.924	0.279	17.04	0.744	0.478	18.42	0.818	0.508	17.08	0.839	0.472	20.21	0.874	0.493
HAIR	24.32	0.824	0.392	25.43	0.903	0.223	17.28	0.798	0.446	18.12	0.812	0.492	16.72	0.862	0.439	18.42	0.854	0.471
UHD-processer	22.72	0.812	0.342	21.82	0.918	0.267	14.32	0.778	0.489	13.28	0.842	0.428	12.38	0.872	0.462	18.28	0.824	0.492
Ours	**28.13**	**0.892**	**0.233**	**28.92**	**0.967**	**0.184**	**19.26**	**0.898**	**0.326**	**20.33**	**0.882**	**0.342**	**19.82**	**0.904**	**0.328**	**24.28**	**0.898**	**0.278**

D.3 Ablation Study on High-Frequency Filtering with Equivariance Loss

The Equivariance Loss L_{Eqv} is introduced with the objective of reducing the decoding process's dependence on excessive high-frequency components. Therefore, to validate the efficacy of this loss, we compare its impact on the Stage One reconstruction performance under varying high-frequency cut-off ratios applied in the latent space. As shown in Figure 6, while reconstruction quality generally degrades as the high-frequency cut-off ratio in the latent space increases (i.e., more high-frequency components are removed from the latent representation), the performance of the model variant incorporating L_{Eqv} declines more slowly. This observation indicates that L_{Eqv} successfully mitigates the decoding process's reliance on high-frequency components, thereby yielding more robust reconstruction results.

D.4 Impact of Loss Weightings and LoRA Rank.

The ablation study on different loss weights λ_{Inv}, λ_{Eqv} and LoRA rank is presented in Figure 7. Based on these results, we set $\lambda_{Inv} = 0.8$, $\lambda_{Eqv} = 0.6$, and the LoRA rank to 0.6.

D.5 Ablation on Different Visual Foundation Model.

Our Invariance Visual Semantic LossL_{Inv} utilizes the DINOv2 model for semantic alignment. As shown in Table 5, consistent performance improvements are achieved when employing different visual foundation models for this loss component, which validates the general applicability and robustness of our semantic alignment strategy.

Table 5: Comparative Results of Employing Different Visual Foundation Models for the Invariance Visual Semantic Loss L_{Inv}: Consistent performance improvements are observed across various visual foundation models.

Method	**MAE** [54]	**SAM** [55]	**CLIP** [56]	**SigLIP** [57]	**DINOv2** [25]
PSNR	29.62	29.66	29.57	29.64	**29.70**
SSIM	0.872	0.880	0.875	**0.882**	0.877

E Limitations And Broader impacts

E.1 Broader impacts

Latent Harmony's advancements in Ultra-High Definition (UHD) image restoration offer significant societal benefits by enhancing visual content quality and utility across diverse high-resolution-dependent domains such as professional media, digital archiving, and specialized imaging. Its efficiency, generalization, and controllable fidelity-perception balance make it a valuable tool for improved image analysis, aiding human observation and downstream automated systems, and potentially driving progress in broader computer vision applications.

However, potential negative impacts warrant consideration. A primary concern is training data bias, which could lead to suboptimal or inequitable restoration outcomes if not addressed through meticulous data curation and fairness evaluations. Additionally, the high quality of restored UHD images raises the possibility of misuse, for instance, in making manipulated media appear more authentic. This underscores the need for transparency, robust detection methods, and careful ethical considerations regarding privacy and responsible deployment, especially in sensitive contexts. While Latent Harmony is fundamentally a restoration tool aimed at removing degradation rather than a de novo content synthesizer—which frames its direct risks as perfecting existing imagery rather than fabricating realities—vigilance remains essential.

E.2 Limitation

While the proposed Latent Harmony framework achieves promising results in UHD image restoration, it still has certain limitations. Currently, the fidelity-perception trade-off parameter α within the framework requires manual specification by the user, often based on experience. Determining an optimal α value automatically based on varying input image characteristics and degradation types remains an open challenge. This necessitates the integration of more sophisticated image quality assessment strategies capable of concurrently evaluating both fidelity and perceptual aspects. Future work exploring adaptive strategies to adjust the α parameter based on image content and restoration requirements represents a significant direction for enhancing the framework's practicality and intelligence.

References

[1] C. Li, C.-L. Guo, M. Zhou, Z. Liang, S. Zhou, R. Feng, and C. C. Loy, "Embedding fourier for ultra-high-definition low-light image enhancement," in *ICLR*, 2023.

[2] C. Wang, J. Pan, W. Wang, G. Fu, S. Liang, M. Wang, X.-M. Wu, and J. Liu, "Correlation matching transformation transformers for uhd image restoration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5336–5344.

[3] H. Chen, X. Chen, C. Wu, Z. Zheng, J. Pan, and X. Fu, "Towards ultra-high-definition image deraining: A benchmark and an efficient method," *arXiv preprint arXiv:2405.17074*, 2024.

[4] Y. Liu, D. Li, J. Xiao, Y. Bao, S. Xu, and X. Fu, "Dreamuhd: Frequency enhanced variational autoencoder for ultra-high-definition image restoration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 6, 2025, pp. 5712–5720.

[5] Y. Liu, D. Li, Y. Ma, J. Huang, W. Zhang, X. Fu, and Z.-j. Zha, "Decouple to reconstruct: High quality uhd restoration via active feature disentanglement and reversible fusion," *arXiv preprint arXiv:2503.12764*, 2025.

[6] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-In-One Image Restoration for Unknown Corruption," in *IEEE Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, Jun. 2022.

[7] Y. Cui, S. W. Zamir, S. Khan, A. Knoll, M. Shah, and F. S. Khan, "Adair: Adaptive all-in-one image restoration via frequency mining and modulation," 2024.

[8] V. Potlapalli, S. W. Zamir, S. Khan, and F. Khan, "Promptir: Prompting for all-in-one image restoration," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[9] X. Kong, C. Dong, and L. Zhang, "Towards effective multiple-in-one image restoration: A sequential and prompt learning strategy," *arXiv preprint arXiv:2401.03379*, 2024.

[10] M. Yao, R. Xu, Y. Guan, J. Huang, and Z. Xiong, "Neural degradation representation learning for all-in-one image restoration," *IEEE Transactions on Image Processing*, 2024.

[11] X. Yu, S. Zhou, H. Li, and L. Zhu, "Multi-expert adaptive selection: Task-balancing for all-in-one image restoration," 2024. [Online]. Available: https://arxiv.org/abs/2407.19139

[12] B. Lin, Z. Tang, Y. Ye, J. Cui, B. Zhu, P. Jin, J. Huang, J. Zhang, Y. Pang, M. Ning *et al.*, "Moe-llava: Mixture of experts for large vision-language models," *arXiv preprint arXiv:2401.15947*, 2024.

[13] H. Gao, J. Yang, N. Wang, J. Yang, Y. Zhang, and D. Dang, "Prompt-based all-in-one image restoration using cnns and transformer," *arXiv preprint arXiv:2309.03063*, 2023.

[14] J. Ma, T. Cheng, G. Wang, Q. Zhang, X. Wang, and L. Zhang, "Prores: Exploring degradation-aware visual prompt for universal image restoration," *arXiv preprint arXiv:2306.13653*, 2023.

[15] Z. Li, Y. Lei, C. Ma, J. Zhang, and H. Shan, "Prompt-in-prompt learning for universal image restoration," 2023. [Online]. Available: https://arxiv.org/abs/2312.05038

[16] E. Zamfir, Z. Wu, N. Mehta, Y. Tan, D. P. Paudel, Y. Zhang, and R. Timofte, "Complexity experts are task-discriminative learners for any image restoration," 2024.

[17] Y. Liu, D. Li, X. Fu, X. Lu, J. Huang, and Z. jun Zha, "Uhd-processer: Unified uhd image restoration with progressive frequency learning and degradation-aware prompts," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[18] X. Lin, Y. Li, J. Hsiao, C. Ho, and Y. Kong, "Catch missing details: Image reconstruction with frequency augmented variational autoencoder," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[19] A. Kiruluta, "Wavelet-based variational autoencoders for high-resolution image generation," *arXiv preprint arXiv:2504.13214*, 2025.

[20] S. Sadat, J. Buhmann, D. Bradley, O. Hilliges, and R. M. Weber, "Litevae: Lightweight and efficient variational autoencoders for latent diffusion models," *arXiv preprint arXiv:2405.14477*, 2024.

[21] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2017.

[22] T. Kouzelis, I. Kakogeorgiou, S. Gidaris, and N. Komodakis, "Eq-vae: Equivariance regularized latent space for improved generative image modeling," in *arxiv*, 2025.

[23] Y. Zhou, Z. Xiao, S. Yang, and X. Pan, "Alias-free latent diffusion models:improving fractional shift equivariance of diffusion latent space," 2025. [Online]. Available: https://arxiv.org/abs/2503.09419

[24] L. Zhao, S. Woo, Z. Wan, Y. Li, H. Zhang, B. Gong, H. Adam, X. Jia, and T. Liu, "Epsilon-vae: Denoising as visual decoding," 2025. [Online]. Available: https://arxiv.org/abs/2410.04081

[25] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024, featured Certification. [Online]. Available: https://openreview.net/forum?id=a68SUt6zFt

[26] J. Yao, B. Yang, and X. Wang, "Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[27] S. Yu, S. Kwak, H. Jang, J. Jeong, J. Huang, J. Shin, and S. Xie, "Representation alignment for generation: Training diffusion transformers is easier than you think," in *International Conference on Learning Representations*, 2025.

[28] X. Leng, J. Singh, Y. Hou, Z. Xing, S. Xie, and L. Zheng, "Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers," 2025. [Online]. Available: https://arxiv.org/abs/2504.10483

[29] S. Deng, W. Ren, Y. Yan, T. Wang, F. Song, and X. Cao, "Multi-scale separable network for ultra-high-definition video deblurring," in *the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 14 030–14 039.

[30] Z. Zheng, W. Ren, X. Cao, X. Hu, T. Wang, F. Song, and X. Jia, "Ultra-high-definition image dehazing via multi-guided bilateral learning," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 180–16 189.

[31] H. Chen, X. Chen, C. Wu, Z. Zheng, J. Pan, and X. Fu, "Towards ultra-high-definition image deraining: A benchmark and an efficient method," 2024. [Online]. Available: https://arxiv.org/abs/2405.17074

[32] L. Wang, C. Wang, J. Pan, X. Liu, W. Zhou, X. Sun, W. Wang, and Z. Su, "Ultra-high-definition image restoration: New benchmarks and a dual interaction prior-driven solution," 2024. [Online]. Available: https://arxiv.org/abs/2406.13607

[33] K. Zhang, D. Li, W. Luo, W. Ren, B. Stenger, W. Liu, H. Li, and Y. Ming-Hsuan, "Benchmarking ultra-high-definition image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[34] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.

[35] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[36] X. Yu, P. Dai, W. Li, L. Ma, J. Shen, J. Li, and X. Qi, "Towards efficient and scale-robust ultra-high-definition image demoiréing," in *European Conference on Computer Vision*. Springer, 2022, pp. 646–662.

[37] S. Rajagopalan, N. G. Nair, J. N. Paranjape, and V. M. Patel, "Gendeg: Diffusion-based degradation synthesis for generalizable all-in-one image restoration," *arXiv preprint arXiv:2411.17687*, 2024.

[38] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[39] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *CVPR*, 2022.

[40] T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, and T. Lu, "Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 2654–2662.

[41] W. Yu, J. Huang, B. Li, K. Zheng, Q. Zhu, M. Zhou, and F. Zhao, "Empowering resampling operation for ultra-high-definition image enhancement with model-aware guidance," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 25 722–25 731.

[42] S. Sun, W. Ren, X. Gao, R. Wang, and X. Cao, "Restoring images in adverse weather conditions via histogram transformer," in *European Conference on Computer Vision*. Springer, 2025, pp. 111–129.

[43] J. Cao, Y. Cao, L. Pang, D. Meng, and X. Cao, "Hair: Hypernetworks-based all-in-one image restoration," 2024. [Online]. Available: https://arxiv.org/abs/2408.08091

[44] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 17 683–17 693.

[45] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin, "Stripformer: Strip transformer for fast image deblurring," in *European conference on computer vision*. Springer, 2022, pp. 146–162.

[46] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, "Efficient frequency domain-based transformers for high-quality image deblurring," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 5886–5895.

[47] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE Transactions on Image Processing*, vol. 32, pp. 1927–1941, 2023.

[48] Y. Qiu, K. Zhang, C. Wang, W. Luo, H. Li, and Z. Jin, "Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing," 2023.

[49] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," *arXiv preprint arXiv:2204.04676*, 2022.

[50] Z. Yue, J. Wang, and C. C. Loy, "Resshift: Efficient diffusion model for image super-resolution by residual shifting," *Advances in Neural Information Processing Systems*, vol. 36, pp. 13 294–13 307, 2023.

[51] Y. Liu, Z. Ke, F. Liu, N. Zhao, and R. W. Lau, "Diff-plugin: Revitalizing details for diffusion-based low-level tasks," in *CVPR*, 2024.

[52] Z. Wan, B. Zhang, D. Chen, P. Zhang, D. Chen, F. Wen, and J. Liao, "Old photo restoration via deep latent space translation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2071–2087, 2023.

[53] J. K. Sifei Liu, Shalini De Mello, "Cosae: Learnable fourier series for image restoration," *NeurIPS*, 2024.

[54] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[55] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[56] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[57] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.