Lumina-OmniLV: A Unified Multimodal Framework for General Low-Level Vision

Anonymous ECCV 2026 Submission

Paper ID #3952

Abstract. We present **Lunima-OmniLV** (abbreviated as **OmniLV**), a universal multimodal multi-task framework for low-level vision that addresses over 100 sub-tasks across four major categories, including image restoration, image enhancement, weak-semantic dense prediction, and stylization. OmniLV leverages both textual and visual prompts to offer flexible, user-friendly interactions. Built on Diffusion Transformer (DiT)-based generative priors, our framework supports arbitrary resolutions — achieving optimal performance at 1K resolution — while preserving fine-grained details and high fidelity. Through extensive experiments, we demonstrate that separately encoding text and visual instructions, combined with co-training using shallow feature control, is essential to mitigate task ambiguity and enhance multi-task generalization. Our findings also reveal that integrating high-level generative tasks into low-level vision models can compromise detail-sensitive restoration. These insights pave the way for more robust and generalizable low-level vision systems.

1 Introduction

The rapid evolution of large-scale foundation models has revolutionized artificial intelligence, demonstrating remarkable generalization and multi-task capabilities across various domains. Unified frameworks such as GPT-4V [2], InternVL [17–19], Flamingo [5], OmniGen [62, 64], and Bagel [20] have showcased impressive performance by leveraging large-scale pretraining on multi-modal datasets. These models excel in semantic-driven high-level vision tasks, such as image classification, image understanding, visual generation and editing. In contrast, the development of unified models for low-level vision remains largely fragmented and underexplored.

Low-level vision encompasses a broad spectrum of tasks, including image restoration [11, 14, 21, 38, 39, 70, 76] , image enhancement [8, 13, 15, 56, 73], style transfer [25, 29], and weak-semantic dense prediction [33, 49, 66] (e.g., edge detection, depth estimation, normal map estimation). Unlike high-level vision tasks that rely on predefined semantic understanding, most low-level vision tasks do not require explicit object-level reasoning. Instead, they focus on pixel-level fidelity, fine-grained texture reconstruction, and feature extraction. This distinction makes the unification of low-level vision tasks particularly challenging, as different tasks often operate in vastly different output domains.

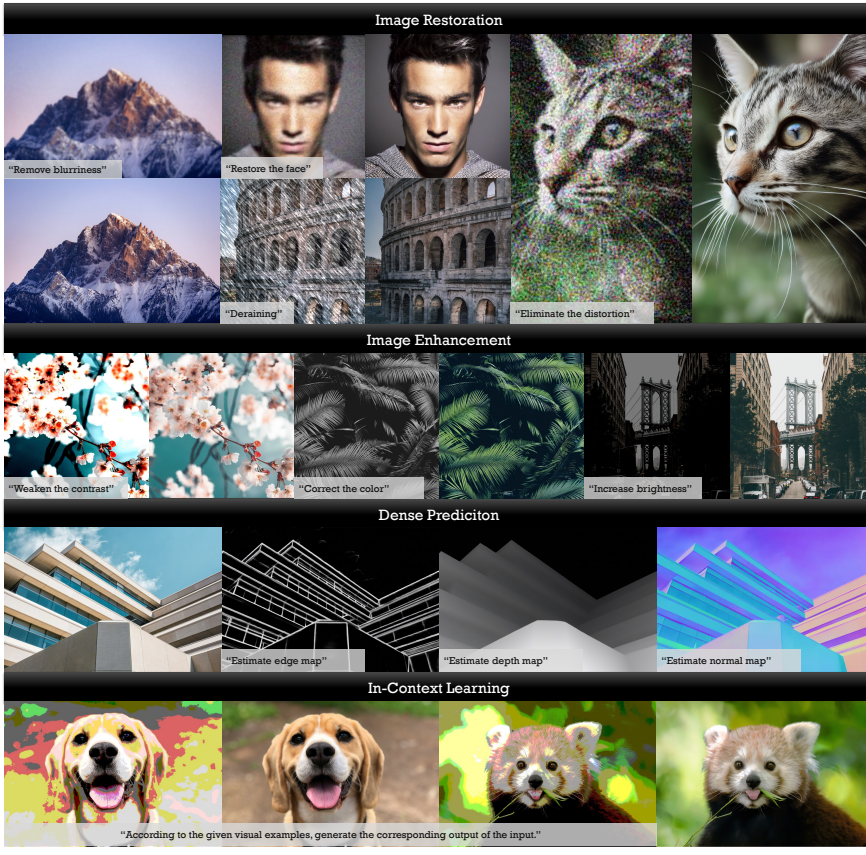

Fig. 1: Illustration of OmniLV's versatile capabilities. As a universal framework, OmniLV is capable of handling a wide variety of low-level vision tasks within a single model, which adapts to diverse input-output domains and generates high-fidelity results.

Existing approaches to low-level vision remain limited in generalization, usability, and scalability. Task-specific models [14,36] are designed to handle a single task (e.g., denoising, deblurring, super-resolution), requiring extensive model redesigning and retraining to adapt to new tasks. All-in-one restoration models, such as AirNet [36], PromptIR [48], and OneRestore [26], integrate multiple restoration tasks within a single framework, yet remain restricted to in-domain restoration, unable to generalize to cross-domain tasks such as feature extraction or style transfer. Visual-prompt-based models, such as PromptGIP [40] and GenLV [12, 16], extend to cross-domain tasks using image prompt pairs, but require carefully crafted prompts, making them less intuitive and user-friendly compared to text-driven interaction. Furthermore, many existing methods operate only on fixed-resolution images, severely limiting their flexibility and real-

world applicability. To summarize, high-resolution image processing still remains challenging, leaving ample room for improvement in task adaptability.

Given the inherent complexity and diversity of low-level vision, developing a truly universal model must handle multiple task domains while reliably preserving fine-grained details and high fidelity. A key requirement for such a model is flexible interaction mechanisms. While text-based instructions offer a convenient and intuitive way to specify tasks (e.g., "remove noise from this image", "enhance brightness", and "estimate the Canny edge"), certain tasks — such as style transfer — are difficult to define using text alone. Visual prompts, provided in the form of exemplar image pairs, provide an effective alternative by allowing the model to infer complex, task-specific transformations through visual analogy. Thus, an ideal general low-level vision model should integrate both textual and visual prompts for versatile and user-friendly task execution.

To address these challenges, we propose **OmniLV**, a universal multimodal multi-task framework for low-level vision, capable of handling over 100 sub-tasks via both textual and visual prompts. Built on Diffusion Transformer (DiT)-based generative priors [22, 28, 47, 51], our model significantly improves generalization and output quality across tasks. Unlike prior models constrained to fixed resolutions, our framework supports arbitrary resolutions, achieving optimal performance at 1K resolution. We systematically explore multimodal fusion strategies and propose a simple yet effective design that prevents task misinterpretation issues. Fig. 1 presents the versatile capabilities of OmniLV.

Throughout the development of OmniLV, we have gained several key insights that shape the design of a robust and generalizable low-level vision model. First, we find that separately encoding text-based and visual instructions is crucial for preventing task ambiguity, as naive fusion can lead to task misinterpretations (Sec. 3.1). Additionally, co-training the base model with shallow feature control proves to be an effective strategy for enhancing multi-task generalization (Sec. 3.1). Furthermore, incorporating high-level generative or editing tasks into a low-level vision model significantly compromises fidelity, particularly in detail-sensitive restoration tasks (Sec. 4.2). These findings highlight the need for dedicated multimodal architectures tailored for low-level vision tasks.

In summary, our work makes the following key contributions. (1) We present the first unified multimodal framework capable of handling four major low-level vision categories (over 100 sub-tasks) through both text and image interactions. (2) We introduce an effective multimodal fusion mechanism that aligns text and image prompts, mitigating task misalignment issues. (3) We provide new empirical insights into challenges of building multi-task low-level vision generalists, revealing how the integration of high-level generative and editing tasks can adversely impact fidelity-critical tasks.

2 Related Work

2.1 Image Restoration with Generative Prior

Diffusion-based methods have emerged as a robust framework for image restoration, converting degraded inputs into high-quality outputs through reverse denoising. Several key works illustrate the versatility of this approach [4, 9, 38, 55,

63, 67, 70, 72]. StableSR [55] leverages the generative priors of pre-trained text-to-image diffusion models for blind super-resolution, employing a time-aware encoder and feature wrapping to balance quality and fidelity while accommodating arbitrary resolutions. DiffBIR [38] uses a two-stage pipeline where the first stage reduces degradations and the second stage employs a latent diffusion model to generate missing details, proving effective in denoising and face restoration. PASD [67] extends the Stable Diffusion framework for realistic super-resolution and personalized stylization by integrating a pixel-aware mechanism that improves both resolution precision and style adaptability. SUPIR [70] scales up large diffusion models such as StableDiffusion-XL, incorporating a trained adapter and a massive high-resolution dataset to enable text-guided, photo-realistic restoration in complex scenes. However, the limitation of these approaches is that they are confined to image restoration tasks and cannot address other challenges in low-level vision.

2.2 All-in-one Generative Models

Developing all-in-one models is an exciting yet challenging pursuit. In the realm of image generation, various studies have sought to build versatile systems [27, 35, 37, 45, 58, 64]. For example, OmniGen [64] encodes text and images into a unified tensor, utilizing causal attention for text tokens and bidirectional attention for image tokens. Pixwizard [37] introduces task-specific embeddings for image editing and understanding, while ACE [27, 45] offers a conditioning module that accepts diverse input images and processes them concurrently with a transformer. UniReal [10] employs a video generation framework that treats images as individual frames, providing a universal solution for various image generation and editing tasks.

Despite these advances, most of these approaches focus on image generation and editing, leaving universal models for low-level vision relatively unexplored. Visual prompt-based approaches [12, 16, 40] tackle cross-domain tasks by utilizing pairs of image prompts. However, their dependence on meticulously crafted prompts renders them less intuitive and user-friendly compared to text-driven alternatives. Moreover, many current methods are restricted to fixed-resolution outputs, limiting their practical applicability.

3 Method

3.1 Building OmniLV Step-by-Step

In this section, we detail the key design choices and learned insights in developing a universal low-level vision model, outlining our step-by-step thinking process.

Preliminary: Base Model Selection Unlike most foundational image restoration models [22, 28, 47, 51] that are trained from scratch using deterministic regression objectives, we leverage a pre-trained text-to-image diffusion model as a strong initialization. Pre-trained diffusion models [22, 24, 34, 65, 80], trained on

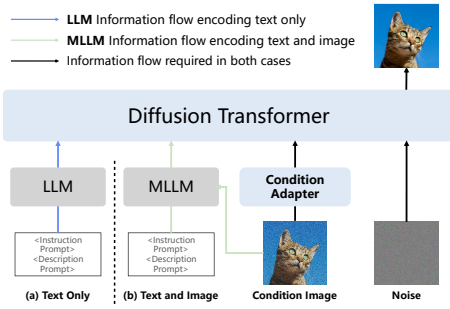

Fig. 2: Comparison between MLLM guided and LLM guided framework.

	Compression		Quantization		Noise		Inpainting	
	PSNR↑	MUSIQ↑	PSNR↑	MUSIQ↑	PSNR↑	MUSIQ↑	PSNR↑	MUSIQ↑
Addition	21.92	**56.31**	**18.72**	**55.71**	22.34	**59.31**	**19.94**	**56.96**
Concat	**21.93**	55.99	18.18	55.11	**22.35**	57.13	19.86	55.95

Table 1: Ablation study on whether to use addition or concatenation in in-context learning.

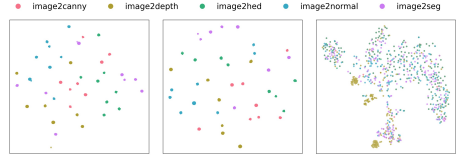

Fig. 3: t-SNE visualization of the feature space of LLM and MLLM. Each dot represents a task instruction.

billions of images, offer rich visual priors that enhance generalization, support diverse resolutions and aspect ratios, and effectively capture the uncertainty inherent in multi-task image restoration. These properties contribute to a more robust and versatile low-level vision model.

For our base model, we initialize with Lumina-Next [24,80], a flow-based diffusion transformer that introduces several architectural improvements over traditional DiT-based models [47], including 2D Rotary Positional Encoding, QK Normalization, and Sandwich Normalization. Additionally, Lumina-Next adopts a flow-matching formulation, improving training stability and accelerating convergence. To adapt this model for general low-level vision, we introduce a condition adapter that integrates inputs to enable effective task conditioning, which is illustrated in subsequent sections. The modified model is trained using a flow-matching loss to learn a conditional time-dependent velocity field, facilitating the transformation between noisy and clean image distributions. Please refer to the supplementary material for details of training loss. We use Gemma-2B to encode the instruction and description prompts, which matches the base model's text interface and provides a lightweight textual conditioning pathway for low-level tasks.

Encoding Multimodal Information Given an input image x, our goal is to generate the target image using both textual instructions and in-context visual exemplars. We explore two different encoding strategies: (1) **Separate encoding**, where text prompts are processed using a large language model (LLM), while visual exemplars are encoded independently. (2) **Unified encoding**, where both text and visual inputs are fused within a multimodal language model (MLLM). Fig. 2 illustrates the architectural differences between these two approaches. While unified encoding benefits from parameter efficiency and leverages cross-modal correlations, we observe that it introduces critical limita-

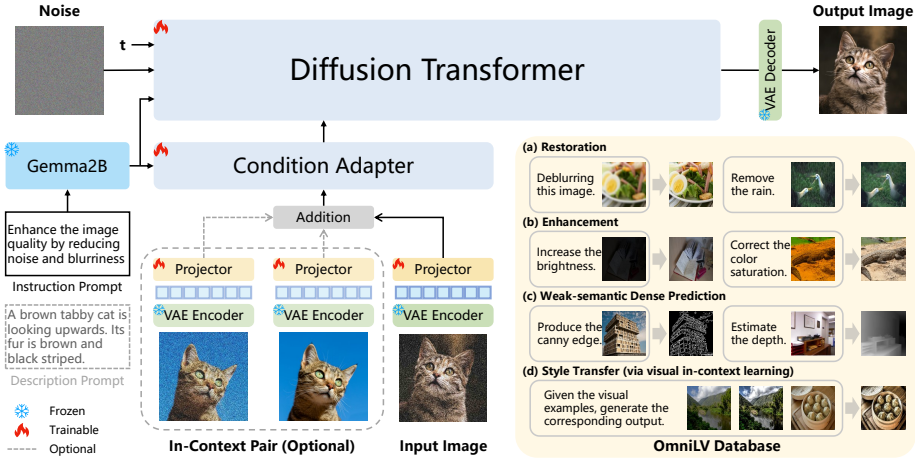

Fig. 4: Overall framework of OmniLV. First, input images are encoded into latent space by VAE encoder. Then, we patchify the image latent and noise latent into visual tokens. Optionally, in-context pairs can be added to visual tokens to handle complex scenarios. At the same time, the instruction prompt and description prompt are processed by Gemma-2B. Finally, we decode the denoised results to get the desired output images.

tions especially when applied to dense prediction tasks. Specifically, multimodal encoders often misinterpret task instructions, leading to inconsistencies in generated outputs. To better understand this issue, we visualize the encoded feature distributions in Fig. 3. Our findings indicate that mixing text and image prompts within a single encoder leads to severe task ambiguity. Since visual tokens dominate the shared feature space, text-based instructions often get overshadowed, leading to misalignment and incorrect outputs. Please refer to the supplementary material for an illustration of task misalignment.

Based on these observations, we adopt a separate encoding strategy: text instructions are processed via an LLM, while image exemplars are encoded using a visual VAE. This ensures clearer task separation, preventing interference between textual and visual guidance, and improves task accuracy across a vast number of low-level vision tasks.

Design Choices of Condition Integration Integrating condition images into diffusion models is commonly achieved through two primary approaches: (1) **Feature Injection:** A trainable adapter injects feature maps into a frozen diffusion model [46, 77]. (2) **Input Concatenation:** Condition images are concatenated with inputs, and the entire model is fine-tuned. These designs have been widely used in in-domain single task (e.g. image restoration, canny2image), achieving remarkable results [38, 70]. To systematically investigate condition integration strategies for general low-level vision tasks, we conduct comparative

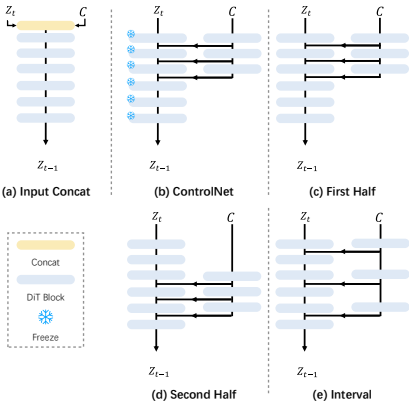

Fig. 5: Schematic structures of five different variants.

Table 2: Ablation study on condition integration.

	Position	Train DM?	SIDD		RealBlurJ		SR	
			PSNR↑	MUSIQ↑	PSNR↑	MUSIQ↑	PSNR↑	MUSIQ↑
(a)	Input	✓	32.40	21.91	22.98	51.66	22.89	**56.89**
(b)	First Half	✗	25.52	23.53	22.28	44.41	21.11	50.44
(c)	First Half	✓	**34.09**	**23.96**	**24.05**	**57.42**	**22.93**	56.72
(d)	Second Half	✓	29.60	23.38	23.15	54.35	22.96	56.60
(e)	Interval	✓	34.07	23.06	22.77	53.50	22.80	56.38

Table 3: Effects of Various Prompt Formats. Our approach supports text prompts, visual prompts, or a combination of both.

Text	Visual	Blur		Noise		Contrast Adju.		Saturate Adju.	
		PSNR↑	MUSIQ↑	PSNR↑	MUSIQ↑	PSNR↑	MUSIQ↑	PSNR↑	MUSIQ↑
✓	✗	**22.57**	68.95	**23.53**	69.23	**20.90**	69.95	**21.79**	70.90
✗	✓	21.99	67.59	22.71	67.66	20.09	66.58	20.14	68.08
✓	✓	22.50	**68.99**	23.07	**69.37**	20.50	**70.09**	21.00	**70.91**

Table 4: Ablation study for the training data.

High-semantic?	Blur		Noise		Contrast Adju.		Saturate Adju.	
	PSNR↑	MUSIQ↑	PSNR↑	MUSIQ↑	PSNR↑	MUSIQ↑	PSNR↑	MUSIQ↑
✗	**21.13**	**58.78**	**22.34**	57.39	**19.03**	**56.95**	**18.64**	**56.60**
✓	20.97	57.06	22.18	**59.31**	18.76	55.80	18.58	56.55

experiments evaluating different design choices (see Fig. 5). Our findings, summarized in Tab. 2, are as follows: (1) Training only the adapter is suboptimal (settings (b) & (c)), indicating that fine-tuning the base model is necessary for adapting generative priors to diverse low-level tasks. (2) While input concatenation is efficient(setting(a)), adding additional parameters to process the condition image enhances performance (setting (c)), suggesting that explicitly modeling condition images helps extract more relevant structural and contextual information. (3) The injection position significantly influences the performance (settings (c), (d), & (e)). Integrating condition information in the first half of the network leads to better results, likely because early-stage modulation ensures stronger feature guidance throughout the process.

Based on these findings, we propose a co-training condition adapter, which jointly optimizes the adapter and base model. Unlike ControlNet-like architectures, which keep the base model frozen, our approach ensures deeper feature alignment, improving multi-task generalization and fidelity.

Enabling In-Context Learning While text prompts can effectively guide tasks, many low-level vision tasks (e.g., stylization) require precise visual instructions that are difficult to express linguistically. To address this, we compare two paradigms for visual prompt integration: (1) **Input Concatenation** [10,60,64], where visual prompts are concatenated along the token dimension:

$$\mathbf{H}_{\text{fused}} = [\mathbf{H}_{\text{img}};\ \mathbf{H}_{\text{prompt}_1};\ \ldots;\ \mathbf{H}_{\text{prompt}_n}], \tag{1}$$

where $\mathbf{H}_{\text{img}}$ and $\mathbf{H}_{\text{prompt}_i}$ denote latent representation of input image and latent representation of i-th visual prompt, and $\mathbf{H}_{\text{fused}}$ denotes the combined latent

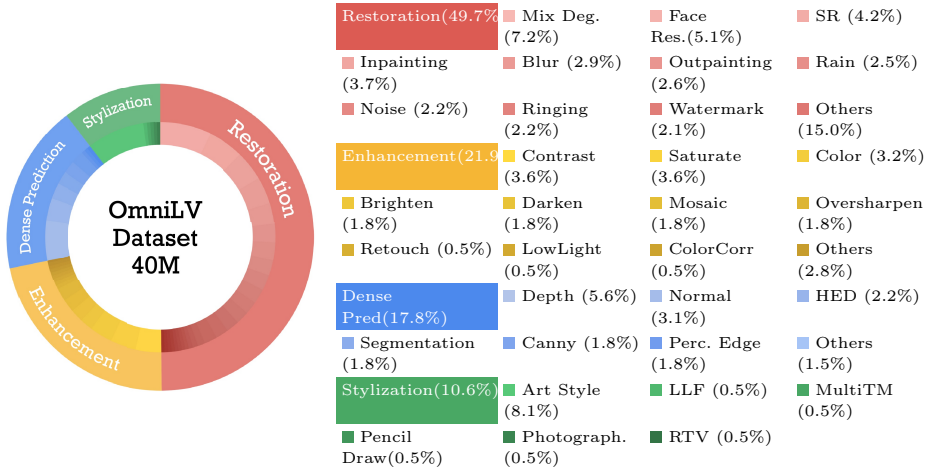

Fig. 6: OmniLV dataset distribution with main categories.

representation. (2) **Projection-Addition** [61], which employs lightweight projectors to align visual prompts with the latent space before summation:

$$\mathbf{H}_{\text{fused}} = \mathbf{H}_{\text{image}} + \sum_{i=1}^{n} \phi_i(\mathbf{H}_{\text{prompt}_i}), \tag{2}$$

where $\phi(\cdot)$ denotes linear projectors. Fig. 4 illustrates the architectural differences between these two approaches, where the concatenation method can be seen as a variation where the "Projector-Addition" module is replaced with a Concatenation operation. Tab. 1 presents the quantitative comparison, demonstrating that projection-addition outperforms input concatenation across different tasks. This suggests that projection-based alignment better preserves task-relevant information.

Final Architecture. Based on these insights, we design the final architecture of OmniLV, as illustrated in Fig. 4. Our approach unifies diverse low-level vision tasks while ensuring strong multimodal conditioning and in-context learning capabilities.

3.2 Large-Scale OmniLV Dataset

To build a universal low-level vision model, we construct a large-scale multi-task dataset containing 40 million instances over 100 sub-tasks across four major domains: image restoration, image enhancement, dense prediction, and stylization. The main categories and distribution of OmniLV dataset are illustrated in Fig. 6. The dataset is sourced from publicly available collections and synthetically generated pairs, with additional high-quality data created through internal pipelines.

Image Restoration. The restoration dataset covers 23 major tasks with a total of 45 sub-tasks, addressing various degradation types such as motion blur, noise, and weather-induced distortions. It consists of both real-world degraded images and synthetic degradation pairs, carefully processed alongside high-quality ground truth images to ensure realism and diversity.

Image Enhancement. The enhancement dataset includes 14 major tasks with a total of 25 sub-tasks, covering tasks such as low-light correction, contrast enhancement, and saturation refinement. The dataset is composed of professionally edited reference images alongside algorithmically generated enhancement pairs, ensuring controlled transformations that align with perceptual quality.

Weak-semantic Dense Prediction. For dense prediction tasks, we compile annotated datasets for 10 tasks, including edge detection, depth estimation, and surface normal prediction. Each sample contains pixel-level ground truth annotations paired with descriptive task-specific instructions, facilitating multimodal learning.

Image Stylization. The stylization dataset spans 20 tasks, covering artistic transformations across various styles and techniques. It includes both real-world artistic works and style-transferred images generated by neural algorithms, ensuring a diverse range of stylish effects. We implement in-context learning on image stylization tasks due to the difficulty of defining task prompt.

Dataset Summary and Test Set Construction. OmniLV dataset comprises four major task categories with over 100 sub-tasks and approximately 40 million training instances. For publicly available datasets, we directly adopt their test sets for evaluation. For our synthesized tasks, we construct test sets based on DIV2K-val [3], forming OmniLV-Test (OLV-T). OLV-T consists of 44 task-specific test sets, each containing 100 images, totaling 4,400 test images with 1K resolution. Further details on dataset partitioning and evaluation can be found in the supplementary material.

3.3 Model Training and Sampling Settings

The training of OmniLV is divided into three stages: In the first stage, we train the model with images at a resolution of 512^2, focusing solely on single-image tasks. We use a constant learning rate of 1e-4 and train for 100k steps with a batch size of 512. The second stage adds in-context learning (ICL) tasks. We continue training for another 100k steps, maintaining the same learning rate of 1e-4 and batch size of 512. Finally, in the third stage, we increase the resolution to 1024^2 and train on all tasks. The batch size is reduced to 128, and the learning rate remains at 1e-4. This final stage ensures that the model is trained to handle a variety of tasks and image sizes effectively. The model is trained using 16 A100 GPUs.

Table 5: Quantitative comparison on restoration tasks. Red and blue colors represent the best and second best performance, respectively, excluding specialized models. All values are reported as PSNR↑ / MUSIQ↑. For specialized models, if a model achieves the best value, the corresponding number is highlighted in **bold**.

Category	Method	Deblur		Compression	Denoise		Derain		Desnow	BIR	Face
		OLV-T(6 types)	RealBlur-J	OLV-T(2 types)	OLV-T(6 types)	SIDD	Synthetic	Rain1400	Snow100K-L	DIV2K	CelebA
Specialized	X-Restormer	21.18/45.17	26.57/50.41	–	**27.19**/63.67	31.95/22.04	27.10/71.34	32.35/70.34	–	–	–
	MPRNet	20.33/43.35	26.51/48.45	–	24.53/45.66	39.63/22.34	25.16/69.40	32.04/69.98	–	–	–
	MAXIM	21.39/44.13	**29.99**/55.68	–	24.75/49.30	**39.68**/22.39	25.82/71.15	32.25/70.27	–	–	–
	DiffBIR	–	–	–	–	–	–	–	–	**22.77**/67.01	–
	GFPGAN	–	–	–	–	–	–	–	–	–	**25.80**/69.76
	CodeFormer	–	–	–	–	–	–	–	–	–	25.15/**75.55**
All-in-One Restoration	X-Restormer	21.44/39.73	26.23/38.84	–	25.96/62.42	24.06/20.84	23.28/69.00	32.12/70.26	–	–	–
	DA-CLIP	19.94/34.98	18.82/39.22	–	22.99/44.89	26.40/29.25	23.15/53.18	26.44/67.78	–	–	–
	AutoDIR	20.09/45.07	19.10/49.63	–	26.46/57.80	22.19/28.72	25.33/64.59	26.21/70.75	–	–	–
Visual-Prompt-based	Painter	17.05/28.74	15.37/28.79	17.84/34.43	18.04/37.11	38.65/21.57	17.84/34.43	27.92/62.38	20.30/47.60	–	–
	PromptGIP	20.01/31.26	22.94/29.65	21.93/35.15	22.80/35.58	26.16/22.79	21.93/35.15	23.87/50.62	20.29/40.21	–	–
	GenLV	22.15/33.00	25.53/29.12	23.59/35.96	23.51/38.21	30.41/28.10	23.59/35.96	26.26/56.99	20.21/45.61	–	–
Text-Prompt-based	PromptFix	20.32/43.75	26.14/39.37	18.10/54.01	14.59/51.77	24.25/21.22	18.10/54.01	21.61/63.07	21.12/53.83	13.77/29.49	–
	Pixwizard	17.90/64.19	23.34/55.97	18.99/62.40	17.22/63.05	27.60/23.63	18.99/62.40	23.84/66.89	21.12/61.41	19.03/59.90	–
Multi-Modal Instruction	OmniLV	22.57/68.95	28.24/36.09	22.93/68.99	23.53/69.23	32.96/22.42	22.93/68.99	24.98/65.66	24.57/61.19	22.36/69.55	25.04/70.70

Table 6: Quantitative comparison on enhancement tasks.

Category	Method	Brighten	Darken	Low light	Photoretouching	Contrast Adjust	Saturation Adjust	Oversharpening
		OLV-T(4 types)	OLV-T(4 types)	LOLv2-Real	MIT5K	OLV-T(4 types)	OLV-T(4 types)	OLV-T(1 type)
Specialized	Retinexformer	–	16.72/65.73	22.79/59.30	16.12/63.24	–	–	–
	MIRNet	–	16.35/65.45	28.10/63.35	19.37/**65.59**	–	–	–
	MAXIM	–	16.09/67.16	**34.04/70.75**	14.98/62.90	–	–	–
All-in-One Restoration	DA-CLIP	–	14.91/53.04	26.64/67.73	–	–	–	–
	AutoDIR	–	15.48/64.74	24.16/67.91	–	–	–	–
Visual-Prompt-based	Painter	12.00/36.77	13.96/37.09	29.44/53.82	17.19/58.39	12.55/35.99	13.25/36.66	–
	PromptGIP	15.46/35.43	17.85/33.89	21.35/38.18	16.57/43.02	15.80/33.75	16.63/34.49	16.63/38.74
	GenLV	21.11/40.16	21.70/39.31	21.01/50.84	24.91/56.08	21.58/39.46	20.87/40.29	21.69/37.08
Text-Prompt-based	PromptFix	10.55/57.78	10.15/54.40	17.16/63.54	11.09/52.89	11.34/57.76	12.31/58.49	14.93/56.01
	PixWizard	11.16/64.14	13.81/65.44	14.07/62.11	15.99/63.59	13.12/65.51	12.77/65.13	13.55/71.00
Multi-Modal Instruction	OmniLV	22.58/70.54	20.28/69.77	18.60/58.76	19.78/62.12	20.91/69.95	21.80/70.91	23.64/70.87

4 Experiments

4.1 Comparisons with Existing Works

As a universal model, OmniLV exhibits superior abilities for various low-level vision tasks, even compared with existing task-specific models. We compare our method with task-specific methods (X-Restormer [11], MPRNet [74], MAXIM [54], DiffBIR [38], GFPGAN [59], CodeFormer [79], Retinexformer [8], MIRNet [73], Depth Anything (D.A.) [66]), all-in-one methods (X-Restormer [11], DA-CLIP [44], AutoDIR [31]), visual prompt methods (PromptGIP [40], GenLV [12], Painter [57]), and text-guided diffusion methods (PixWizard [37], PromptFix [71]). Some of them are constrained to generating images of fixed size. In our comparison, we resize the generated image to target image size to facilitate fair comparisons. We selected full-reference metric PSNR and no-reference metric MUSIQ [32] for quantitative comparison. The test sets include both our synthetic OLV-T and public datasets (RealBlur-J [50], SIDD [1], Rain1400 [23], Snow100K-L [41], DIV2K [3], CelebA [42], LOL-v2-Real [68], MIT5K [7]). Detailed results are provided in the supplementary material.

Table 7: Quantitative comparison on stylization tasks and weak-semantic dense prediction tasks.

(a) Stylization Tasks.

Category	Method	LLF		PencilDrawing	
		PSNR↑	FID↓	PSNR↑	FID↓
Visual-Prompt-based	Painter	13.64	120.3	8.434	157.5
	PromptGIP	22.87	50.61	21.35	132.5
	GenLV	25.66	29.53	28.29	38.70
Multi-Modal Instruction	OmniLV	21.72	24.16	20.33	54.18

(b) Dense Prediction Tasks.

Method	Depth Esti.	Normal Esti.	HED	DA-2K	SUNRGB-D	PASCAL-Ctx
	RMSE↓	Mean Angle Error↓	MAE↓	Accuracy↑	RMSE↓	Score↓
PromptGIP	–	–	85.90	–	–	–
GenLV	–	–	44.28	–	–	–
D.A.	0.291	–	–	–	–	–
InvPT	–	19.04	–	–	–	–
PixWizard	0.941	19.65	25.95	85.49	0.3523	34.65
OmniLV	0.525	17.30	20.75	90.86	0.3018	29.91

Image Restoration. Tab. 5 demonstrates that OmniLV achieves the highest PSNR scores across all restoration benchmarks when compared with diffusion-based models. We demonstrate the qualitative comparisons in Fig. 7 on general low-level tasks. In addition, the MUSIQ scores of OmniLV are highly competitive on most benchmarks, further underscoring its strong performance. Notably, on the Blind Image Restoration (BIR) and Face benchmarks, OmniLV, as a universal model, attains performance levels that are comparable to those of state-of-the-art specialized models, thereby validating its effectiveness in handling diverse restoration tasks.

Image Enhancement. As reported in Tab. 6, OmniLV significantly improves upon existing enhancement methods. The improvements highlight OmniLV's ability to effectively enhance image quality while maintaining natural details and color fidelity.

Dense Prediction. Tab. 7b reports depth estimation (RMSE), normal estimation (mean angle error), edge detection (HED MAE), and additional dense benchmarks (DA-2K, SUNRGB-D, and PASCAL-Context). OmniLV improves over generalist/prompt-based baselines, while specialized models still lead on some single-task metrics (e.g., Depth Anything on depth with RMSE 0.291). This indicates both the feasibility of unified dense prediction and remaining headroom versus task-specific systems; evaluation details are provided in the supplementary material.

Stylization. Tab. 7a illustrates that OmniLV also performs well on stylization tasks such as Local Laplacian Filter (LLF) and Pencil Drawing [43]. Since stylization tasks are challenging to describe using natural language, we employed visual prompts to guide the model in processing images. OmniLV obtains balanced results in terms of objective quality and perceptual quality, thus validating its versatility across diverse low-level vision tasks.

4.2 More Exploration

Text Prompt vs. Visual Prompt. Our model supports both text prompt and visual prompt to guide the generation process. In Tab. 3, we present a detailed comparison between the two prompting methods across several low-level

Fig. 7: Comparison results for low-level vision tasks. More results can be found in the Supp.

vision tasks, including deblurring, denoising, contrast adjustment, and saturation adjustment. Adopting both prompts can yield better quality scores.

Relationship with High-Semantic Tasks. We further investigate the relationship between low-level vision tasks and high-semantic tasks such as image generation or image editing tasks [6, 30, 52, 53, 69, 75, 78]. As shown in Tab. 4, when high-semantic tasks are included in the training data, the performance on low-level vision tasks degrades. Specifically, performance for various tasks is consistently lower when high-semantic tasks are incorporated. This degradation arises because high-semantic tasks prioritize conceptual coherence and structural abstraction over pixel-accurate reconstruction, which conflicts with the objectives of low-level vision tasks that demand fine-grained texture recovery and precise detail preservation.

Generalization Exploration. We investigated the generalizability of OmniLV in terms of domain-specific adaptation and real-world robustness. Specifically, we selected images with various real-world degradations to comprehensively assess the model's performance. As shown in Fig. 8, OmniLV effectively restores images in these diverse conditions, demonstrating its robustness and versatility in handling complex real-world degradations, such as real-world restoration, deraining, desnowing, underwater im-

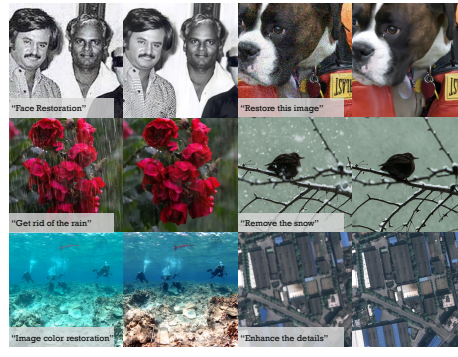

Fig. 8: Examples of image restoration in various scenarios.

age enhancement, and satellite image enhancement.

Prompt Conflict. To study how OmniLV behaves under inconsistent conditioning, we construct deliberately conflicting cases where the text instruction and the visual exemplars specify different tasks. As illustrated in Fig. 9, when both prompts are provided, the output consistently follows the text instruction rather than the visual exemplars. We do not intentionally create such conflicting pairs in our training data. We believe this behavior arises because the data is dominated by samples with text instructions. Consequently, the model has learned to treat the text instruction as the primary source of task specification in the presence of conflicts.

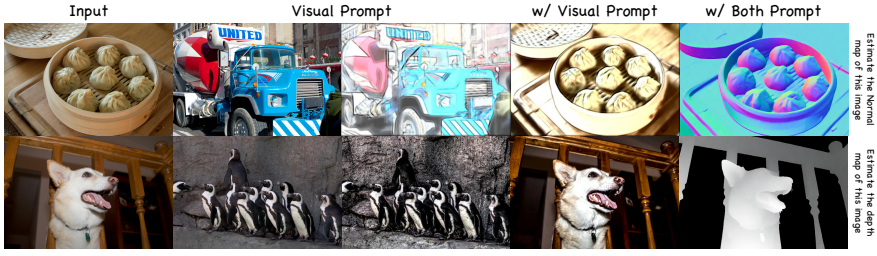

Fig. 9: Results when providing conflict visual prompt and text instruction. Our model tends to follow text instruction when conflict happens.

Cross-domain Task. Following the definition in GenLV [12], we categorize low-level vision tasks into five domains: restoration, enhancement, stylization, dense prediction, and natural images. The cross-domain capability in our setting refers to a model's ability to not only accomplish tasks within a single domain but also handle tasks from multiple domains simultaneously. To assess this ability, we construct composite tasks that involve multiple domains at once. As illustrated in Fig. 10, OmniLV demonstrates strong instruction-following behavior and consistently produces correct results across these mixed-domain scenarios.

Instruction Following Ability. Existing models often fail to adhere to the instructions and perform wrong transformations, especially for dense prediction tasks. In contrast, OmniLV shows strong instruction-following ability courtesy of the dedicated design choice described in Sec. 3.1. To illustrate it more clearly, we compute the success rates of OmniLV on some dense prediction tasks compared to OmniGen [64] and PixWizard [37]. We deem a case as successful when the model performs the intended task, no matter how well. Tab. 8 presents the results, showing the superior task alignment ability of OmniLV.

Human Preference. We conduct pairwise human preference comparisons for evaluating low-level image restoration and enhancement quality. As summarized

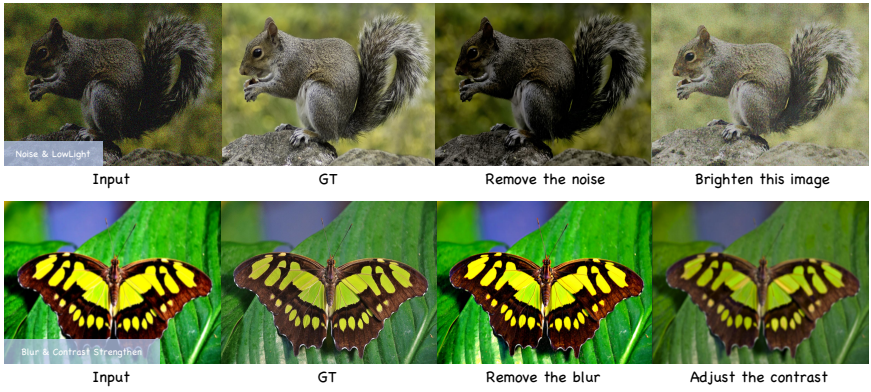

Fig. 10: Results for cross-domain tasks. For cross-domain tasks, OmniLV can generate correct results with regard to the prompts.

Table 8: Successful rates on some dense prediction tasks.

Models	Image2Canny	Image2Depth	Image2HED
OmniGen	80%	75%	-
Pixwizard	100%	93%	97%
OmniLV	100%	100%	100%

Table 9: Human-preference Elo ratings on benchmarks. PromptFix is anchored at 1000 Elo.

Model	**Elo rating**
OmniLV	1,475
PixWizard	1,310
PromptFix	1,000

in Tab. 9, OmniLV achieves the highest human-preference Elo score, outperforming both PixWizard and PromptFix by substantial margins.

Discussion. These results show that careful conditioning is critical for unified low-level vision generalists. OmniLV remains behind specialized systems on some dense prediction and stylization settings (see supplementary `tab:inference_cost` and continued fine-tuning results).

5 Conclusion

In this work, we introduced **OmniLV**, a unified multimodal framework for low-level vision that successfully handles over 100 sub-tasks, including image restoration, enhancement, weak-semantic dense prediction, and stylization. By leveraging both textual and visual prompts with generative priors, OmniLV demonstrates robust generalization, high-fidelity results, and flexibility across arbitrary resolutions. OmniLV achieves state-of-the-art performance in multiple low-level vision tasks and demonstrates promising generalization capabilities in real-world scenarios.

Limitations. Despite OmniLV's extensive capability on a wide range of low-level vision tasks, it does not always achieve optimal performance in certain spe-

cialized scenarios. Future work will focus on enhancing task-specific performance by refining model components and training strategies.

References

1. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1692–1700 (2018) 10
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 1
3. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017) 9, 10
4. Ai, Y., Zhou, X., Huang, H., Han, X., Chen, Z., You, Q., Yang, H.: Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=6eoGVqMiIj 3
5. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022) 1
6. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800 (2022) 12
7. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input / output image pairs. In: The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition (2011) 10
8. Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12504–12513 (2023) 1, 10
9. Chen, B., Li, G., Wu, R., Zhang, X., Chen, J., Zhang, J., Zhang, L.: Adversarial diffusion compression for real-world image super-resolution. arXiv preprint arXiv:2411.13383 (2024) 3
10. Chen, X., Zhang, Z., Zhang, H., Zhou, Y., Kim, S.Y., Liu, Q., Li, Y., Zhang, J., Zhao, N., Wang, Y., et al.: Unireal: Universal image generation and editing via learning real-world dynamics. arXiv preprint arXiv:2412.07774 (2024) 4, 7
11. Chen, X., Li, Z., Pu, Y., Liu, Y., Zhou, J., Qiao, Y., Dong, C.: A comparative study of image restoration networks for general backbone network design. In: European Conference on Computer Vision. pp. 74–91. Springer (2024) 1, 10
12. Chen, X., Liu, Y., Pu, Y., Zhang, W., Zhou, J., Qiao, Y., Dong, C.: Learning a low-level vision generalist via visual task prompt. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2671–2680 (2024) 2, 4, 10, 13
13. Chen, X., Liu, Y., Zhang, Z., Qiao, Y., Dong, C.: Hdrunet: Single image hdr reconstruction with denoising and dequantization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 354–363 (2021) 1
14. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22367–22377 (2023) 1, 2
15. Chen, X., Zhang, Z., Ren, J.S., Tian, L., Qiao, Y., Dong, C.: A new journey from sdrtv to hdrtv. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4500–4509 (2021) 1

16. Chen, X., Zhu, K., Pu, Y., Cao, S., Li, X., Zhang, W., Liu, Y., Qiao, Y., Zhou, J., Dong, C.: Exploring scalable unified modeling for general low-level vision. arXiv preprint arXiv:2507.14801 (2025) 2, 4
17. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024) 1
18. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences **67**(12), 220101 (2024) 1
19. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024) 1
20. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) 1
21. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015) 1
22. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024) 3, 4
23. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3855–3863 (2017) 10
24. Gao, P., Zhuo, L., Liu, C., , Du, R., Luo, X., Qiu, L., Zhang, Y., et al.: Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. arXiv preprint arXiv:2405.05945 (2024) 4, 5
25. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016) 1
26. Guo, Y., Gao, Y., Lu, Y., Liu, R.W., He, S.: Onerestore: A universal restoration framework for composite degradation. In: European Conference on Computer Vision (2024) 2
27. Han, Z., Jiang, Z., Pan, Y., Zhang, J., Mao, C., Xie, C., Liu, Y., Zhou, J.: Ace: All-round creator and editor following instructions via diffusion transformer. arXiv preprint arXiv:2410.00086 (2024) 4
28. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3, 4
29. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017) 1
30. Hui, M., Yang, S., Zhao, B., Shi, Y., Wang, H., Wang, P., Zhou, Y., Xie, C.: Hq-edit: A high-quality dataset for instruction-based image editing. arXiv preprint arXiv:2404.09990 (2024) 12
31. Jiang, Y., Zhang, Z., Xue, T., Gu, J.: Autodir: Automatic all-in-one image restoration with latent diffusion. arXiv preprint arXiv:2310.10123 (2023) 10

32. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021) 10
33. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023) 1
34. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024) 4
35. Le, D.H., Pham, T., Lee, S., Clark, C., Kembhavi, A., Mandt, S., Krishna, R., Lu, J.: One diffusion to generate them all (2024), https://arxiv.org/abs/2411.16318 4
36. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17452–17462 (2022) 2
37. Lin, W., Wei, X., Zhang, R., Zhuo, L., Zhao, S., Huang, S., Xie, J., Qiao, Y., Gao, P., Li, H.: Pixwizard: Versatile image-to-image visual assistant with open-language instructions. arXiv preprint arXiv:2409.15278 (2024) 4, 10, 13
38. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: Diffbir: Toward blind image restoration with generative diffusion prior. In: European Conference on Computer Vision. pp. 430–448. Springer (2024) 1, 3, 4, 6, 10
39. Lin, X., Yu, F., Hu, J., You, Z., Shi, W., Ren, J.S., Gu, J., Dong, C.: Harnessing diffusion-yielded score priors for image restoration. arXiv preprint arXiv:2507.20590 (2025) 1
40. Liu, Y., Chen, X., Ma, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Unifying image processing as visual prompting question answering. arXiv preprint arXiv:2310.10513 (2023) 2, 4, 10
41. Liu, Y.F., Jaw, D.W., Huang, S.C., Hwang, J.N.: Desnownet: Context-aware deep network for snow removal. IEEE Transactions on Image Processing **27**(6), 3064–3073 (2018) 10
42. Liu, Z., Luo, P., Wang, X., Tang, X.: Large-scale celebfaces attributes (celeba) dataset. Retrieved August **15**(2018), 11 (2018) 10
43. Lu, C., Xu, L., Jia, J.: Combining sketch and tone for pencil drawing production. In: Proceedings of the symposium on non-photorealistic animation and rendering. pp. 65–73 (2012) 11
44. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Controlling vision-language models for universal image restoration. arXiv preprint arXiv:2310.01018 (2023) 10
45. Mao, C., Zhang, J., Pan, Y., Jiang, Z., Han, Z., Liu, Y., Zhou, J.: Ace++: Instruction-based image creation and editing via context-aware content filling. arXiv preprint arXiv:2501.02487 (2025) 4
46. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4296–4304 (2024) 6
47. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 3, 4, 5
48. Potlapalli, V., Zamir, S.W., Khan, S., Khan, F.: Promptir: Prompting for all-in-one image restoration. In: Thirty-seventh Conference on Neural Information Processing Systems (2023) 2

49. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024) 1
50. Rim, J., Lee, H., Won, J., Cho, S.: Real-world blur dataset for learning and benchmarking deblurring algorithms. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 10
51. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 3, 4
52. Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu edit: Precise image editing via recognition and generation tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8871–8879 (2024) 12
53. Shi, J., Xu, N., Bui, T., Dernoncourt, F., Wen, Z., Xu, C.: A benchmark and baseline for language-driven image editing. In: Proceedings of the Asian Conference on Computer Vision (2020) 12
54. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxim: Multi-axis mlp for image processing. CVPR (2022) 10
55. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision **132**(12), 5929–5949 (2024) 3, 4
56. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6849–6857 (2019) 1
57. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. arXiv preprint arXiv:2212.02499 (2022) 10
58. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024) 4
59. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 10
60. Wang, Z., Xia, X., Chen, R., Yu, D., Wang, C., Gong, M., Liu, T.: Lavin-dit: Large vision diffusion transformer. arXiv preprint arXiv:2411.11505 (2024) 7
61. Wang, Z., Jiang, Y., Lu, Y., He, P., Chen, W., Wang, Z., Zhou, M., et al.: In-context learning unlocked for diffusion models. Advances in Neural Information Processing Systems **36**, 8542–8562 (2023) 8
62. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025) 1
63. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. Advances in Neural Information Processing Systems **37**, 92529–92553 (2024) 3
64. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. arXiv preprint arXiv:2409.11340 (2024) 1, 4, 7, 13

65. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., Han, S.: Sana: Efficient high-resolution image synthesis with linear diffusion transformer (2024), https://arxiv.org/abs/2410.10629 4
66. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10371–10381 (2024) 1, 10
67. Yang, T., Wu, R., Ren, P., Xie, X., Zhang, L.: Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In: European Conference on Computer Vision. pp. 74–91. Springer (2024) 3, 4
68. Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J.: From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 10
69. Yildirim, A.B., Baday, V., Erdem, E., Erdem, A., Dundar, A.: Inst-inpaint: Instructing to remove objects with diffusion models (2023) 12
70. Yu, F., Gu, J., Li, Z., Hu, J., Kong, X., Wang, X., He, J., Qiao, Y., Dong, C.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25669–25680 (2024) 1, 3, 4, 6
71. Yu, Y., Zeng, Z., Hua, H., Fu, J., Luo, J.: Promptfix: You prompt and we fix the photo. In: NeurIPS (2024) 10
72. Yue, Z., Liao, K., Loy, C.C.: Arbitrary-steps image super-resolution via diffusion inversion. arXiv preprint arXiv:2412.09013 (2024) 3
73. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Learning enriched features for real image restoration and enhancement. In: ECCV (2020) 1, 10
74. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: CVPR (2021) 10
75. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. In: Advances in Neural Information Processing Systems (2023) 12
76. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE transactions on image processing **26**(7), 3142–3155 (2017) 1
77. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 6
78. Zhao, H., Ma, X., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale (2024), https://arxiv.org/abs/2407.05282 12
79. Zhou, S., Chan, K.C., Li, C., Loy, C.C.: Towards robust blind face restoration with codebook lookup transformer. In: NeurIPS (2022) 10
80. Zhuo, L., Du, R., Han, X., Li, Y., Liu, D., Huang, R., Liu, W., et al.: Lumina-next: Making lumina-t2x stronger and faster with next-dit. arXiv preprint arXiv:2406.18583 (2024) 4, 5