L2P: Unlocking Latent Potential for Pixel Generation

Zhennan Chen[1,2*] **Junwei Zhu**[2†] **Xu Chen**[2] **Jiangning Zhang**[2] **Jiawei Chen**[1]
Zhuoqi Zeng[3] **Wei Zhang**[4] **Chengjie Wang**[2] **Jian Yang**[1] **Ying Tai**[1‡]
[1]Nanjing University [2]Tencent Youtu Lab [3]Hainan-biuh [4]Weess Gmbh
https://nju-pcalab.github.io/projects/L2P/

Abstract

Pixel diffusion models have recently regained attention for visual generation. However, training advanced pixel-space models from scratch demands prohibitive computational and data resources. To address this, we propose the **Latent-to-Pixel (L2P)** transfer paradigm, an efficient framework that directly harnesses the rich knowledge of pre-trained LDMs to build powerful pixel-space models. Specifically, L2P discards the VAE in favor of large-patch tokenization and freezes the source LDM's intermediate layers, exclusively training shallow layers to learn the latent-to-pixel transformation. By utilizing LDM-generated synthetic images as the sole training corpus, L2P fits an already smooth data manifold, enabling rapid convergence with zero real-data collection. This strategy allows L2P to seamlessly migrate massive latent priors to the pixel space using only 8 GPUs. Furthermore, eliminating the VAE memory bottleneck unlocks native 4K ultra-high resolution generation. Extensive experiments across mainstream LDM architectures show that L2P incurs negligible training overhead, yet performs on par with the source LDM on DPG-Bench and reaches 93% performance on GenEval.

1 Introduction

Latent Diffusion Models (LDMs) [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11] have recently dominated the field of text-to-image (T2I) generation [12, 13, 14, 15, 16, 17, 18, 19], achieving unprecedented success in synthesizing high-quality images. By compressing images into a lower-dimensional latent space via a Variational Autoencoder (VAE) [20], LDMs significantly reduce computational overhead. Nevertheless, this bipartite paradigm is inherently bounded by VAE-induced limitations. The compression process inevitably discards critical high-frequency details [21, 22, 23, 24, 25], leading to sub-optimal reconstruction and a non-end-to-end training pipeline that decouples representation learning from the generation process. Furthermore, the VAE decoding process imposes severe memory constraints, bottlenecking the scaling to ultra-high resolutions (e.g., native 4K). To circumvent these VAE-induced limitations and achieve uncompromised visual fidelity, pixel-space diffusion models have recently re-emerged as a promising alternative [26, 27, 28, 29, 30, 31].

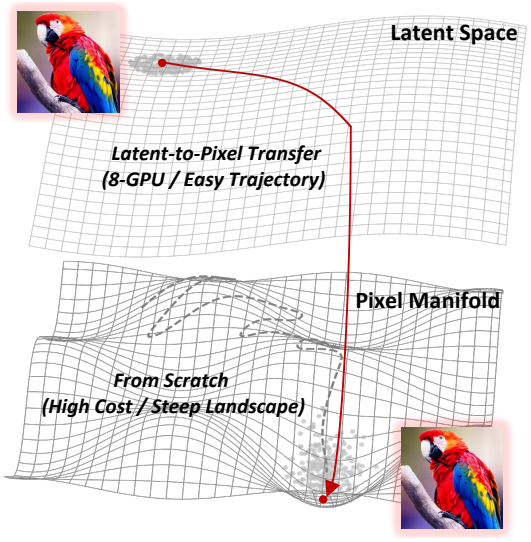

Figure 1: By leveraging the smooth manifold of pre-trained LDMs, L2P bypasses costly from-scratch training, achieving high-quality generation with just 8 GPUs.

*Work done during the internship at Tencent.
†Project Leader. ‡Corresponding Author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Despite their architectural purity and end-to-end appeal, training a state-of-the-art pixel-space T2I model from scratch remains computationally prohibitive, typically demanding hundreds of high-end GPUs and billions of curated image-text pairs. Consequently, nascent pixel-space models [28, 29, 30, 31] frequently exhibit a pronounced gap in semantic comprehension and compositional quality when compared to established LDMs [12, 13, 32, 33], which have already internalized profound world knowledge distilled from massive-scale datasets. This presents a critical cold-start dilemma: *Can we directly transfer the rich semantic priors embedded in pre-trained LDMs to a pixel-space diffusion model, thereby bypassing the astronomical costs of from-scratch training?*

To this end, we propose the **Latent-to-Pixel (L2P)** transfer paradigm, a highly efficient framework designed to bridge the representation gap between latent and pixel spaces at low cost, as shown in Figure 1. Architecturally, we discard the VAE, employ large-patch tokenization for pixel inputs, and utilize a lightweight U-Net to manage the decoding process. To facilitate robust knowledge transfer, we keep the Diffusion Transformer (DiT) architecture unmodified and align the prediction target with the source LDM. This architectural fidelity ensures seamless weight inheritance, while objective consistency allows the frozen intermediate layers to function within their native optimization manifold, thereby preserving the rich semantic priors and world knowledge. Consequently, we freeze the intermediate layers of the DiT backbone and exclusively train the shallow input and output layers to learn the latent-to-pixel modality transformation. Furthermore, rather than collecting massive real-world datasets, we utilize the source LDM to generate high-quality images as our training corpus. Beyond eliminating data curation costs, this strategy forces the new pixel model to fit the smooth data manifold already constructed by the LDM, thereby drastically accelerating convergence. Moreover, eliminating the VAE bottleneck unlocks native 4K generation. We maintain computational efficiency at this scale simply by enlarging the patch size and increasing the noise shift. The resulting heavier noise fully corrupts the dense local correlations of 4K pixels, averting trivial local reconstruction and enforcing global structural learning.

Our contributions are summarized as follows:

• We propose Latent-to-Pixel (L2P), a highly resource-efficient transfer paradigm that harnesses massive pre-trained LDM priors for pixel-space diffusion using merely 8 GPUs, seamlessly transitioning to the pixel space while simultaneously unlocking native 4K ultra-high-resolution generation.

• We construct a comprehensive, multi-dimensional prompt dataset to generate synthetic training pairs, achieving highly efficient training with zero real-data cost.

• Extensive validations demonstrate that L2P robustly inherits the generative priors of the source LDM. It maintains near-lossless semantic alignment on standard benchmarks while simultaneously exhibiting exceptional visual fidelity in native 4K ultra-high-resolution generation.

2 Related Work

Text-to-Image Generation. Text-to-Image (T2I) generation [34, 35, 36, 14, 37, 15, 38, 16, 17, 39, 18, 40, 41, 19, 42, 43, 44, 45, 46, 47, 48] is currently dominated by LDMs [49], which bypass the exorbitant computational costs of early pixel-space models [50, 2] by compressing images into a compact latent space via a Variational Autoencoder (VAE) [20]. Despite encapsulating profound world knowledge and robust semantic alignment, LDMs are inherently bottlenecked by the VAE decoder. The compression-decompression process inevitably incurs high-frequency information loss [22, 23, 24, 25]. Furthermore, the severe quadratic memory footprint of the VAE spatial decoding process imposes rigid hardware constraints, making native ultra-high resolution (e.g., 4K) generation practically intractable for standard LDMs [37, 51, 52, 8, 53, 54, 55, 56].

Pixel Diffusion Models. Early pixel diffusion models (e.g., DDPM [2] and ADM [50]) are severely constrained when processing high-resolution images due to their quadratic complexity bottleneck. Approaches like JiT [27] and PixelGen [30] introduce novel prediction targets. Most relevantly, PixNerd [29], DeCo [28], PixelDiT [31], and DiP [26] efficiently decouple global structural modeling from local detail refinement via lightweight decoders. Despite their architectural advances, these modern models still mandate computationally prohibitive from-scratch training on massive datasets. In contrast, our work fundamentally circumvents these exorbitant pre-training costs. Through our L2P paradigm, we directly transfer the rich priors of existing LDMs into the pixel space, achieving state-of-the-art pixel-based text-to-image generation with minimal computational overhead.

3 Method

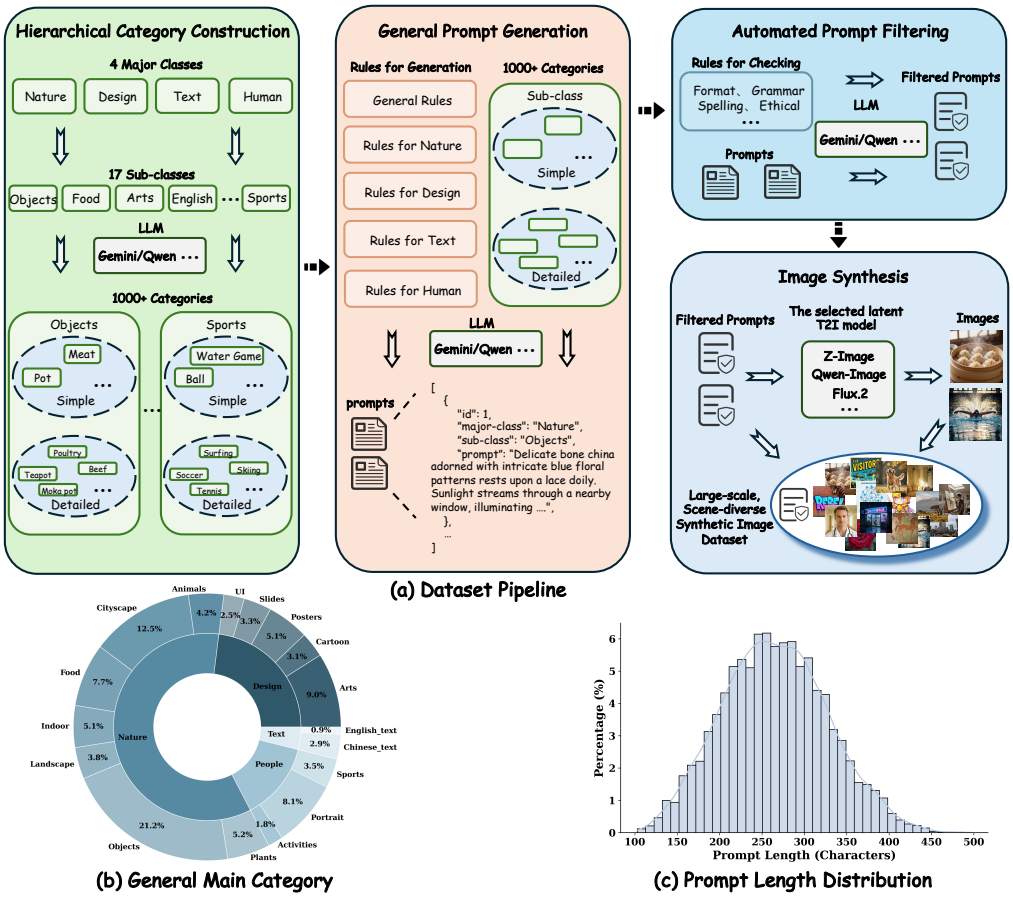

Figure 2: The proposed data construction pipeline. (a) Four-stage construction framework: Hierarchical Category Construction, General Prompt Generation, Automated Prompt Filtering, and Image Synthesis. Further details are provided in the Appendix. (b) Category distribution detailing 4 super-classes and 17 sub-classes. (c) Prompt length distribution, peaking at 200–350 characters to provide rich textual details.

3.1 Preliminary

Diffusion models learn to synthesize data by reversing a progressive noise-injection process. Given an initial sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, the discrete forward process yields a noisy state at step t:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where $\bar{\alpha}_t$ is determined by a predefined variance schedule. As $t \to T$, the marginal distribution $p(\mathbf{x}_T)$ converges to a standard Gaussian $\mathcal{N}(0, \mathbf{I})$. In a continuous-time framework, this corruption process is governed by a stochastic differential equation (SDE) $d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)d\mathbf{w}$, with drift $f(\cdot, t)$ and diffusion coefficient $g(t)$. The generative process corresponds to simulating the reverse-time Probability Flow ODE:

$$d\mathbf{x} = \left[f(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right] dt. \tag{2}$$

Consequently, data generation relies on estimating the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ or the associated vector field. A standard approach (e.g., DDPM) trains a neural network $\epsilon_\theta(\mathbf{x}_t, t)$ to predict the injected noise:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2 \right]. \tag{3}$$

Alternatively, Flow Matching (FM) [32] offers a simulation-free paradigm to directly regress the continuous vector field. By defining a conditional probability path $p_t(\mathbf{x} \mid \mathbf{x}_0)$ and its target vector field $u_t(\mathbf{x})$, a model $v_\theta(\mathbf{x}, t)$ is optimized via:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, p_t(\mathbf{x}|\mathbf{x}_0)} \left[\| u_t(\mathbf{x}) - v_\theta(\mathbf{x}, t) \|^2 \right]. \tag{4}$$

3.2 Dataset Construction

To facilitate the L2P transfer without the prohibitive costs of real-world data collection, we designed a comprehensive dataset pipeline, as shown in Figure 2(a). Through this pipeline, we construct a large-scale, scene-diverse synthetic image dataset. Generating our training corpus directly from the source LDM forces the new pixel-space model to fit the smooth data manifold already constructed by the source model, significantly accelerating convergence and activating its intrinsic prior knowledge. Our data construction process is structured into the following sequential stages:

Hierarchical Category Construction. To ensure comprehensive semantic coverage and diversity, we establish a top-down hierarchical taxonomy. First, drawing upon [13, 57], we define 4 major classes and further divide them into 17 sub-classes, as shown in Figure 2(b). Subsequently, we leverage an LLM to expand these sub-classes into over 1,000 fine-grained categories.

General Prompt Generation. We design a refined set of generation rules to guide the LLM in synthesizing high-quality prompts. Guided by these customized rules and the 1,000+ categories, the LLM generates highly descriptive prompts formatted as structured JSON data. As shown in Figure 2(c), the generated prompts are densely concentrated between 200 and 350 characters, providing abundant textual details for complex scene generation.

Automated Prompt Filtering. To prevent the propagation of low-quality or unsafe data, we implement a rigorous prompt check. The rules for check filter the generated text based on strict criteria. This ensures a high-quality corpus of filtered prompts.

Image Synthesis. Finally, for image generation, we feed the filtered prompts into the source latent T2I model to synthesize the final images.

3.3 L2P Transfer Paradigm

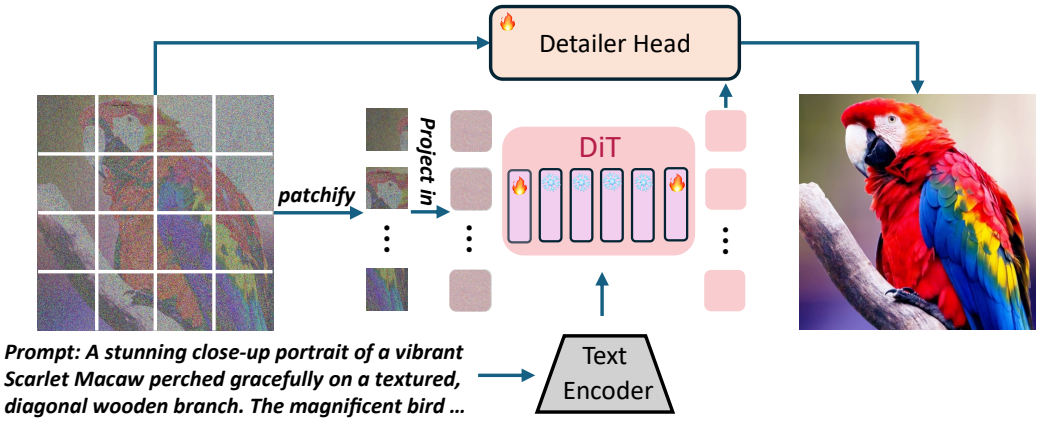

Figure 3: Overview of the L2P framework. L2P operates directly in pixel space via large-patch tokenization without VAE. To efficiently adapt priors, core DiT layers are frozen, while shallow blocks and a Detailer Head are tuned to restore high-frequency spatial details.

To efficiently migrate the rich generative priors embedded in pre-trained LDMs into the pixel space, we introduce the L2P transfer paradigm. The overall architecture is illustrated in Figure 3.

Architectural Adaptation. To facilitate the transition from latent to pixel space without disrupting the internal sequence processing of the pre-trained Diffusion Transformer (DiT), we implement three structural modifications:

1) We discard the VAE and apply a patchification strategy to the input image. To align the sequence length and maintain the computational efficiency equivalent to the original VAE-compressed latent space, we employ a patch size of 16×16.

2) Pre-trained LDMs map latent representations back to images via a VAE decoder. To bypass the VAE decoder bottleneck and enable high-fidelity pixel-level generation, inspired by DiP [26], we replace the final projection layer with a lightweight U-Net, termed the Detailer Head. This module decodes DiT representations to reconstruct dense pixel semantics and restore high-frequency details.

3) To achieve rapid convergence while preventing catastrophic forgetting of the LDM's semantic priors, we employ a selective freezing strategy. During training, the majority of the intermediate DiT blocks are frozen. We only update the initial input projection layer, the first and last n blocks of the DiT, and the newly added Detailer Head. This drastically reduces the computational overhead compared to training from scratch.

Objective Function. To maximize the preservation of pre-trained generative priors, we strictly adhere to the original diffusion training objective of the source LDM. The L2P optimization objective is formulated as:

$$\mathcal{L}_{\text{L2P}} = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\left\|(\epsilon - \mathbf{x}_0) - v_\theta(\mathbf{x}_t, t)\right\|^2\right] \tag{5}$$

By maintaining optimization consistency with the source model, L2P inherently mitigates the catastrophic forgetting of pre-trained knowledge. Furthermore, this architecture-agnostic formulation ensures seamless deployment across diverse LDM frameworks.

3.4 Scaling to Ultra-High Resolution

By bypassing the memory bottlenecks inherent to VAEs, our pure pixel architecture natively supports ultra-high resolution synthesis. When extended to 4K generation, L2P operates with remarkable efficiency, reducing single-step inference latency by 97.67% and peak GPU memory footprint by 38.81% compared to the source latent baseline, as shown in Figure 4. We enable this via two adaptations:

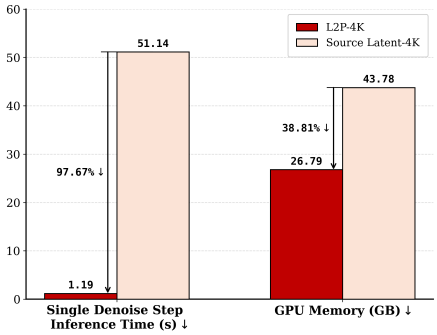

Figure 4: Efficiency comparison for 4K generation. L2P drastically mitigates the computational bottlenecks of high-resolution synthesis, significantly outperforming the source latent model in both inference speed and GPU memory consumption.

First, to maintain computational feasibility and a manageable sequence length for the DiT backbone, we dynamically expand the patch size from 16×16 to 64×64 for 4K inputs. This preserves inference speed without requiring structural modifications.

Second, due to the extremely dense local correlations in 4K pixel space, standard noise schedules fail to fully corrupt the image signal [58, 59]. This inadequate signal destruction causes the model to degenerate into trivial local reconstruction. To mitigate this, we increase the noise shift parameter, skewing the schedule toward higher noise levels. This guarantees sufficient data corruption during the forward process, forcing the model to learn robust global generation.

4 Experiments

4.1 Setup

Implementation Details. To validate the proposed L2P transfer paradigm, we instantiate our framework using Z-Image [12] as the source LDM. For the base transfer training at 1024×1024 resolution, we curate 10k diverse prompts and generate 20k synthetic images from the source model using varying random seeds. We utilize the UltraHR-100K dataset [37] for 4K training, since the source LDM fails to generate reliable 4K synthetic data natively (as shown in Figure 8).

Evaluation Metrics. At the 1024×1024 resolution, we employ DPG-Bench [60] and GenEval [61] to assess semantic alignment and overall generation quality. For 4K generation, evaluations are conducted on the UltraHR-eval4k [37]. We comprehensively assess the performance using Fréchet

Inception Distance (FID) [62] and FID-patch to measure global quality and local details, Inception Score (IS) [63] for generation diversity, as well as Long CLIP Score [64] and Fine-Grained CLIP (FG-CLIP) [65] to evaluate image-text consistency.

Table 1: Comparison of the performance of different methods on DPG-Bench and Geneval. The best results among the pixel text-to-image are highlighted in **bold**.

Method	DPG-Bench						GenEval
	Global↑	Entity↑	Attribute↑	Relation↑	Other↑	Average↑	Overall↑
Latent Text-to-Image Model							
FLUX.1 [Dev] [33]	74.35	90.00	88.96	90.87	88.33	83.84	0.66
SD3 Medium [32]	87.90	91.01	88.83	80.70	88.68	84.08	0.62
Qwen-Image [13]	91.32	91.56	92.02	94.31	92.73	88.32	0.87
Seedream 3.0 [66]	94.31	92.65	91.36	92.78	88.24	88.27	0.84
Z-Image-turbo [12]	91.29	89.59	90.14	92.16	88.68	84.86	0.82
Pixel Text-to-Image Model							
PixelFlow [67]	-	-	-	-	-	77.93	0.60
PixelGen [30]	85.61	86.84	89.35	85.69	87.85	80.01	0.79
Deco [28]	84.91	88.58	87.61	89.70	88.18	81.25	**0.86**
PixNerd [29]	87.16	89.53	88.94	89.26	88.70	82.65	0.73
PixelDiT [31]	-	-	-	-	-	83.50	0.74
L2P	**92.02**	**90.84**	**89.48**	**93.00**	**91.55**	**86.00**	0.76

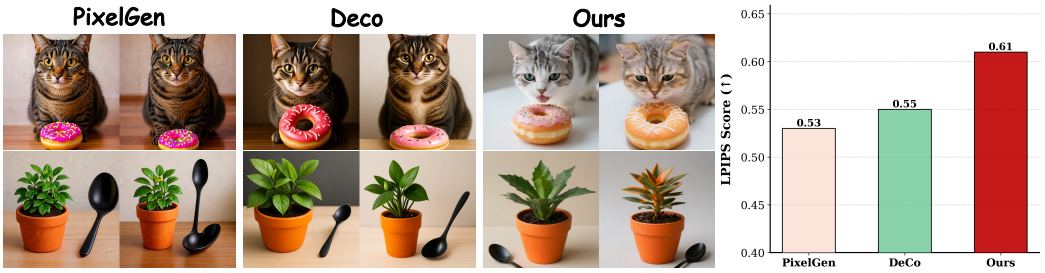

Figure 5: Comparison of generative diversity on GenEval. Compared to PixelGen and Deco, which generally produce visually similar images across various seeds, our approach offers a broader range of structural diversity, yielding higher LPIPS scores.

4.2 Main Result

Quantitative Experiment. To comprehensively evaluate the generative capabilities and text-alignment of the L2P framework, we conduct benchmark testing on DPG-Bench and GenEval, as shown in Table 1. *1) Comparison with Latent Text-to-Image Models:* Results validate L2P's efficient migration of massive latent priors to the pixel space. Notably, despite discarding the VAE and requiring minimal training overhead, L2P achieves a score of 86.00 on DPG-Bench. This slightly exceeds its source LDM, Z-Image-turbo (84.86), effectively maintaining performance on par with the original latent baseline. On GenEval, it retains approximately 93.6% of the source model's performance. *2) Comparison with Pixel Text-to-Image Models:* While L2P establishes a new SOTA among pixel models on DPG-Bench, it yields a lower GenEval score than Deco and PixelGen. However, as shown in Figure 5, across different random seeds, Deco and PixelGen produce highly homogenized, nearly identical images, drastically sacrificing generative diversity, a characteristic also reflected in their low LPIPS scores. In contrast, L2P inherits rich prior knowledge to successfully balance accurate complex attribute binding with high structural diversity. Furthermore, we believe that different instruction tuning settings [68] significantly affect the performance on GenEval. Therefore, we fine-tuned the model following the BLIP3o-60k [69] data style, after which L2P's GenEval score rises from **0.76** to **0.86**. This shows that, with comparable aligned data, the upper limit of L2P's instruction-following capability can be substantially improved.

Qualitative Experiment. Figure 6 qualitatively compares L2P with open-source pixel baselines. Baselines frequently fail at complex attribute binding and text rendering, resulting in structural distortion (Rows 1, 2, 4). In contrast, L2P not only achieves superior text-alignment but also exhibits strong zero-shot generalization. For instance, although transferred on merely 20K English/Chinese samples, L2P seamlessly renders completely unseen Korean text (Row 3). This confirms that rather than overfitting to the transfer data, L2P successfully avoids catastrophic forgetting and harnesses the extensive prior knowledge of the source LDM.

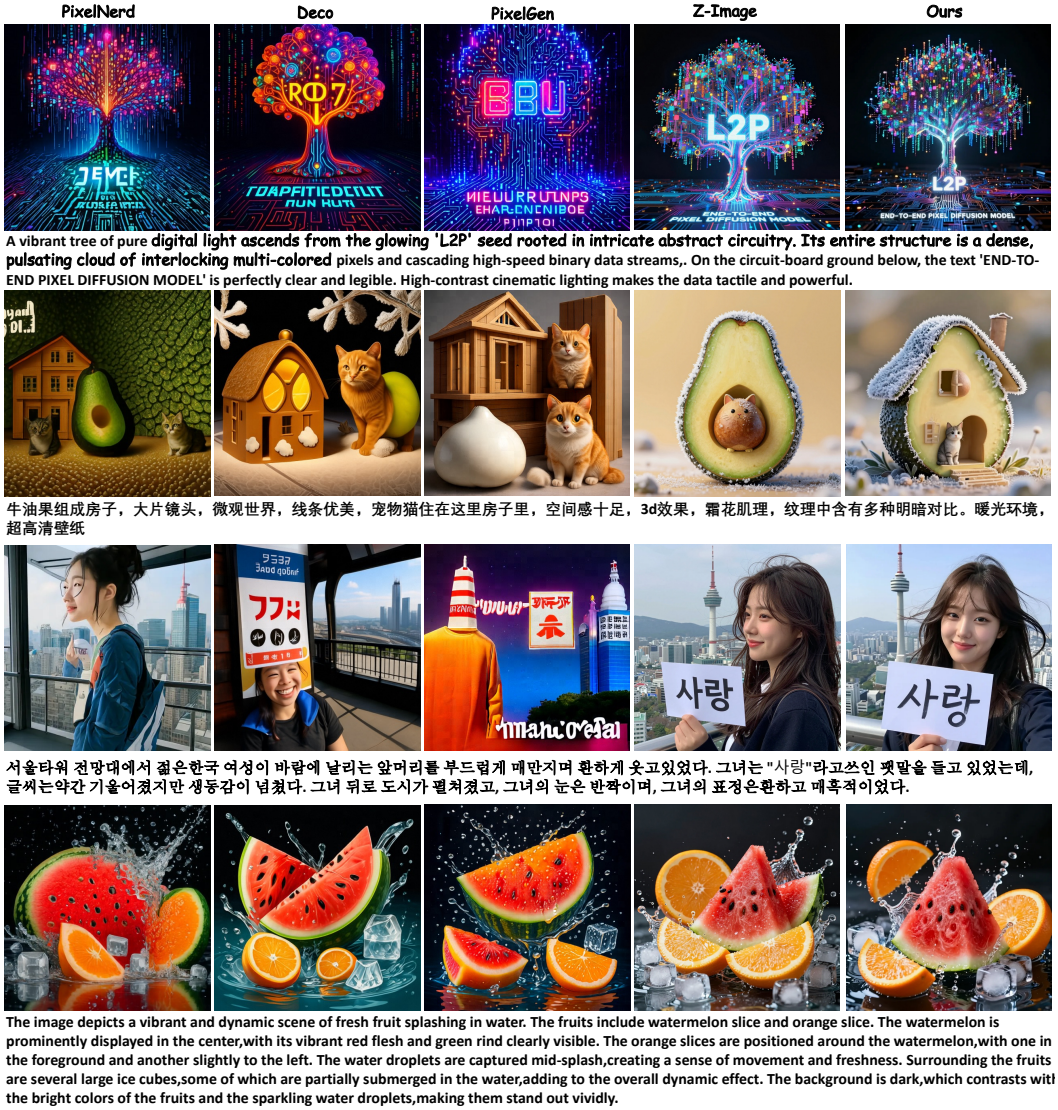

Figure 6: Qualitative comparison of different text-to-image generation models.

4.3 Unlocking Native 4K Generation

Quantitative Experiment. We evaluate the 4K image synthesis capability of our L2P framework against existing 4K solutions. As summarized in Table 2, L2P achieves superior performance in both global visual quality and local structural coherence. Specifically, our method establishes advanced performance in image fidelity, yielding the lowest FID and $\text{FID}_{\text{patch}}$. Furthermore, L2P attains the highest Inception Score, reflecting excellent generation diversity. In terms of semantic alignment, L2P preserves the rich conditional priors of the source LDM, posting highly competitive CLIP and FG-CLIP scores.

Qualitative Experiment. *1) Comparison with different 4K Models*: As shown in Figure 7, compared to existing baselines, L2P effectively mitigates the common issues of over-smoothing and artificial artifacts, ensuring the faithful synthesis of exquisite micro-details. *2) Unlocking Native 4K from LDMs*: As illustrated in Figure 8, the source Z-Image fails to generate semantic content directly at 4K, and merely upsampling its 1K outputs severely blurs high-frequency details. In contrast, L2P natively generates crisp 4K outputs. This indicates that L2P not only seamlessly inherits the rich priors of the source LDM but also effectively expands its generative boundaries, elevating the resolution ceiling with minimal training overhead.

Table 2: Quantitative comparison of 4K ultra-high-resolution image generation. Best results are highlighted in **bold**.

Model	FID↓	FID_{patch}↓	IS↑	CLIP↑	FG-CLIP↑
Training-Free Generation Method					
I-Max [53]	37.12	34.42	11.78	31.53	27.90
HiFlow [54]	38.54	22.74	10.62	31.49	27.84
Training-Based Generation Method					
Pixart-σ [51]	35.60	23.63	12.23	31.76	28.74
SANA [8]	37.26	28.94	12.05	31.79	**29.31**
Diffusion4K [52]	41.69	39.67	12.06	31.44	27.03
Ours	**33.46**	**21.77**	**12.28**	**31.88**	28.22



Figure 7: Qualitative comparison of 4K image generation.

4.4 Ablation Study

Impact of Training Data Source. Figure 9(a) compares source, real, and cross-model (GLM) [70] data. Source data achieves rapid convergence and optimal performance. Conversely, collecting uniformly distributed natural images incurs prohibitive costs; thus, we employ a random 20k subset from UltraHR-100K as the real-data baseline. This variant suffers from sluggish convergence and degraded quality. Such a stark contrast inversely highlights the comprehensive diversity and well-aligned distribution of our curated source dataset. Cross-model data yields intermediate results, trailing source data due to imperfect prior alignment.

Effectiveness of Shallow Layer Tuning. We ablate the number of trainable layers by comparing our default shallow tuning (5 layers) against mid-level (10 layers) and full-layer tuning. As illustrated in Figure 9(b), shallow tuning yields steady performance improvements across training steps. In



Figure 8: Unlocking native 4K generation with L2P.

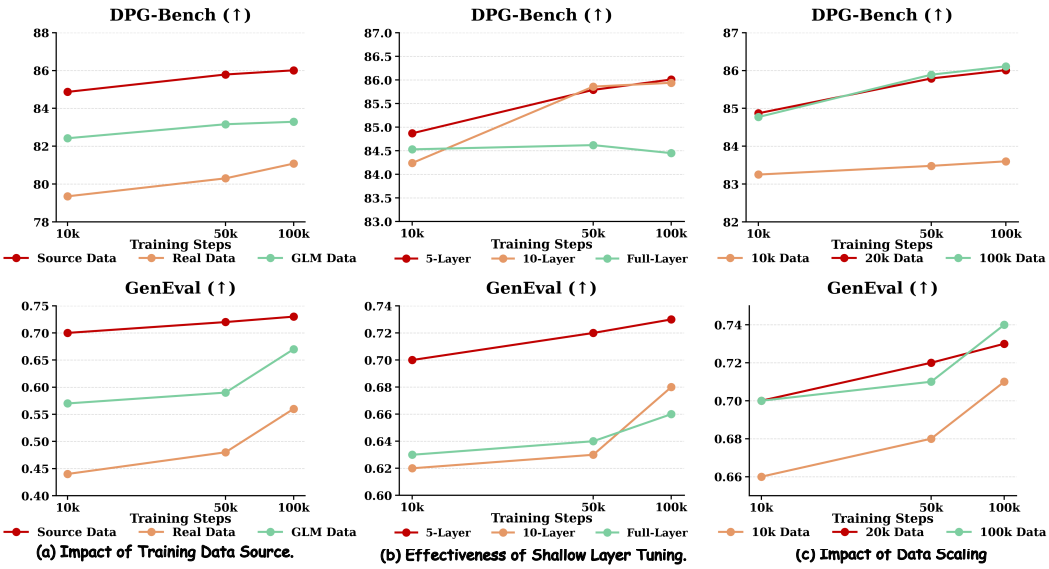

Figure 9: Ablation studies of our proposed L2P framework.

stark contrast, full-layer tuning suffers from clear performance stagnation and degraded generation quality. This phenomenon indicates that unconstrained parameter updates severely disrupt the rich pre-trained priors residing in the deeper layers. By exclusively tuning the shallowest layers, our method successfully learns the latent-to-pixel mapping while optimally preserving the source model's core generative knowledge.

Impact of Data Scaling. Figure 9(c) shows performance across 10k, 20k, and 100k synthetic samples. While increasing data from 10k to 20k yields substantial gains, performance clearly saturates beyond 20k. This early convergence demonstrates L2P's extreme data efficiency. It confirms that our synthetic dataset is sufficiently diverse to comprehensively cover the data manifold, enabling rapid and low-cost adaptation.

5 Conclusion

In this paper, we propose the Latent-to-Pixel (L2P) transfer paradigm to overcome the VAE-induced limitations of LDMs and bypass the prohibitive costs of training pixel-space models from scratch. By discarding the VAE in favor of large-patch tokenization, freezing core intermediate layers, and constructing a multi-dimensional prompt dataset to fit a smooth synthetic data manifold, L2P

successfully migrates deep semantic priors to the pixel space. Notably, this is achieved using only 8 GPUs with zero real-data cost. Furthermore, eliminating the VAE memory bottleneck seamlessly unlocks native 4K ultra-high resolution generation. Experiments demonstrate that L2P robustly maintains state-of-the-art generative performance. Ultimately, L2P lowers the barrier to developing advanced pixel-space diffusion models, offering a practical strategy for exploring VAE-free, high-resolution generation under limited resources.

6 Acknowledge

This work was supported by Natural Science Foundation of China: No. 62406135, Natural Science Foundation of Jiangsu Province: BK20241198, the Gusu Innovation and Entrepreneur Leading Talents: No. ZXL2024362, and CCF-Tencent Rhino-Bird Open Research Fund.

References

[1] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[3] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[4] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[5] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[6] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[7] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[8] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Yujun Lin, Zhekai Zhang, Muyang Li, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

[9] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[11] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

[12] Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

[13] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[14] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[15] Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Ragd: Regional-aware diffusion model for text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19331–19341, 2025.

[16] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6818–6828, 2024.

[17] Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3dis: Depth-driven decoupled instance synthesis for text-to-image generation. *arXiv preprint arXiv:2410.12669*, 2024.

[18] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[19] Nikai Du, Zhennan Chen, Shan Gao, Zhizhou Chen, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv preprint arXiv:2503.23461*, 2025.

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[21] Xin Cai, Zhiyuan You, Zhoutong Zhang, and Tianfan Xue. Da-vae: Plug-in latent compression for diffusion via detail alignment. *arXiv preprint arXiv:2603.22125*, 2026.

[22] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712, 2025.

[23] Maciej Kilian, Varun Jampani, and Luke Zettlemoyer. Computational tradeoffs in image synthesis: Diffusion, masked-token, and next-token prediction. *arXiv preprint arXiv:2405.13218*, 2024.

[24] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv preprint arXiv:2410.10733*, 2024.

[25] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, pages 393–411. Springer, 2024.

[26] Zhennan Chen, Junwei Zhu, Xu Chen, Jiangning Zhang, Xiaobin Hu, Hanzhen Zhao, Chengjie Wang, Jian Yang, and Ying Tai. Dip: Taming diffusion models in pixel space. *arXiv preprint arXiv:2511.18822*, 2025.

[27] Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

[28] Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. *arXiv preprint arXiv:2511.19365*, 2025.

[29] Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. Pixnerd: Pixel neural field diffusion. *arXiv preprint arXiv:2507.23268*, 2025.

[30] Zehong Ma, Ruihan Xu, and Shiliang Zhang. Pixelgen: Pixel diffusion beats latent diffusion with perceptual loss. *arXiv preprint arXiv:2602.02493*, 2026.

[31] Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. *arXiv preprint arXiv:2511.20645*, 2025.

[32] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[33] BlackForest. Black forest labs; frontier ai lab, 2024.

[34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[35] Zhennan Chen, Rongrong Gao, Tian-Zhu Xiang, and Fan Lin. Diffusion model for camouflaged object detection. In *ECAI 2023*, pages 445–452. IOS Press, 2023.

[36] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[37] Chen Zhao, En Ci, Yunzhe Xu, Tiehan Fan, Shanyan Guan, Yanhao Ge, Jian Yang, and Ying Tai. Ultrahr-100k: Enhancing uhr image synthesis with a large-scale high-quality dataset. *Advances in Neural Information Processing Systems*, 2025.

[38] Ji Xie, Dewei Zhou, Xinyu Huang, Zhennan Chen, and Xun Wang. Paint-anything: Unified any-color control for image generation and editing. *arXiv preprint arXiv:2609.20816*, 2026.

[39] Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8281–8291, 2024.

[40] Zhanxin Gao, Beier Zhu, Liang Yao, Jian Yang, and Ying Tai. Subject-consistent and pose-diverse text-to-image generation. *arXiv preprint arXiv:2507.08396*, 2025.

[41] Shaoqi Dong, Chaoyou Fu, Haihan Gao, Yi-Fan Zhang, Chi Yan, Chu Wu, Xiaoyu Liu, Yunhang Shen, Jing Huo, Deqiang Jiang, et al. Vita-vla: Efficiently teaching vision-language models to act via action expert distillation. *arXiv preprint arXiv:2510.09607*, 2025.

[42] Dewei Zhou, You Li, Zongxin Yang, and Yi Yang. Refineanything: Multimodal region-specific refinement for perfect local details. *arXiv preprint arXiv:2604.06870*, 2026.

[43] Chen Zhao, Chenyu Dong, Weiling Cai, and Yueyue Wang. Learning a physical-aware diffusion model based on transformer for underwater image enhancement. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

[44] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International conference on learning representations*, volume 2025, pages 1045–1064, 2025.

[45] Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Zhenheng Yang, Chaoyou Fu, Xiang Li, Jian Yang, and Ying Tai. Instancecap: Improving text-to-video generation via instance-aware structured caption. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28974–28983. IEEE, 2025.

[46] Haojun Chen, Haoyang He, Chengming Xu, Qingdong He, Junwei Zhu, Yabiao Wang, Zhucun Xue, Xianfang Zeng, Zhennan Chen, Xiaobin Hu, et al. Pixverve: Advancing native uhr image generation to 100mp with a large-scale high-quality dataset. *arXiv preprint arXiv:2605.20147*, 2026.

[47] Zhanxin Gao, Beier Zhu, Jian Yang, and Ying Tai. Codi: Subject-consistent and pose-diverse text-to-image generation. In *International Conference on Learning Representations*, volume 2026, pages 113718–113740, 2026.

[48] Zhizhou Chen, Shanyan Guan, Zhanxin Gao, En Ci, Yanhao Ge, Wei Li, Zhenyu Zhang, Jian Yang, and Ying Tai. Vins-120k: Ultra high-resolution image editing with a large-scale dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15302–15312, June 2026.

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

[50] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[51] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[52] Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23464–23473, 2025.

[53] Ruoyi Du, Dongyang Liu, Le Zhuo, Qin Qi, Hongsheng Li, Zhanyu Ma, and Peng Gao. I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow. 2024.

[54] Jiazi Bu, Pengyang Ling, Yujie Zhou, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Hiflow: Training-free high-resolution image generation with flow-aligned guidance. *arXiv preprint arXiv:2504.06232*, 2025.

[55] Chen Zhao, Yunzhe Xu, Zhizhou Chen, Enxuan Gu, Kai Zhang, Xiaoming Liu, Jian Yang, and Ying Tai. From zero to detail: A progressive spectral decoupling paradigm for uhd image restoration with new benchmark. *arXiv preprint arXiv:2604.15654*, 2026.

[56] Haojun Chen, Haoyang He, Chengming Xu, Qingdong He, Junwei Zhu, Yabiao Wang, Zhucun Xue, Xianfang Zeng, Zhennan Chen, Xiaobin Hu, Hao Zhao, Yong Liu, Jiangning Zhang, and Dacheng Tao. Pixverve: Advancing native uhr image generation to 100mp with a large-scale high-quality dataset. *arXiv preprint*, 2026.

[57] Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

[58] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023.

[59] Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion (sid2): 1.5 fid on imagenet512 with pixel-space diffusion. *arXiv preprint arXiv:2410.19324*, 2024.

[60] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

[61] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

[62] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[63] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[64] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024.

[65] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. *arXiv preprint arXiv:2505.05071*, 2025.

[66] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.

[67] Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025.

[68] Ji Xie, Luke Zettlemoyer, Xudong Wang, et al. Reconstruction alignment improves unified multimodal models. In *International Conference on Learning Representations*, volume 2026, pages 120095–120137, 2026.

[69] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025.

[70] Z.ai. Glm-image: Auto-regressive for dense-knowledge and high-fidelity image generation, 2026.

Supplementary Material: L2P: Unlocking Latent Potential for Pixel Generation

1 More Implementation Details

Table 1: Architectural configurations and hyperparameter settings.

DiT Architecture	
Input Channels	3
Patch Size	16×16
Hidden Dimension	3840
Transformer Layers	30
Attention Heads	30
Head Dimension	128
Detailer Head Architecture	
DownSampling Path	16→8→4→2→1
UpSampling Path	1→2→4→8→16
DownSampling Channel	3→64→128→256→512
Bottleneck	(512+3840)→512
UpSampling Channel	512→256→128→64→64
Output Layer	64→3
Optimization	
Optimizer	AdamW
Learning Rate	0.00005
Weight Decay	0.01
Batch Zize	8
Gradient Accumulation Steps	1

Table 1 details the precise architectural configurations and training hyperparameters for our L2P framework.

DiT Architecture. To seamlessly inherit the pre-trained latent priors, our DiT backbone strictly preserves the structural configurations of the source Latent Diffusion Model (LDM). The only structural modifications occur at the input stage: we adjust the input channels for raw RGB images and replace the standard VAE with a large-patch tokenization strategy. This allows the model to directly process pixel-space inputs while utilizing the frozen intermediate transformer blocks.

Detailer Head Architecture: For the shallow layers responsible for the latent-to-pixel transformation, we adopt the Detailer Head architecture proposed in DiP [26]. It employs a symmetric encoder-decoder paradigm for spatial downsampling and upsampling. To integrate this head into our framework, we specifically adapt its bottleneck dimension to align with our DiT backbone, enabling the concatenation and fusion of features extracted from the frozen intermediate layers.

Optimization. During the L2P transfer phase, we freeze the massive intermediate layers of the source LDM and exclusively train the shallow layers. The network is optimized using AdamW with a learning rate of 5×10^{-5} and a weight decay of 0.01. We utilize a batch size of 8 with no gradient accumulation (steps set to 1).

```
# Role:
You are a Synthetic Data Generator specialized in creating high-quality, diverse
text-to-image prompts for training state-of-the-art diffusion models. Your goal is
to produce natural English prompts that read like professional image captions or
detailed visual descriptions.
```

```
# Global Constraints
1. **Format**: Output strictly valid JSON.
2. **Language**: Prompts must be in fluent, natural English. Avoid "tag soup"
(comma-separated lists of tags). Use complete sentences.
3. **Escaping**: Since prompts may contain double quotes for text rendering…
4. **Safety**: Do not reference NSFW content…

# Prompt Types
- **Detailed**: 3-5 sentences. Elaborate on minute details…
```

```
## 1. Nature
- **Objects**: The input is the **Object Name**. Focus on material…
- **Cityscape**: The input is a **City/Location Type**. Mention time…
- **Food**: Appetizing descriptions. Mention steam…
- **Plants**: Specific species, foliage details, macro shots of veins/dew…
- **Indoor**: Interior design. Mention furniture style…
- **Landscape**: Vast scenes. Define foreground/background…
- **Animals**: Specific species, fur/feather texture…

## 2. Design
- **Arts**: Traditional or digital art. Specify medium…
- **Posters**: **MUST** include text inside escaped quotes…
- **Slides**: Presentation slides. **MUST** include a title/heading…
- **Cartoon**: Illustration styles. Anime, western comic…
- **UI**: App or Web interfaces. **MUST** include button…
- **Others**: Abstract patterns, surrealism, mixed media…

## 3. People
- **Portrait**: Close-ups or half-body. Focus on skin texture…
- **Sports**: Dynamic action. Freeze-frame or motion blur…
- **Activities**: Daily life scenarios. Focus on interaction with tools…
- **Others**: Crowds, silhouettes, or abstract human representations…

## 4. Synthetic (Text Rendering Focus)
- **Chinese Text**: **MUST** include specific Chinese characters…
- **English Text**: **MUST** include specific English phrases…
- **Others**: Numbers, code snippets, or mathematical formulas…
```

```
# Diversity & Randomization Checklist
To ensure the dataset is non-repetitive:
- **Vary Viewpoints**: Low angle, bird's eye, dutch angle, macro, wide shot.
- **Vary Lighting**: Golden hour, blue hour, cinematic lighting, volumetric fog,
harsh sunlight, studio softbox.
- **Vary Styles**: Photorealistic, Cinematic, Analog Film (Kodak Portra), Digital Art,
Minimalist.

# Anti-Pattern Constraints
- **NO REPETITION**: Do not use the same sentence structure for consecutive prompts.
- **VARIETY**: Do not start every prompt with "A photo of..." or "A cinematic shot
of..." or "A ...". Use diverse sentence openers.
```

```
# Output Structure
Each prompt should be on a separate line, without any prefixes or formatting.
[
  "prompt1",
  "prompt2",
...
]
```

```
# Task
Generate prompt for each subject on the following list of subjects.
**Category**: {{Nature}} - {{Animals}}
**Subject List**:
{}
```

Figure 1: The template for General Prompt Generation. This template guides the LLM to synthesize high-quality, diverse image descriptions by enforcing strict stylistic and structural constraints across multiple predefined categories.

```
# Role:
You are a strict QA Auditor for a Text-to-Image dataset.
Your task is to evaluate the following image prompt based on these strict criteria.
```

```
1. **Format & Structure**:
- NO "tag soup" (comma-separated lists).
- Must use fluent, complete English sentences.

2. **Text Rendering Syntax (CRITICAL)**:
- If the prompt describes specific text to be written/displayed (English or Chinese),
that text **MUST** be enclosed in escaped double quotes (e.g., \\"Hello\\").
- **Correct**: A neon sign reading \"JAZZ NIGHT\".
- **Incorrect**: A neon sign reading JAZZ NIGHT.

3. **Safety & Ethics**:
- NO NSFW, violence, sexual content, or political sensitivity.

4. **Copyright**:
- NO trademarked brand names or copyrighted characters.
```

```
Analyze the prompt below.
Output strictly in JSON format:
{
"is_valid": true/false,
"reason": "If false, explain specifically why. If true, put 'Pass'."
}
```

Figure 2: The template for Automated Prompt Filtering. This template systematically evaluates generated prompts to discard instances violating formatting, text-rendering syntax, safety, or copyright standards.

2 More Data Construction Details

To provide further transparency into our data construction pipeline, Figures 1 and 2 present the system prompts used for General Prompt Generation and Automated Prompt Filtering. Specifically, the generation template enforces strict guidelines on prompt diversity, structural formatting, and text-rendering syntax to elicit complex, high-quality scene descriptions from the LLM. Furthermore, we detail the filtering mechanism designed to automatically discard prompts that fail to meet our safety, ethical, and copyright standards, thereby guaranteeing the overall integrity of the training corpus.

3 More Experimental Results

3.1 Quantitative Results

Figure 3 presents the ablation study of the noise shift parameter for 4K generation. As the parameter increases from 1 to 4, the FID score drops significantly, confirming that skewing the noise schedule toward higher noise levels is essential to fully corrupt the dense 4K image signals. Beyond this optimal point (shift parameter=5), performance slightly degrades due to over-corruption.

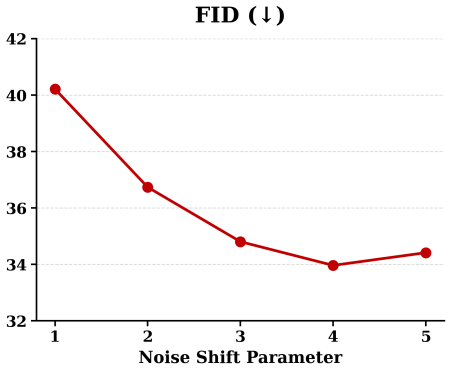

Figure 3: Impact of the noise shift parameter after 100k training steps.

3.2 Visualization Results

We provide extended qualitative results to demonstrate the visual fidelity and scalability of the L2P framework. Figure 4 presents diverse text-to-image generations at 1K resolution, validating the successful transfer of the

source LDM's powerful generative priors. As shown in Figures 5, L2P unlocks native 4K ultra-high resolution generation. This capability directly stems from eliminating the VAE memory bottleneck, allowing the pixel-space model to render exquisite details at extreme resolutions without prohibitive computational overhead.

Figures 6 present the visual results of zero-shot resolution extrapolation. Benefiting from the pure pixel-space formulation and the elimination of VAE-induced coupling, L2P extends its generative boundaries far beyond its training resolution. The generated 8K images exhibit exceptional global structural consistency and faithful micro-details, robustly validating the extrapolation capabilities of our L2P paradigm.

4 Limitations and Future Work

While the proposed L2P paradigm enables highly efficient latent-to-pixel transfer, it naturally entails certain limitations. First, our reliance on synthetic images generated by the source LDM implies that the semantic and compositional capabilities of our model are fundamentally upper-bounded by the source model's priors. Although incorporating real-world datasets could theoretically circumvent this knowledge bottleneck, it would reintroduce a strong dependency on data curation and quality, diverging from our cost-effective, self-contained transfer objective. Second, a pivotal advantage of operating natively in pixel space is the straightforward integration of fine-grained, task-specific loss functions (e.g., pixel-level perceptual or physics-based constraints) for downstream applications. To preserve the simplicity and universality of the L2P framework, we deliberately omit the exploration of such tailored objectives in this work. Leveraging direct pixel-space regularizations for specialized generation tasks remains a highly promising avenue for future research.



Figure 4: More text-to-image generation results at 1024×1024 resolution. The synthesized images exhibit diverse stylistic rendering and complex semantic alignment, demonstrating the successful migration of latent priors to the pixel space via our L2P framework.



Figure 5: More native 4K ultra-high resolution generation results. By eliminating the VAE memory bottleneck inherent in traditional LDMs, the L2P framework seamlessly scales to generate 4K images with exquisite details and crisp textures.



Figure 6: Visualizations of 8K ultra-high resolution zero-shot extrapolation.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction have clearly included the motivations, important assumptions, and contributions made in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The authors have discussed the limitations of the work in the Limitations and Future Work section.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The authors have clearly provided the full set of assumptions and complete proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The authors have presented all the experimental details in the paper.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: All data, models and codes will be made public after being sorted and reviewed.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The authors have presented all the training and test details in the Experiments section and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The authors have presented all the experimental details in the paper with sufficient instructions provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The authors have provided sufficient information on the computer resources in the Experiments section and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This paper explores the fundamental architecture of the model.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The authors have described safeguards about dataset licensing in the paper.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The authors have cited the original paper that produced the code package or dataset and included the licenses of existing assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The author has provided detailed information in the text regarding training, license, limitations, etc.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The authors have presented the LLM usage details in supp.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.