GUIDED IMAGE ENHANCEMENT PATH DIFFUSION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion models generate coarse structure and fine details within a single denoising process, leaving their coarse-to-fine hierarchy only implicitly modeled. We introduce **GIEP** (Guided Image Enhancement Path), which decomposes an image into ordered coarse-to-fine components and models their conditional dependencies. We show that the score of each partial reconstruction can be expressed as a sum of reusable per-level score contributions. This result yields a cumulative score-matching objective that supervises all levels jointly in a single shared network. Causal attention along the path axis enforces dependencies from coarse structure to fine detail, while level-specific text conditioning guides different levels of visual granularity. Experiments demonstrate leading perceptual quality and competitive compositional fidelity at a substantially smaller parameter scale; human evaluators also prefer GIEP for detail realism and saliency.

1 INTRODUCTION

Text-to-image (T2I) diffusion models have achieved remarkable success in generating realistic images from natural language descriptions Rombach et al. (2022); Podell et al. (2023); Esser et al. (2024). The field has evolved across successive generative paradigms Zhang & Tang (2024). GAN-based approaches Goodfellow et al. (2014); Reed et al. (2016); Zhang et al. (2017); Xu et al. (2018); Karras et al. (2019; 2020); Kang et al. (2023) established feasibility but suffered from mode collapse, and autoregressive Transformer Vaswani et al. (2017) models Ramesh et al. (2021); Esser et al. (2021); Yu et al. (2022) imposed a rigid, slow raster-scan order, before denoising diffusion Sohl-Dickstein et al. (2015); Ho et al. (2020) became dominant Dhariwal & Nichol (2021). Text-guided diffusion has since advanced along every axis, spanning CLIP Radford et al. (2021) conditioning Nichol et al. (2022); Ramesh et al. (2022) and language-model text encoders Saharia et al. (2022), denoising in a compressed latent space Rombach et al. (2022), improved prompt adherence and scale Ho & Salimans (2022); Podell et al. (2023); Betker et al. (2023), a shift from U-Net Ronneberger et al. (2015) to Diffusion Transformer Peebles & Xie (2023); Chen et al. (2024); Li et al. (2024b); Labs (2024) backbones, and flow-matching objectives Liu et al. (2022); Lipman et al. (2023); Tong et al. (2023), culminating in Stable Diffusion 3 Esser et al. (2024). Despite this progress, diffusion models capture the coarse-to-fine hierarchy only implicitly. Their denoising process operates on a monolithic latent that *entangles the full frequency spectrum*, with coarse structure, fine textures, and boundaries represented within a single undifferentiated signal Si et al. (2024); Rissanen et al. (2023); Sun et al. (2026).

This entanglement bears directly on fine-grained attributes such as delicate textures and intricate boundaries, which must be resolved jointly with the coarse structure they overlay and can overwhelm it Si et al. (2024). The coarse-to-fine ordering governing this interaction is present in diffusion only as an implicit inductive bias of the Gaussian forward process, one the model must discover from data rather than expose as a structure it can control Rissanen et al. (2023). Absent such a handle, the resulting texture-structure misalignment surfaces most visibly on compositionally complex prompts Chefer et al. (2023); Feng et al. (2023). Recent frequency-space diffusion exploits this ordering for better generation Ning et al. (2025). The same limitation propagates to restoration, super-resolution, and editing, where structure must be preserved while fine detail is regenerated Sun et al. (2026). In each case, the limitation traces to a common root: the entanglement of global semantics and local detail in the spatial domain Sun et al. (2026), leaving the denoising process without explicit control over the level of detail.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To overcome this limitation, GIEP reformulates image generation as *sampling from a structured enhancement path*, decomposing the target into K ordered spectral components from coarse structure to fine detail and modeling their joint distribution as a factored score field with causal dependencies. First, we derive a *cumulative score matching* objective (Theorem 3.1) based on an additive decomposition of the global score along a frequency-based path into K atomic scores reusable across cumulative levels. A single shared network jointly estimates all path-level scores in one forward pass. This design obviates the cascaded per-stage models of prior hierarchical pipelines. Second, to model coarse-to-fine dependencies among jointly estimated levels, we introduce *causal path attention*. This causally masked self-attention operates along the path axis, architecturally restricting each level to depend on coarser ones. Third, *hierarchical text conditioning* routes multi-granularity text features to path-level-specific representations, granting explicit semantic control at each frequency band. Together, these components form a dual-axis framework: the *diffusion axis* performs iterative denoising over time, while the *path axis* models coarse-to-fine dependencies across path levels (Fig. 1). Extensive experiments demonstrate that GIEP delivers leading perceptual quality and competitive compositional fidelity at a substantially smaller parameter scale, with ablations validating our core design choices.

2 Related Work

Diffusion generative models. Diffusion models generate samples by reversing gradual noising. DDPMs Ho et al. (2020) use a Gaussian Markov chain to corrupt data and predict added noise for iterative denoising. Score-based models Song & Ermon (2019); Song et al. (2021) formulate continuous-time SDEs whose reverse dynamics use scores of noise-perturbed densities. These scores are learned through denoising score matching Vincent (2011) without an intractable normalization constant. Flow matching Lipman et al. (2023) and rectified flow Liu et al. (2022) learn velocity fields along arbitrary probability paths, yielding simpler objectives and straighter trajectories. Fast ODE solvers Song et al. (2020); Lu et al. (2022) reduce sampling steps by orders of magnitude. Backbones have shifted from convolutional U-Nets Rombach et al. (2022) to DiTs Peebles & Xie (2023), powering Stable Diffusion 3 Esser et al. (2024) and FLUX Labs (2024). Across these advances the state remains a monolithic latent on one axis, its coarse-to-fine structure an implicit inductive bias the model must discover rather than a handle it exposes Rissanen et al. (2023). GIEP decomposes the global score into K atomic per-level components, jointly estimated via cumulative score matching (Theorem 3.1), making every level of detail separately addressable and coarse-to-fine control explicit.

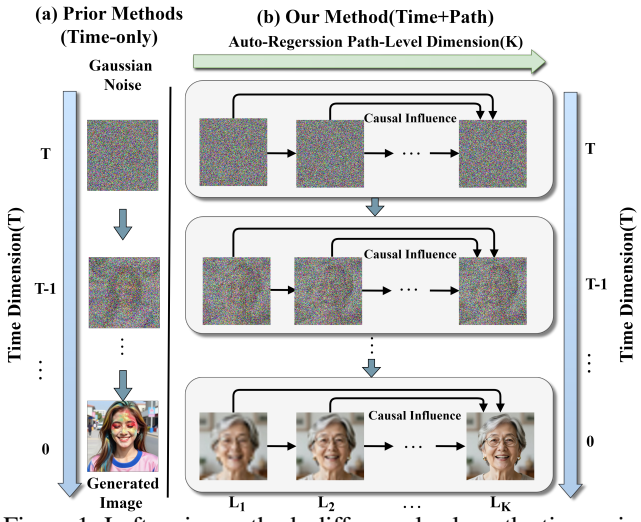

Figure 1: Left: prior methods diffuse only along the time axis. Right: GIEP adds a path axis, jointly modeling temporal denoising and coarse-to-fine refinement.

Multi-scale and frequency-aware image synthesis. Natural images exhibit a spectral hierarchy of low-frequency structure and high-frequency detail Burt & Adelson (1983), and the isotropic schedule that hides it is one member of a broader non-isotropic family Hoogeboom & Salimans (2023). Left implicit, the hierarchy can in some cases become a liability, with high-frequency features overwhelming low-frequency semantics Si et al. (2024). Cascaded pipelines Ho et al. (2022); Pernias et al. (2023); Saharia et al. (2022); Team (2023) split the hierarchy across separately trained stages that propagate error, while inference-time methods rebalance an unchanged model by rescaling skips Si et al. (2024), low-pass filtering Huang et al. (2024); Koh et al. (2025), or reweighting the loss Choi et al. (2022). Wavelet diffusion Phung et al. (2023); Guth et al. (2022) and W-Edit Sun et al. (2026) go further, separating structural anchors from editable detail in coefficient space. Each confirms that

diffusion's frequency structure is real and better made explicit, yet all bands remain jointly denoised under one schedule, leaving the level of detail a byproduct rather than an indexable coordinate. GIEP instead makes frequency a second explicit axis alongside diffusion time, ordering levels causally, scoring them atomically, and conditioning them independently, so that the detail hierarchy becomes the structural backbone of generation and cross-stage error propagation is eliminated by construction.

Text conditioning and compositional control. Modern T2I systems use CLIP Radford et al. (2021), large-language-model features Saharia et al. (2022); Raffel et al. (2020), SDXL's dual encoders Podell et al. (2023), or SD3's triple-encoder stack Esser et al. (2024); these features are concatenated and applied uniformly across denoising stages and spatial locations. Structured methods address this uniformity along complementary axes. P+ Voynov et al. (2023) routes tokens to U-Net resolutions, and Attend-and-Excite Chefer et al. (2023) enforces per-token attention coverage. Structured Diffusion Feng et al. (2023) uses parsed noun phrases, while ELLA Hu et al. (2024) injects timestep-aware LLM features. RPG Yang et al. (2024) assigns regional prompts via multimodal LLMs, and DiT-ST Zhang et al. (2025) hierarchically schedules split-text prompts. TextCraftor Li et al. (2024a) fine-tunes CLIP with differentiable rewards, while DATE Na et al. (2025) adapts text embeddings to intermediate latents at each denoising step without retraining. None of these axes explicitly corresponds to visual detail levels. GIEP instead aligns text granularity with the levels of detail, a coarse-to-fine hierarchy (e.g., spectral decomposition) in which global descriptions condition coarse levels and fine-grained attributes condition detail levels.

3 Theoretical Framework

3.1 Problem Formulation

We reformulate image generation as sampling from a *path distribution* $p(\mathbf{P})$, where $\mathbf{P}$ is a structured sequence of progressive refinements from coarse to fine.

Definition 3.1 (Image Decomposition Operator)**.** *Let $\mathcal{T} : \mathbb{R}^D \xrightarrow{\sim} V_0 \times V_1 \times \cdots \times V_{K-1}$ be a linear, invertible operator that decomposes an image $\mathbf{x}_0$ into K ordered spectral components $\mathcal{T}(\mathbf{x}_0) = (I_0, I_1, \ldots, I_{K-1})$, where the ordering progresses from coarse (k=0) to fine (k=$K-1$). The inverse satisfies $\mathbf{x}_0 = \sum_{k=0}^{K-1} I_k$.*

Concrete choices (FFT bandpass, Laplacian pyramids, wavelet transforms) are detailed in Sec. 3.3.

Definition 3.2 (Enhancement Path)**.** *The* Enhancement Path *of $\mathbf{x}_0$ under $\mathcal{T}$ is the ordered sequence of cumulative partial images $\mathbf{P} = (\mathcal{I}_0, \mathcal{I}_1, \ldots, \mathcal{I}_{K-1})$ with $\mathcal{I}_k = \sum_{j=0}^{k} I_j$. By the invertibility of $\mathcal{T}$, $\mathcal{I}_{K-1} = \mathbf{x}_0$ (Completeness).*

Proposition 3.1 (Sampling Equivalence)**.** *Since $\mathcal{T}$ is a linear invertible map with constant Jacobian, the change-of-variables formula gives $p(\mathbf{x}_0) = p(\mathbf{P}) \cdot |\det \mathbf{J}_{\mathcal{T}}|^{-1}$, so sampling $\mathbf{x}_0 \sim p(\mathbf{x}_0)$ reduces to sampling from $p(\mathbf{P})$.*

Assumption 3.1 (Detail Hierarchy)**.** *Under the coarse-to-fine ordering induced by $\mathcal{T}$, fine-scale detail (edges, textures) is statistically conditioned on coarse-scale structure (shapes, illumination), while the reverse dependency is weak.*

This well-known property of natural images motivates a coarse-to-fine autoregressive factorization. Applying the chain rule:

$$p(\mathbf{P}) = p(I_0, I_1, \ldots, I_{K-1}) = \prod_{k=0}^{K-1} p(I_k \mid I_{<k}), \tag{1}$$

where $I_{<k} = \{I_0, \ldots, I_{k-1}\}$. Each factor generates the k-th detail level conditioned on all coarser levels, naturally enabling causal masking (Sec. 4.1). However, modeling $p_{\mathrm{AR}}(\mathbf{P}) = \prod_k p(I_k \mid I_{<k})$ subject to $\int p_{\mathrm{AR}}(\mathbf{P})\, d\mathbf{P} = 1$ requires evaluating an intractable partition function over high-dimensional image path space.

3.2 From Auto-Regression to Score-Based Path Generation

Score-based generative modeling Song et al. (2021); Song & Ermon (2019) sidesteps this intractability by targeting the score $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma)$, which requires no partition function. Via denoising score

matching Vincent (2011); Karras et al. (2022), $\nabla_{\mathbf{x}} \log p(\mathbf{x};\sigma) = (D_\theta(\mathbf{x};\sigma) - \mathbf{x})/\sigma^2$, and sampling proceeds by integrating the probability flow ODE Karras et al. (2022) from noise to data. We now show that the global score decomposes into components matching the autoregressive factors.

Proposition 3.2 (Additive Score Decomposition)**.** *Let $\mathcal{T}$ be a linear invertible decomposition (Def. 3.1). The score of the noisy state decomposes as:*

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) = \sum_{k=0}^{K-1} \nabla_{\mathbf{x}_t} \log p_t\Big(I_k^{(t)} \mid I_{<k}^{(t)}\Big). \tag{2}$$

Proof sketch. From Eq. 1, $\log p_t(\mathbf{x}_t) = \sum_k \log p_t(I_k^{(t)} \mid I_{<k}^{(t)}) + C$ with constant $C = -\log|\det \mathbf{J}_{\mathcal{T}}|$. Applying $\nabla_{\mathbf{x}_t}$ yields the result. Full proof in Appendix A. □

The global score thus decomposes into K additive components. We next show these are reusable across cumulative levels, enabling a single set of atomic scores for the entire path.

Definition 3.3 (Hierarchical Perturbation)**.** *For each level k, let $\boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and define the* level-k noisy state *as $\mathbf{x}_t^{(k)} = \alpha_t \mathcal{I}_k + \sigma_t \boldsymbol{\epsilon}^{(k)}$, with perturbation kernel:*

$$p\Big(\mathbf{x}_t^{(k)} \mid \{I_j\}_{j=0}^k\Big) = \mathcal{N}\Big(\mathbf{x}_t^{(k)};\ \alpha_t \mathcal{I}_k,\ \sigma_t^2 \mathbf{I}\Big). \tag{3}$$

The conditional score $f_j(\mathbf{x}) \triangleq \log p_t(I_j^{(t)} \mid I_{<j}^{(t)})$ depends on $\mathbf{x}$ only through projections $(\Pi_{V_0}(\mathbf{x}), \ldots, \Pi_{V_j}(\mathbf{x}))$. For any $k \geq j$, these projections decompose as $\alpha_t I_i + \sigma_t \Pi_{V_i}(\boldsymbol{\epsilon}^{(k)})$ for $i \leq j$. Since projecting $\mathcal{N}(\mathbf{0}, \mathbf{I})$ onto orthogonal subspaces yields independent Gaussians determined entirely by the marginal, the expected gradient of f_j is cumulative-level invariant (for all $k \geq j$):

Proposition 3.3 (Expected Score Reuse)**.** *Let $\boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ marginally for each k, with arbitrary joint distribution across levels. Define $\mathbf{x}_t^{(k)} = \alpha_t \mathcal{I}_k + \sigma_t \boldsymbol{\epsilon}^{(k)}$. Then for all $j \leq k \leq k'$:*

$$\mathbb{E}_{\boldsymbol{\epsilon}^{(k)}}\Big[\nabla_{\mathbf{x}_t^{(k)}} f_j\Big(\mathbf{x}_t^{(k)}\Big)\Big] = \mathbb{E}_{\boldsymbol{\epsilon}^{(k')}}\Big[\nabla_{\mathbf{x}_t^{(k')}} f_j\Big(\mathbf{x}_t^{(k')}\Big)\Big], \tag{4}$$

where $f_j(\mathbf{x}) = \log p_t(I_j^{(t)} \mid I_{<j}^{(t)})$. Proof in Appendix B.

The Bayes-optimal predictor thus satisfies atomic score reuse: each contribution is identical in expectation across all cumulative levels.

Definition 3.4 (Atomic Score)**.** *The* atomic score *at band j is the expected conditional score, cumulative-level invariant across $k \geq j$:*

$$\mathbf{a}^{(j)} \triangleq \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\nabla_{\mathbf{x}_t^{(k)}} \log p_t\Big(I_j^{(t)} \mid I_{<j}^{(t)}\Big)\Big], \quad \textit{for any } k \geq j, \tag{5}$$

where the equality across k is guaranteed by Proposition 3.3.

Corollary 3.1 (Cumulative Score)**.** *Combining Propositions 3.2 and 3.3, the expected score of any cumulative noisy state decomposes as $\mathbb{E}_{\boldsymbol{\epsilon}^{(k)}}[\nabla_{\mathbf{x}_t^{(k)}} \log p_t(\mathbf{x}_t^{(k)})] = \sum_{j=0}^k \mathbf{a}^{(j)}$.*

This enables a single network with K output heads operating on the ordered collection $\mathbf{x}_t \triangleq (\mathbf{x}_t^{(0)}, \ldots, \mathbf{x}_t^{(K-1)})$ of the level-wise noisy states (Def. 3.3). Causal path attention (Sec. 4.1) restricts the j-th head to the causal prefix $\mathbf{x}_t^{(\leq j)} \triangleq (\mathbf{x}_t^{(0)}, \ldots, \mathbf{x}_t^{(j)})$, so it estimates $\mathbf{s}_\theta^{(j)}(\mathbf{x}_t^{(\leq j)}, t) \approx \mathbf{a}^{(j)}$, and $\mathbf{S}_\theta^{(k)}(\mathbf{x}_t^{(\leq k)}, t) = \sum_{j=0}^k \mathbf{s}_\theta^{(j)}(\mathbf{x}_t^{(\leq j)}, t)$ provides a valid cumulative score for any partial image $\mathbf{x}_t^{(k)}$.

Partial Image Estimation. Applying Tweedie's formula Efron (2011) to the k-th perturbation kernel (Def. 3.3) with cumulative score $\mathbf{S}_\theta^{(k)}$ yields $\hat{\mathcal{I}}_k = (\mathbf{x}_t^{(k)} + \sigma_t^2\, \mathbf{S}_\theta^{(k)}(\mathbf{x}_t^{(\leq k)}, t))/\alpha_t$, connecting the score field to pixel-level reconstruction (derivation in Appendix C). This leads to our main result:

Theorem 3.1 (GIEP Cumulative Score Matching)**.** *Let $\mathbf{x}_t^{(k)} = \alpha_t \mathcal{I}_k + \sigma_t \boldsymbol{\epsilon}^{(k)}$ with $\boldsymbol{\epsilon}^{(k)} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$, collected into $\mathbf{x}_t = (\mathbf{x}_t^{(0)}, \ldots, \mathbf{x}_t^{(K-1)})$ with causal prefixes $\mathbf{x}_t^{(\leq j)} = (\mathbf{x}_t^{(0)}, \ldots, \mathbf{x}_t^{(j)})$ (Sec. 4.1). The training objective*

$$\mathcal{L}_{\text{GIEP}} = \mathbb{E}_{t,\,\{\boldsymbol{\epsilon}^{(k)}\}}\left[\sum_{k=0}^{K-1} \lambda_k \left\| \mathcal{I}_k - \frac{\mathbf{x}_t^{(k)} + \sigma_t^2 \sum_{j=0}^k \mathbf{s}_\theta^{(j)}(\mathbf{x}_t^{(\leq j)}, t)}{\alpha_t} \right\|^2\right] \tag{6}$$

decomposes level-wise into a weighted sum of Fisher divergences:

$$\mathcal{L}_{\text{GIEP}} \propto \sum_{k=0}^{K-1} \lambda_k \, \mathbb{E}\left[\left\| \mathbf{S}_\theta^{(k)} - \nabla_{\mathbf{x}_t^{(k)}} \log p_t(\mathbf{x}_t^{(k)}) \right\|^2 \right], \tag{7}$$

since the Fisher divergence at each level depends only on $\mathbf{x}_t^{(k)} \sim \mathcal{N}(\alpha_t \mathcal{I}_k, \sigma_t^2 \mathbf{I})$. *By Proposition 3.3, the minimizer satisfies atomic score reuse in expectation.*

Proof sketch. Substituting $\mathbf{x}_t^{(k)} = \alpha_t \mathcal{I}_k + \sigma_t \boldsymbol{\epsilon}^{(k)}$ into the reconstruction error gives $\|\mathcal{I}_k - \hat{\mathcal{I}}_k\|^2 \propto \|\mathbf{S}_\theta^{(k)} + \boldsymbol{\epsilon}^{(k)}/\sigma_t\|^2$. The denoising score matching equivalence Vincent (2011) identifies $-\boldsymbol{\epsilon}^{(k)}/\sigma_t$ with the transition kernel score, yielding the Fisher divergence. Full derivation in Appendix D. □

This completes the reformulation: the intractable joint $p(I_0, \ldots, I_{K-1})$ is replaced by atomic scores $\{\mathbf{s}_\theta^{(j)}\}$ trained via cumulative score matching, with all K levels supervised in a single forward pass and the causal dependency encoded through cumulative summation and causal masking (Sec. 4.1).

3.3 Enhancement Path Design and Construction

We present three instantiations of the decomposition operator $\mathcal{T}$ (Def. 3.1), each linear, invertible, and coarse-to-fine ordered. *1) FFT bandpass*: radial filtering gives $I_k = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{x}_0) \odot \mathbf{B}_k)$ with masks $\{\mathbf{B}_k\}$ partitioning the frequency plane ($\sum_k \mathbf{B}_k = \mathbf{1}$) from DC ($\mathbf{B}_0$) to the finest textures ($\mathbf{B}_{K-1}$), so $\mathcal{I}_k = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{x}_0) \odot \sum_{j \le k} \mathbf{B}_j)$; Parseval's theorem yields $\langle I_j, I_k \rangle = 0$ for $j \neq k$. *2) Laplacian pyramid*: with Gaussian pyramid $\{G_k\}$ from anti-aliased downsampling, set $I_0 = G_0$ and $I_k = G_k - \text{Up}(G_{k-1})$ for $k \ge 1$, giving $\mathcal{I}_k = G_k$, trading exact orthogonality for spatial-frequency localization. *3) PCA subspaces*: given eigenvectors $\{\mathbf{v}_j\}$ of the data covariance ($\lambda_1 \ge \cdots \ge \lambda_D$) with an index partition $\{\mathcal{G}_k\}$, define $I_k = \sum_{j \in \mathcal{G}_k} \langle \mathbf{x}_0 - \boldsymbol{\mu}, \mathbf{v}_j \rangle \mathbf{v}_j$ and $\mathcal{I}_k = \boldsymbol{\mu} + \sum_{j \le k} I_j$; orthogonality is exact, and leading groups capture global modes (illumination, pose) before fine residuals.

Theoretical guarantees. All three satisfy Def. 3.1: linearity yields the constant Jacobian in Prop. 3.1, invertibility ensures completeness $\mathcal{I}_{K-1} = \mathbf{x}_0$, and the coarse-to-fine ordering aligns with Assumption 3.1. FFT and PCA give exact subspace orthogonality, directly enabling atomic score reuse (Prop. 3.3); the Laplacian pyramid achieves approximate orthogonality with strong spatial localization, offering a practical trade-off.

4 Model Architecture and Training Objective

4.1 Path Score Network Implementation

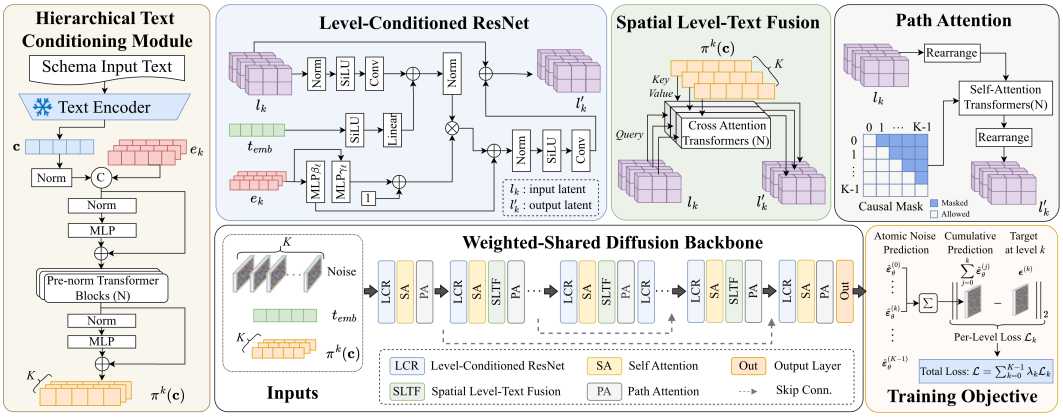

Figure 2: Architecture of GIEP.

Theorem 3.1 requires a single score network $\mathbf{s}_\theta$ that produces K composable atomic scores $\{\mathbf{s}_\theta^{(j)}\}_{j=0}^{K-1}$, any prefix of which sums to a valid cumulative score. As shown in Figure 2, we realize this with a

weight-shared backbone, learnable level embeddings specializing each head to its detail level, and causal path attention enforcing the coarse-to-fine dependency of the autoregressive factorization.

Weight-Shared Backbone. Our denoising backbone follows the SDXL U-Net architecture (Podell et al., 2023) ($\sim$2.6B parameters). We add a path dimension K to the activations while sharing spatial weights across levels. The same construction applies to DiT (Peebles & Xie, 2023) and other backbones.

Level-Conditioned Score Heads. To specialize each head $\mathbf{s}_\theta^{(j)}$ to its detail level (Corollary 3.1) while keeping convolutional weights shared, we inject a learnable level embedding $\mathbf{e}_k = \mathbf{E}[k]$, $\mathbf{E} \in \mathbb{R}^{K \times d_\ell}$, into every ResNet block via affine modulation, $\mathbf{h}_k \leftarrow \big(1 + \gamma_\ell(\mathbf{e}_k)\big) \odot \mathrm{Norm}\big(\mathbf{h}_k + \phi(\mathbf{t}_{\mathrm{emb}})\big) + \beta_\ell(\mathbf{e}_k)$, where $\mathbf{h}_k$ denotes the post-convolution pre-activation feature inside the block, ϕ is a learned timestep projection, and γ_ℓ, β_ℓ are scale/shift MLPs. Since $\mathbf{e}_k$ varies along the path axis but broadcasts spatially, a single forward pass emits all K atomic scores. Unlike cascaded pipelines that train one model per stage Ho et al. (2022); Saharia et al. (2022), the design is agnostic to K.

Causally Masked Path Attention. Convolutions apply identically across the path axis and cannot encode the coarse-to-fine ordering of Eq. 1. We insert *path attention* layers performing causally masked self-attention along K, with mask $\mathbf{M}[i, j] = 0$ for $j \le i$ and $-\infty$ otherwise, so level i attends only to $\{0, \dots, i\}$—i.e. the atomic head at level i reads only the causal prefix $\mathbf{x}_t^{(\le i)}$—and directly realizes $p(I_i \mid I_{<i})$. By forbidding fine levels from influencing coarse ones, the mask is the architectural embodiment of Assumption 3.1 and structurally rules out the high-to-low leakage diagnosed by FreeU Si et al. (2024), in which unmodulated high-frequency features overshadow low-frequency semantics under joint generation. This learned routing replaces inference-time hand-tuned frequency reweighting and is grounded in Proposition 3.2 (ablated in Sec. 5.3).

4.2 Hierarchical Text Conditioning

To align textual guidance with each path level's spectral granularity, we derive level-specific text contexts from an enriched prompt (Sec. 5.1) containing global scene context and fine-grained detail. A lightweight conditioning transformer (Fig. 2, left) converts $\mathbf{c}$ and the learnable level embedding $\mathbf{e}_k$ into level-conditioned context $\mathbf{c}_k^{\text{level}} = \pi^{(k)}(\mathbf{c})$ by first fusing token features with $\mathbf{e}_k$ through an MLP, then passing the fused sequence through N pre-norm transformer blocks that reweight inter-token dependencies for the target frequency band, and finally projecting through a residual MLP. Each level then queries its own $\mathbf{c}_k^{\text{level}}$ in the shared denoising backbone through a cross-attention block we term **Spatial Level-text Fusion** (Fig. 2, top), requiring no additional branches while enabling hierarchically coherent text control. Sec. 5.4 (Fig. 7) verifies this empirically: the same token spreads over global structure at coarse levels and localizes on fine-grained cues at later levels, confirming that level-specific contexts steer the intended granularity at each band.

4.3 Training Objective

Under the standard ε-prediction reparametrization $\mathbf{s}_\theta^{(j)} = -\hat{\boldsymbol{\varepsilon}}_\theta^{(j)}/\sigma_t$, the GIEP loss of Theorem 3.1 reduces to (derivation in Appendix D):

$$\mathcal{L} = \mathbb{E}_{t,\,\{\boldsymbol{\epsilon}^{(k)}\},\mathbf{c}} \left[\sum_{k=0}^{K-1} \lambda_k \left\| \sum_{j=0}^{k} \hat{\boldsymbol{\varepsilon}}_\theta^{(j)}\big(\mathbf{x}_t^{(\le j)}, t, \pi^{(j)}(\mathbf{c})\big) - \boldsymbol{\epsilon}^{(k)} \right\|^2 \right], \tag{8}$$

where $\boldsymbol{\epsilon}^{(k)} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the noise target at level k, $\hat{\boldsymbol{\varepsilon}}_\theta^{(j)}$ is the atomic noise prediction at level j, and $\pi^{(j)}(\mathbf{c})$ the level-conditioned text features (Sec. 4.2). All K levels are supervised in one forward pass, with causal path attention (Sec. 4.1) ensuring each head depends only on coarser levels.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Implementation Details. We train GIEP from scratch on a 500M-image high-resolution corpus, using $K=5$ FFT path levels. All components are jointly optimized with AdamW, and inference uses 75-step DDPM with a uniform classifier-free guidance scale of 7.5 across all five path levels. Appendix E provides the dataset preprocessing, module configurations, and complete optimization settings; Appendix F details prompt construction.

Metrics. We report ManIQA (Yang et al., 2022) and HPSv2 (Wu et al., 2023) for perceptual quality, CLIPScore (Hessel et al., 2021) on COCO-30K (Lin et al., 2014) for text–image alignment, and GenEval (Ghosh et al., 2023) for compositional prompt following. We additionally conduct a blinded human study evaluating overall realism, detail realism, detail saliency, and semantic alignment, and report FID (Heusel et al., 2017) for ablations.

5.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Quantitative Comparison. Table 1 presents a system-level comparison of GIEP with representative T2I methods, grouped into open-source T2I peers (G1), frequency-aware or multi-stage methods (G2), and text-control alternatives (G3). Each method is evaluated as a complete generation pipeline under its recommended inference configuration.

Table 1: System-level comparison of representative T2I methods. CLIPScore and ManIQA use COCO-30K. Best results are **bold**; second-best are underlined.

Method	CLIP Score↑	HPSV2↑	ManIQA ↑
G1: Open-source T2I peers			
SD1.5 (Rombach et al., 2022)	32.85	27.41	70.22
SDXL (Podell et al., 2023)	33.93	30.31	72.11
SD3.5-Large (Esser et al., 2024)	34.06	31.48	73.03
FLUX.1-dev (Labs, 2024)	33.09	32.46	72.45
Z-Image-Turbo (Cai et al., 2025)	33.66	31.90	73.38
FLUX.2-dev (Labs, 2025)	**34.83**	32.59	74.45
G2: Frequency / multi-stage alternatives			
Stable Cascade (Pernias et al., 2023)	33.15	31.48	72.96
FreeU (Si et al., 2024)	33.63	29.06	72.79
SDXL-Refiner (Podell et al., 2023)	33.73	30.71	72.59
G3: Text-control alternatives			
TextCraftor (Li et al., 2024a)	32.23	28.04	71.05
DATE (Na et al., 2025)	32.68	26.05	70.60
GIEP (ours, $K=5$)	34.74	**32.77**	**74.47**

GIEP attains the best HPSv2 and ManIQA in Table 1 and remains competitive on CLIPScore. Against G1, this indicates that explicit coarse-to-fine path decomposition improves perceptual quality without scaling up the backbone. Against G2 and G3, the gains suggest that learnable path decomposition and hierarchical text conditioning outperform fixed frequency heuristics (Si et al., 2024) and text-side enhancement alone. Figure 3 further illustrates that GIEP preserves the prompt semantics while producing more realistic material texture in the sandwich example and clearer whiskers, ear contours, and striped fur detail in the cat example.

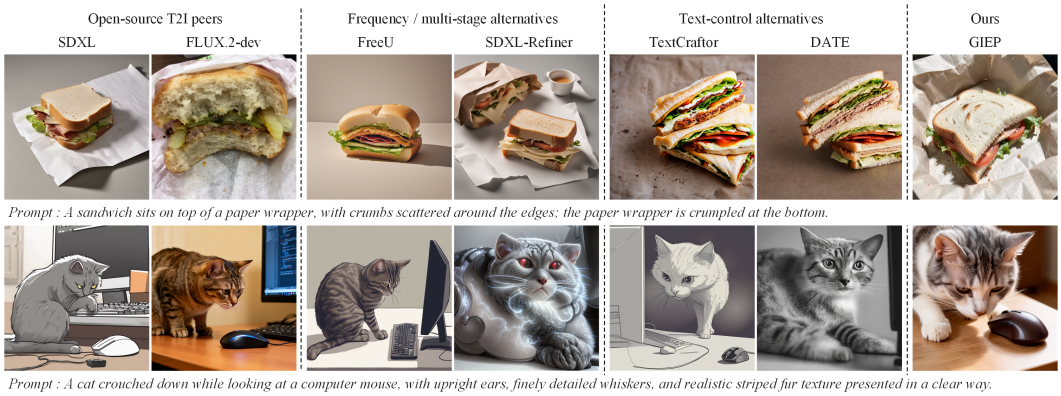

Figure 3: Visual quality comparison. GIEP produces sharper textures and finer details than representative baselines.

Compositional Prompt Following. To assess compositional prompt adherence beyond holistic CLIPScore, we report per-category GenEval results in Table 2. GIEP is highly competitive on GenEval, trailing only the much larger FLUX.2-dev overall while using roughly $8\times$ fewer parameters, and leading or tying on Single Object, Colors, and Position. These gains indicate that the enhancement

Table 2: Results of GenEval Benchmark. Best results are in **bold**, and second-best results are underlined.

Model	Params	Overall	Single object	Two object	Counting	Colors	Position	Attribute
SD1.5 (Rombach et al., 2022)	0.9B	0.43	0.97	0.38	0.35	0.76	0.04	0.06
SDXL (Podell et al., 2023)	2.6B	0.55	0.98	0.74	0.39	0.85	0.15	0.23
SD3 Medium (Esser et al., 2024)	2B	0.62	0.98	0.74	0.63	0.67	0.34	0.36
FLUX.1-dev (Labs, 2024)	12B	0.66	0.98	0.81	0.74	0.79	0.22	0.45
SD3.5-Large (Esser et al., 2024)	8B	0.71	0.98	0.89	0.73	0.83	0.34	0.47
GLM_Image (Z.ai, 2026)	18B	0.74	0.99	0.89	0.69	0.73	0.67	0.49
Z-Image-Turbo (Cai et al., 2025)	6B	0.82	**1**	0.95	0.77	0.89	0.65	0.68
FLUX.2-dev (Labs, 2025)	32B	**0.87**	**1**	**0.97**	0.78	**0.92**	0.75	**0.81**
GIEP (ours, $K=5$)	4B	0.86	**1**	0.95	**0.83**	**0.92**	**0.76**	0.73

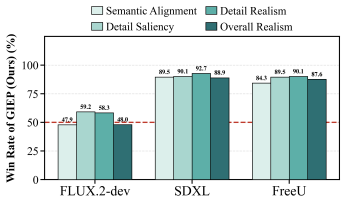

Figure 4: Human preference results: pairwise win rate of Ours against baselines.

path establishes scene layout before refining details, while hierarchical text conditioning binds attributes and spatial cues at the appropriate generation stage.

Human Evaluation. Since automatic metrics undercapture high-frequency detail, we conduct a double-blind pairwise study on 100 diverse prompts. For each prompt, one sample per method is generated with matched seeds and shown in randomized left-right order; 20 raters assess four axes: overall realism, detail realism, detail saliency, and semantic alignment.

Figure 4 shows a consistent trend. Against FLUX.2-dev, GIEP is preferred on detail realism and detail saliency while remaining comparable on overall realism and semantic alignment. Against SDXL and FreeU, GIEP wins on all four axes, with the largest margins again on fine-detail quality. This matches our main claim—the structured enhancement path improves high-frequency realism and saliency by separating coarse scene organization from fine-detail refinement—and Figure 5 provides representative examples. In the tennis-racket example, GIEP produces a more regular and tightly woven string pattern with clearer line thickness and spacing; in the baby example, it recovers sharper wisps of hair and eyebrow texture with more natural skin detail.

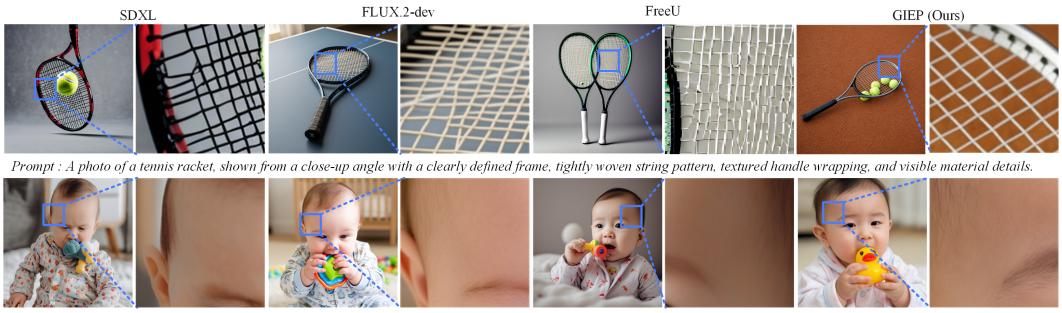

Figure 5: Qualitative examples from the human study.

5.3 Ablation Study

Core Component Ablations. We ablate three design axes: core components, decomposition operator $\mathcal{T}$, and path depth K. For efficiency, all variants are trained on a 1M-image corpus subset and evaluated on COCO-5K. Unless noted, they inherit the hyperparameters of Sec. 5.1 and differ only in the factor under study.

Table 3: Structural ablations on the core components of GIEP, evaluated on COCO-5K.

Setting	CLIP Score↑	FID↓
GIEP ($K=1$)	31.66	25.03
GIEP ($K=5$), w/o causal mask	32.07	23.57
GIEP ($K=5$), w/o hier. text cond.	31.69	23.12
GIEP ($K=5$)	**32.81**	**22.61**

Table 3 isolates each ingredient of GIEP under a parameter-matched protocol. All multi-level variants use a fixed path depth of $K=5$, whereas GIEP ($K=1$) collapses the enhancement path to a single active level while retaining the same 4.0B parameterization. The $K=1$ control yields the weakest CLIPScore and FID, supporting the contribution of coarse-to-fine factorization to the observed gains. Removing the causal mask relaxes the dependency structure in Eq. 1 and degrades both metrics.

This supports the empirical benefit of explicitly enforcing the coarse-to-fine conditional structure rather than relying on unconstrained path attention. Replacing hierarchical text conditioning with shared prompt features likewise degrades both alignment and fidelity, demonstrating its complementarity to the path decomposition. The full model attains the best performance on both metrics.

Path Construction Strategy. Figure 6 compares three $K{=}5$ path constructions with path-collapsed GIEP ($K{=}1$) under the same setup. All outperform the control and show similar convergence, supporting multi-level coarse-to-fine factorization while indicating that the gains are not specific to one choice of $\mathcal{T}$. Although the Laplacian pyramid satisfies the orthogonality condition of Prop. 3.3 only approximately, it remains competitive with orthogonal FFT- and PCA-based constructions, demonstrating GIEP's robustness to approximate orthogonality.

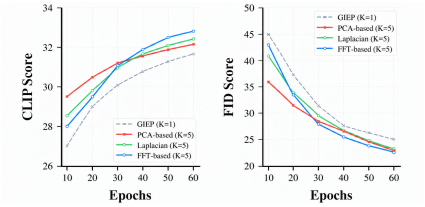

Figure 6: Path-strategy ablation with dual-metric training dynamics, evaluated on COCO-5K.

Number of Path Levels. Table 4 shows that increasing the path depth from $K{=}1$ to $K{=}3$ and $K{=}5$ consistently improves both CLIPScore and FID at $512{\times}512$ and $1024{\times}1024$ resolutions. The gains are more pronounced at $1024{\times}1024$, where richer high-frequency content renders a deeper enhancement path more beneficial. This observation supports our core design: finer path factorization is especially advantageous when high-resolution generation requires both stable global structure and progressively refined detail.

5.4 Mechanistic Analysis

Figure 7 visualizes the level-wise response of hierarchical text conditioning. A clear coarse-to-fine pattern emerges: early levels respond more broadly to global scene semantics, while later levels become increasingly concentrated on local regions such as facial expression, blurred backgrounds, and other fine details. This trend is consistent with our design, where text guidance is distributed across path levels rather than injected uniformly, allowing the enhancement path to integrate global organization with progressive detail refinement.

Table 4: Ablation on path depth K at 512/1024 resolutions, evaluated on COCO-5K.

Res.	K	CLIP Score↑	FID↓
512	1	30.91	22.65
512	3	31.12	21.07
512	5	**31.31**	**20.73**
1024	1	31.66	25.03
1024	3	32.03	24.12
1024	5	**32.81**	**22.61**

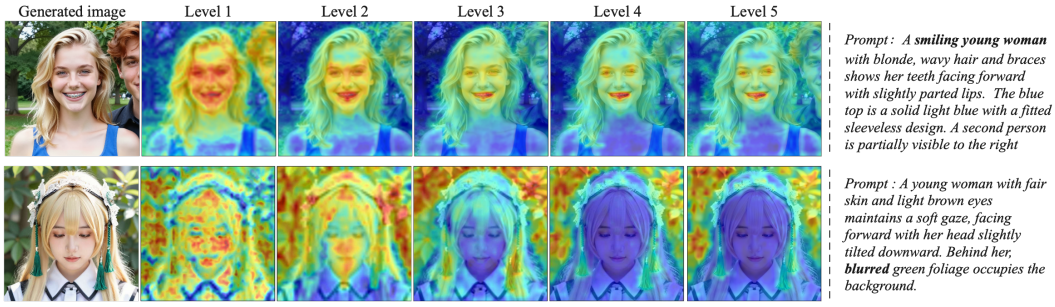

Figure 7: Hierarchical text-conditioning heatmaps. 'Smiling young women' (top) spreads over the face at early levels and sharpens onto the teeth and eyes—the cues that carry the expression—at later levels. 'Blurred' (bottom) covers flat regions of face and background early, then localizes on the foliage. The same word thus acts on different regions as detail resolves.

6 Conclusion

We presented GIEP, a diffusion framework that recasts image generation as sampling from a structured enhancement path over ordered levels of detail, from coarse structure to fine texture. Its theoretical core combines additive score decomposition, atomic score reuse, and cumulative score matching. Together, these allow a single shared network to jointly estimate all per-level scores in one forward pass. Causal path attention and hierarchical text conditioning realize the coarse-to-fine factorization architecturally. Experiments show GIEP matches or surpasses substantially larger baselines, with human raters consistently preferring it on detail realism and saliency. The explicit decoupling of levels of detail along the path axis also indicates clear potential for downstream tasks such as image restoration, super-resolution, and controllable editing, which we leave to future work.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data sets and to clean and reformat datasets. We have not used generative AI tools to assist with translation, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, or interpret results, and support for qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for sourcing and searching for information, editing the paper to improve readability, and formatting references. We have reviewed all AI-assisted work: synthetic data and dataset cleaning or reformatting steps were inspected against the original data and our experimental protocol; literature suggestions were verified against the primary sources, and we discarded any unverifiable or potentially fabricated references; wording edits were accepted only when they preserved the authors' intended claims; and reference formatting was checked against the cited works. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631. License: Apache-2.0.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Peter J Burt and Edward H Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (SIGGRAPH)*, 42(4):1–10, 2023.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image generation. In *International Conference on Learning Representations*, 2024.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11472–11481, 2022.

Prajwal Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorber, Dustin Podell, Tim Dockhorn, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *International Conference on Learning Representations*, 2023.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Florentin Guth, Simon Coste, Valentin De Bortoli, and Stéphane Mallat. Wavelet score-based generative modeling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 478–491, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

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. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

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

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

Emiel Hoogeboom and Tim Salimans. Blurring diffusion models. In *International Conference on Learning Representations*, 2023.

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.

Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *European conference on computer vision*, pp. 196–212. Springer, 2024.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10134, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022.

Sungho Koh, SeungJu Cha, Hyunwoo Oh, Kwanyoung Lee, and Dong-Jin Kim. ScaleDiff: Higher-resolution image synthesis via efficient and model-agnostic diffusion. In *Advances in Neural Information Processing Systems*, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Yanyu Li, Xian Liu, Anil Kag, Ju Hu, Yerlan Idelbayev, Dhritiman Sagar, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. TextCraftor: Your text encoder can be image quality controller. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014. License: annotations and website under Creative Commons Attribution 4.0; images under Flickr Terms of Use.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.

Byeonghu Na, Minsang Park, Gyuwon Sim, Donghyeok Shin, HeeSun Bae, Mina Kang, Se Jung Kwon, Wanmo Kang, and Il-Chul Moon. Diffusion adaptive text embedding for text-to-image diffusion models. *arXiv preprint arXiv:2510.23974*, 2025. Accepted at NeurIPS 2025.

Alexander Quinn Nichol, Prajwal Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pp. 16784–16804, 2022.

Mang Ning, Mingxiao Li, Jianlin Su, Haozhe Jia, Lanmiao Liu, Martin Beneš, Wenshuo Chen, Albert Ali Salah, and Itir Onal Ertugrul. DCTdiff: Intriguing properties of image generative modeling in the DCT space. In *International Conference on Machine Learning*, 2025.

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

Pablo Pernias, Dominic Rampas, Mats L. Richter, Christopher J. Pal, and Marc Aubreville. Wuerstchen: An efficient architecture for large-scale text-to-image diffusion models, 2023.

Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10199–10208, 2023.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831, 2021.

Aditya Ramesh, Prajwal Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Scott Reed, Zeynep Akata, Xinchen Yan, Lakshminarayanan Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pp. 1060–1069, 2016.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *International Conference on Learning Representations*, 2023.

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*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, pp. 234–241, 2015.

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.

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4733–4743. IEEE, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning*, pp. 2256–2264, 2015.

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

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

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*, 2021.

Jiahui Sun, Weining Wang, Mingzhen Sun, Peiyao Wang, Xinxin Zhu, and Jing Liu. W-Edit: A wavelet-based frequency-aware framework for text-driven image editing. In *International Conference on Learning Representations*, 2026.

DeepFloyd Team. Deepfloyd if: A pixel-based text-to-image diffusion model with deep text understanding and high resolution generation. *GitHub repository*, 2023. Available at: https://github.com/deep-floyd/IF.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching for generative modeling. *arXiv preprint arXiv:2302.00482*, 2023.

Unsplash. Unsplash lite dataset. `https://unsplash.com/data`, 2025. Version 1.3.0; License: Unsplash Dataset Terms.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1316–1324, 2018.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal LLMs. In *International Conference on Machine Learning*, 2024.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayers, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.

Z.ai. GLM-Image: Auto-regressive for dense-knowledge and high-fidelity image generation. `https://z.ai/blog/glm-image`, 2026. Official technical blog.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5907–5915, 2017.

Nonghai Zhang and Hao Tang. Text-to-image synthesis: A decade survey. *arXiv preprint arXiv:2411.16164*, 2024.

Yu Zhang, Jialei Zhou, Xinchen Li, Qi Zhang, Zhongwei Wan, Duoqian Miao, Changwei Wang, and Longbing Cao. Enhancing text-to-image diffusion transformer via split-text conditioning. In *Advances in Neural Information Processing Systems*, 2025.

A Proof of Proposition 3.2 (Additive Score Decomposition)

We restate the proposition for convenience.

Proposition 3.2. *Let $\mathcal{T}$ be a linear invertible decomposition (Def. 3.1). Then:*

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) = \sum_{k=0}^{K-1} \nabla_{\mathbf{x}_t} \log p_t\Big(I_k^{(t)} \mid I_{<k}^{(t)}\Big).$$

Proof. We proceed in four steps.

Step 1: Change of Variables. The operator $\mathcal{T}$ is a linear bijection from $\mathbb{R}^D$ to $V_0 \times V_1 \times \cdots \times V_{K-1}$, with $\sum_k \dim(V_k) = D$. Since $\mathcal{T}$ is linear, its Jacobian matrix $\mathbf{J}_{\mathcal{T}}$ is constant (independent of $\mathbf{x}_t$). By the change-of-variables formula for probability densities:

$$p_t(\mathbf{x}_t) = p_t\Big(I_0^{(t)}, I_1^{(t)}, \ldots, I_{K-1}^{(t)}\Big) \cdot \left|\det \mathbf{J}_{\mathcal{T}}\right|^{-1}, \tag{9}$$

where $(I_0^{(t)}, \ldots, I_{K-1}^{(t)}) = \mathcal{T}(\mathbf{x}_t)$.

Step 2: Chain Rule of Probability. We apply the chain rule to the joint density of the components, ordered from the lowest frequency band (k=0) to the highest (k=K−1):

$$p_t\Big(I_0^{(t)}, \ldots, I_{K-1}^{(t)}\Big) = \prod_{k=0}^{K-1} p_t\Big(I_k^{(t)} \mid I_{<k}^{(t)}\Big), \tag{10}$$

where $I_{<k}^{(t)} = \{I_0^{(t)}, \ldots, I_{k-1}^{(t)}\}$ and $I_{<0}^{(t)} = \emptyset$.

Step 3: Log-Likelihood Decomposition. Substituting Eq. (10) into Eq. (9) and applying the natural logarithm:

$$\begin{aligned} \log p_t(\mathbf{x}_t) &= \log \prod_{k=0}^{K-1} p_t\Big(I_k^{(t)} \mid I_{<k}^{(t)}\Big) - \log \left|\det \mathbf{J}_{\mathcal{T}}\right| \\ &= \sum_{k=0}^{K-1} \log p_t\Big(I_k^{(t)} \mid I_{<k}^{(t)}\Big) + C, \end{aligned} \tag{11}$$

where $C = -\log \left|\det \mathbf{J}_{\mathcal{T}}\right|$ is a constant with respect to $\mathbf{x}_t$.

Step 4: Gradient. Applying the gradient operator $\nabla_{\mathbf{x}_t}$ to both sides of Eq. (11):

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) = \sum_{k=0}^{K-1} \nabla_{\mathbf{x}_t} \log p_t\Big(I_k^{(t)} \mid I_{<k}^{(t)}\Big) + \nabla_{\mathbf{x}_t} C. \tag{12}$$

The gradient distributes over the sum by linearity. Since C is constant with respect to $\mathbf{x}_t$, $\nabla_{\mathbf{x}_t} C = 0$, yielding the desired result. □

Remark A.1. Each term $\nabla_{\mathbf{x}_t} \log p_t(I_k^{(t)} \mid I_{<k}^{(t)})$ is the gradient of the k-th conditional log-likelihood with respect to the *full* state $\mathbf{x}_t$, not merely with respect to $I_k^{(t)}$. This is because $\mathbf{x}_t$ determines all components simultaneously through $\mathcal{T}$: perturbing $\mathbf{x}_t$ changes every $I_k^{(t)}$, including both the "variable" $I_k^{(t)}$ and the "conditioning" components $I_{<k}^{(t)}$ in the k-th term. The decomposition is nonetheless exact because the sum of all such gradient contributions recovers the full score.

B Proof of Proposition 3.3 (Expected Score Reuse)

We prove that the expected atomic score is cumulative-level invariant: it does not depend on the cumulative level k at which it is evaluated, for any $k \geq j$.

Proposition 3.3. *Let $\boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ marginally for each k, with arbitrary joint distribution across levels. Then for all $j \leq k \leq k'$:*

$$\mathbb{E}_{\boldsymbol{\epsilon}^{(k)}}\left[\nabla_{\mathbf{x}_t^{(k)}} f_j\left(\mathbf{x}_t^{(k)}\right)\right] = \mathbb{E}_{\boldsymbol{\epsilon}^{(k')}}\left[\nabla_{\mathbf{x}_t^{(k')}} f_j\left(\mathbf{x}_t^{(k')}\right)\right].$$

Proof. Define $f_j(\mathbf{x}) \triangleq \log p_t(I_j^{(t)} \mid I_{<j}^{(t)})$ where $I_i^{(t)} = \Pi_{V_i}(\mathbf{x})$. This function depends on $\mathbf{x}$ solely through the projections $(\Pi_{V_0}(\mathbf{x}), \ldots, \Pi_{V_j}(\mathbf{x}))$: the conditional density of $I_j^{(t)}$ given $I_{<j}^{(t)}$ involves only the first $j{+}1$ components. By the chain rule, $\nabla_{\mathbf{x}} f_j = \sum_{i=0}^{j} \Pi_{V_i}^{\top} \frac{\partial f_j}{\partial I_i^{(t)}}$, so the gradient is a function of these projections alone:

$$\nabla_{\mathbf{x}} f_j(\mathbf{x}) = g_j\big(\Pi_{V_0}(\mathbf{x}), \ldots, \Pi_{V_j}(\mathbf{x})\big) \tag{13}$$

for some measurable function $g_j : V_0 \times \cdots \times V_j \to \mathbb{R}^D$.

Step 1: Decompose into signal and noise. For any level $k \geq j$ and index $i \leq j$:

$$\Pi_{V_i}\left(\mathbf{x}_t^{(k)}\right) = \Pi_{V_i}(\alpha_t \mathcal{I}_k + \sigma_t \boldsymbol{\epsilon}^{(k)}) = \alpha_t I_i + \sigma_t \Pi_{V_i}(\boldsymbol{\epsilon}^{(k)}), \tag{14}$$

where $\Pi_{V_i}(\mathcal{I}_k) = I_i$ for $i \leq k$ follows from the orthogonality of the decomposition $\mathcal{T}$ and the cumulative structure $\mathcal{I}_k = \sum_{m=0}^{k} I_m$.

Step 2: Identify the noise projection distribution. Since $\boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ marginally, and projection is a linear operation, the projections onto orthogonal subspaces are mutually independent Gaussians:

$$\Pi_{V_i}(\boldsymbol{\epsilon}^{(k)}) \sim \mathcal{N}(\mathbf{0}, \Pi_{V_i}), \quad i = 0, \ldots, j, \qquad \text{mutually independent.} \tag{15}$$

This joint distribution is determined entirely by the marginal $\mathcal{N}(\mathbf{0}, \mathbf{I})$ of $\boldsymbol{\epsilon}^{(k)}$—it is identical for every k.

Step 3: Evaluate the expected gradient. The expected gradient at level k is:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}^{(k)}}\left[\nabla f_j(\mathbf{x}_t^{(k)})\right] &= \mathbb{E}\Big[g_j\Big(\alpha_t I_0 + \sigma_t \Pi_{V_0}(\boldsymbol{\epsilon}^{(k)}),\ \ldots,\ \alpha_t I_j + \sigma_t \Pi_{V_j}(\boldsymbol{\epsilon}^{(k)})\Big)\Big] \\ &= \int g_j(\alpha_t I_0 + \sigma_t \mathbf{z}_0,\ \ldots,\ \alpha_t I_j + \sigma_t \mathbf{z}_j) \prod_{i=0}^{j} \phi_{V_i}(\mathbf{z}_i)\, d\mathbf{z}_i,\end{aligned} \tag{16}$$

where ϕ_{V_i} denotes the Gaussian density $\mathcal{N}(\mathbf{0}, \Pi_{V_i})$.

Since the signal terms $\alpha_t I_i$ (for $i \leq j$) are identical for all $k \geq j$, and the noise density $\prod_i \phi_{V_i}$ is independent of k (by Step 2), the integral in Eq. (16) evaluates to the same value for any $k \geq j$. Replacing k by k' changes neither the integrand nor the measure. □

C Derivation of Hierarchical Perturbation Kernels and Partial Tweedie Estimator

C.1 Hierarchical Perturbation Kernels

In the standard diffusion framework, the forward process perturbs a clean signal $\mathbf{x}_0$ via the transition kernel $p(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t;\, \alpha_t \mathbf{x}_0,\, \sigma_t^2 \mathbf{I})$. We extend this by defining a kernel for each cumulative level k.

Let $\mathcal{I}_k = \sum_{j=0}^{k} I_j$ denote the cumulative clean signal up to level k, and let $\mathbf{x}_t^{(k)}$ be the corresponding noisy state. The cumulative perturbation kernel is:

$$\begin{aligned} p(\mathbf{x}_t^{(0)} \mid I_0) &= \mathcal{N}\Big(\mathbf{x}_t^{(0)};\ \alpha_t I_0,\ \sigma_t^2 \mathbf{I}\Big), \\ p(\mathbf{x}_t^{(1)} \mid I_0, I_1) &= \mathcal{N}\Big(\mathbf{x}_t^{(1)};\ \alpha_t (I_0 + I_1),\ \sigma_t^2 \mathbf{I}\Big), \\ &\vdots \\ p\Big(\mathbf{x}_t^{(k)} \mid \{I_j\}_{j=0}^{k}\Big) &= \mathcal{N}\Big(\mathbf{x}_t^{(k)};\ \alpha_t \mathcal{I}_k,\ \sigma_t^2 \mathbf{I}\Big). \end{aligned} \tag{17}$$

Each kernel is a Gaussian centered on the scaled cumulative signal $\alpha_t \mathcal{I}_k$ with isotropic variance σ_t^2. This confirms that $\mathbf{x}_t^{(k)}$ is the correct noisy variate for the partial signal $\mathcal{I}_k$, rather than for the full image $\mathbf{x}_0$. At $k = K-1$, this reduces to the standard kernel $p(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t;\ \alpha_t \mathbf{x}_0,\ \sigma_t^2 \mathbf{I})$ since $\mathcal{I}_{K-1} = \mathbf{x}_0$.

C.2 Partial Tweedie's Estimator

Tweedie's formula (Efron, 2011) states that for $z \sim \mathcal{N}(\mu,\ \sigma^2 \mathbf{I})$, the posterior mean of μ given z is:

$$\mathbb{E}[\mu \mid z] = z + \sigma^2 \nabla_z \log p(z). \tag{18}$$

We apply this to the k-th cumulative kernel (Eq. 17), where $z = \mathbf{x}_t^{(k)}$, $\mu = \alpha_t \mathcal{I}_k$, and $\sigma^2 = \sigma_t^2$:

$$\mathbb{E}\Big[\alpha_t \mathcal{I}_k \mid \mathbf{x}_t^{(k)}\Big] = \mathbf{x}_t^{(k)} + \sigma_t^2 \nabla_{\mathbf{x}_t^{(k)}} \log p_t\Big(\mathbf{x}_t^{(k)}\Big). \tag{19}$$

Dividing both sides by α_t and substituting the cumulative score $\mathbf{S}_\theta^{(k)} = \sum_{j=0}^{k} \mathbf{s}_\theta^{(j)}$ (Corollary 3.1):

$$\hat{\mathcal{I}}_k = \frac{\mathbf{x}_t^{(k)} + \sigma_t^2\, \mathbf{S}_\theta^{(k)}(\mathbf{x}_t^{(\le k)}, t)}{\alpha_t}, \tag{20}$$

where $\mathbf{S}_\theta^{(k)} = \sum_{j=0}^{k} \mathbf{s}_\theta^{(j)}$ is the cumulative score estimator. This yields a valid estimator for the partial image $\mathcal{I}_k$ at every cumulative level k.

C.3 Chain of Reasoning

For the reader's convenience, Table 5 summarizes the complete logical chain from definitions to the final loss.

Table 5: Logical chain of the theoretical development.

Step	Result	Reference
1	Decompose image into components	Def. 3.1
2	Define enhancement path	Def. 3.2
3	Image sampling ≡ component sampling	Prop. 3.1
4	Factor joint into conditionals	Eq. 1
5	Model scores instead of densities	§3.2
6	Global score = sum of conditional scores	Prop. 3.2
7	Hierarchical perturbation kernels	Def. 3.3
8	Expected score reuse across levels	Prop. 3.3
9	Score → partial image via Tweedie	§3.2
10	GIEP cumulative score matching loss	Thm. 3.1

D Fisher Divergence Equivalence of the GIEP Loss

Theorem 3.1. *The GIEP loss $\mathcal{L}_{\mathrm{GIEP}}$ is proportional to the weighted Fisher divergence (Eq. 7).*

Proof. The cumulative noisy state at level k is:

$$\mathbf{x}_t^{(k)} = \alpha_t \mathcal{I}_k + \sigma_t \boldsymbol{\epsilon}^{(k)}, \qquad \boldsymbol{\epsilon}^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \text{ (marginally).} \tag{21}$$

The following derivation proceeds *level by level* and uses only the marginal distribution of $\boldsymbol{\epsilon}^{(k)}$.

Step 1: Expand the reconstruction error. The k-th term of the GIEP loss compares the ground truth partial image $\mathcal{I}_k$ with the Tweedie estimator $\hat{\mathcal{I}}_k$:

$$\begin{aligned} \mathcal{I}_k - \hat{\mathcal{I}}_k &= \mathcal{I}_k - \frac{\mathbf{x}_t^{(k)} + \sigma_t^2 \, \mathbf{S}_\theta^{(k)}}{\alpha_t} \\ &= \frac{\alpha_t \mathcal{I}_k - \mathbf{x}_t^{(k)} - \sigma_t^2 \, \mathbf{S}_\theta^{(k)}}{\alpha_t}. \end{aligned} \tag{22}$$

Step 2: Substitute the forward process. From Eq. (21), $\alpha_t \mathcal{I}_k - \mathbf{x}_t^{(k)} = -\sigma_t \boldsymbol{\epsilon}^{(k)}$. Substituting:

$$\mathcal{I}_k - \hat{\mathcal{I}}_k = \frac{-\sigma_t \boldsymbol{\epsilon}^{(k)} - \sigma_t^2 \, \mathbf{S}_\theta^{(k)}}{\alpha_t} = \frac{-\sigma_t}{\alpha_t} \left(\boldsymbol{\epsilon}^{(k)} + \sigma_t \, \mathbf{S}_\theta^{(k)} \right). \tag{23}$$

Step 3: Identify the true score. For the Gaussian perturbation kernel $\mathbf{x}_t^{(k)} \sim \mathcal{N}(\alpha_t \mathcal{I}_k, \, \sigma_t^2 \mathbf{I})$, the score of the transition kernel is:

$$\nabla_{\mathbf{x}_t^{(k)}} \log p\left(\mathbf{x}_t^{(k)} \mid \mathcal{I}_k \right) = -\frac{\mathbf{x}_t^{(k)} - \alpha_t \mathcal{I}_k}{\sigma_t^2} = -\frac{\boldsymbol{\epsilon}^{(k)}}{\sigma_t}. \tag{24}$$

Step 4: Relate to the Fisher divergence. Taking the squared norm of Eq. (23):

$$\begin{aligned} \left\| \mathcal{I}_k - \hat{\mathcal{I}}_k \right\|^2 &= \frac{\sigma_t^2}{\alpha_t^2} \left\| \boldsymbol{\epsilon}^{(k)} + \sigma_t \, \mathbf{S}_\theta^{(k)} \right\|^2 \\ &= \frac{\sigma_t^4}{\alpha_t^2} \left\| \mathbf{S}_\theta^{(k)} - \left(-\frac{\boldsymbol{\epsilon}^{(k)}}{\sigma_t} \right) \right\|^2. \end{aligned} \tag{25}$$

Using Eq. (24) to identify $-\boldsymbol{\epsilon}^{(k)}/\sigma_t$ as the score of the transition kernel:

$$\left\| \mathcal{I}_k - \hat{\mathcal{I}}_k \right\|^2 = \frac{\sigma_t^4}{\alpha_t^2} \left\| \mathbf{S}_\theta^{(k)} - \nabla_{\mathbf{x}_t^{(k)}} \log p\left(\mathbf{x}_t^{(k)} \mid \mathcal{I}_k \right) \right\|^2. \tag{26}$$

Step 5: Sum over levels and take expectation. Summing over all cumulative levels with weights λ_k and taking the expectation over t and $\{\boldsymbol{\epsilon}^{(k)}\}$:

$$\begin{aligned} \mathcal{L}_{\text{GIEP}} &= \mathbb{E}_{t, \{\boldsymbol{\epsilon}^{(k)}\}} \left[\sum_{k=0}^{K-1} \lambda_k \left\| \mathcal{I}_k - \hat{\mathcal{I}}_k \right\|^2 \right] \\ &= \sum_{k=0}^{K-1} \lambda_k \, \mathbb{E}_{t, \boldsymbol{\epsilon}^{(k)}} \left[\frac{\sigma_t^4}{\alpha_t^2} \left\| \mathbf{S}_\theta^{(k)} - \nabla_{\mathbf{x}_t^{(k)}} \log p\left(\mathbf{x}_t^{(k)} \mid \mathcal{I}_k \right) \right\|^2 \right]. \end{aligned} \tag{27}$$

The sum and expectation interchange because the k-th term depends on $\boldsymbol{\epsilon}^{(k)}$ only through the marginal $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

By the denoising score matching equivalence (Vincent, 2011; Song et al., 2021), minimizing the expected squared distance to the score of the transition kernel $\nabla \log p(\mathbf{x}_t^{(k)} \mid \mathcal{I}_k)$ is equivalent (up to a constant independent of θ) to minimizing the distance to the marginal score $\nabla \log p_t(\mathbf{x}_t^{(k)})$. Therefore:

$$\mathcal{L}_{\text{GIEP}} \propto \sum_{k=0}^{K-1} \lambda_k \, \mathbb{E} \left[\left\| \mathbf{S}_\theta^{(k)} - \nabla_{\mathbf{x}_t^{(k)}} \log p_t\left(\mathbf{x}_t^{(k)} \right) \right\|^2 \right], \tag{28}$$

which is the weighted Fisher divergence across all cumulative levels. The prefactor σ_t^4/α_t^2 is absorbed into the expectation over t (it acts as a time-dependent reweighting of the score matching objective, consistent with the standard diffusion training framework (Karras et al., 2022)). □

E Detailed Experimental Setup

Data and preprocessing. The 500M-image high-resolution training corpus combines our in-house data with public sources, including the Unsplash Lite Dataset (Unsplash, 2025). We retain images with shorter sides above 1024 pixels, remove corrupted and near-duplicate samples, and apply object-centered cropping at the target resolution. Captions are rewritten as enriched prompts covering global scenes and fine-grained details; the complete prompt pipeline is described in Sec. F.

Architecture and optimization. The SDXL U-Net backbone and all GIEP-specific modules are randomly initialized. We use $K{=}5$ FFT path levels. Path-attention modules attached to the denoising blocks and the hierarchical text-conditioning modules use 4-layer transformers. All generative components are jointly optimized with AdamW using a learning rate of 1×10^{-5}, a global batch size of 2048, and uniform path-loss weights $\lambda_k{=}1$. Inference uses 75-step DDPM with a uniform per-level classifier-free guidance scale of 7.5.

F Training Prompt Construction

This section complements Sec. 5.1 by providing concrete examples for two practical cases in the prompt pipeline: source-caption construction when no human-written caption is available, and representative examples of prompt rewriting by the VLM expander.

Source-caption construction without human annotations. When a training image does not come with a human-written caption, we first generate a source caption using a Qwen3-VL-based captioner (Bai et al., 2025), as illustrated in Figure 8. This source caption is then passed to a VLM expander, which rewrites it into the enriched prompt used during GIEP training. The resulting source-caption/structured-prompt pairs are further used to train our in-house prompt-expansion LLM for inference-time prompt generation.

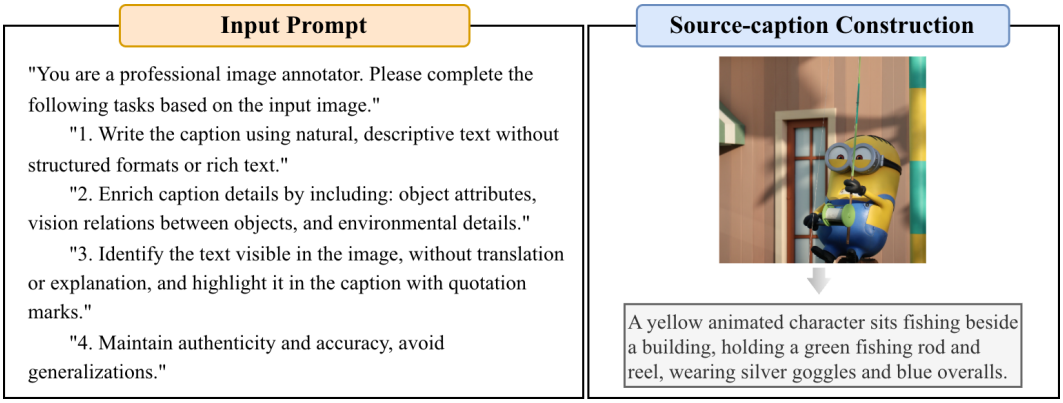

Figure 8: Source-caption generation when no human-written caption is available. The figure shows the prompt given to the Qwen3-VL-based captioner and the resulting source caption generated from the image.

Representative prompt rewriting examples. To support hierarchical conditioning, we feed the image and its source caption to the VLM expander, which rewrites the source caption into a more detailed description that preserves the original semantics while enriching fine-grained visual details, object attributes, and scene context. During training, we generate two admissible rewritten variants per image and randomly sample one at each iteration; the resulting source-caption/rewritten-caption pairs are used to train the in-house prompt-expansion model for inference. Figure 9 illustrates the prompt format given to the VLM expander and representative rewritten captions with enriched details.

Input Prompt

You are a professional "Image Prompt Expander." Your task is to expand, refine, and enhance the input original prompt by combining it with the provided reference image, then output the polished English prompt for image generation models.
Expansion requirements:
1. Do not change the original keywords, and do not delete or replace the core elements in the original sentence.
2.Only add content that can be objectively represented visually and appears in the input image; do not include emotional language, evaluative adjectives, or abstract quality descriptions.
3. Refer to and supplement the following dimensions:
- Subject completeness (full / partial / cropped, etc.)
- Action and pose
- Spatial structure
- Composition
- Lighting direction and shadow relationships
- Focal point
- Overall color palette
- Local color details
- Material and surface texture of subjects and objects (skin, fabric, metal, plastic, etc.)

Rewriting Example



Original Caption

Woodpecker on a tree stump

Woodpecker on a tree stump, close side view, beak inserted into pale cut wood, feet gripping the rough stump. A large bark trunk fills the background, with blurred bare twigs and blue sky. Upper-left sunlight highlights red crown, black-white wings, red belly, bark ridges, and fibrous wood.

Figure 9: Examples of caption rewriting by the VLM expander. The figure illustrates the prompt format provided to the VLM expander, together with representative outputs in which an image and its source caption are converted into rewritten captions with enriched details.

G Computational Cost Analysis

We profile GIEP at 1024×1024 resolution using BF16 and batch size 1. Inference is measured on a single NVIDIA A800 GPU with identical prompts, the 75-step DDPM sampler, and classifier-free guidance scale 7.5 for both path depths. The quality metrics are obtained from the matched 1M-image ablations evaluated on COCO-5K.

Table 6: Computational cost and generation quality at different path depths.

Setting	Training FLOPs/sample	Peak inference memory	Latency/image	CLIPScore ↑	FID ↓
GIEP (K=1)	54.23 TFLOPs	14 GB	20 s	31.66	25.03
GIEP (K=5)	270.57 TFLOPs	33 GB	74 s	32.81	22.61
Cost ratio / quality gain	4.99×	2.36×	3.70×	+1.15	−2.42

Increasing the path depth from K=1 to K=5 improves CLIPScore by 1.15 and reduces FID by 2.42, while increasing training FLOPs by $4.99\times$, peak inference memory by $2.36\times$, and inference latency by $3.70\times$. Cross-level weight sharing keeps the total parameter count at 4.0B for K=5, one eighth of the 32B parameters of FLUX.2-dev, while achieving comparable generation quality. The additional path levels thus improve quality without a proportional increase in model size.

H Human Study Details

We provide additional details for the pairwise human evaluation reported in Sec. 5.

Protocol. 20 raters evaluated 100 prompts in a double-blind pairwise setup. For each prompt, samples generated under matched seeds by the compared methods were displayed side-by-side in

randomized left/right order, and raters selected the preferred sample along four axes: overall realism, detail realism, detail saliency, and semantic alignment.

Instructions to raters. "You will see two anonymized images generated from the same text prompt. The image order is randomized and model names are hidden. Please choose the better image for each criterion."

Rating interface. Figure 10 shows a screenshot of the interface used by raters.

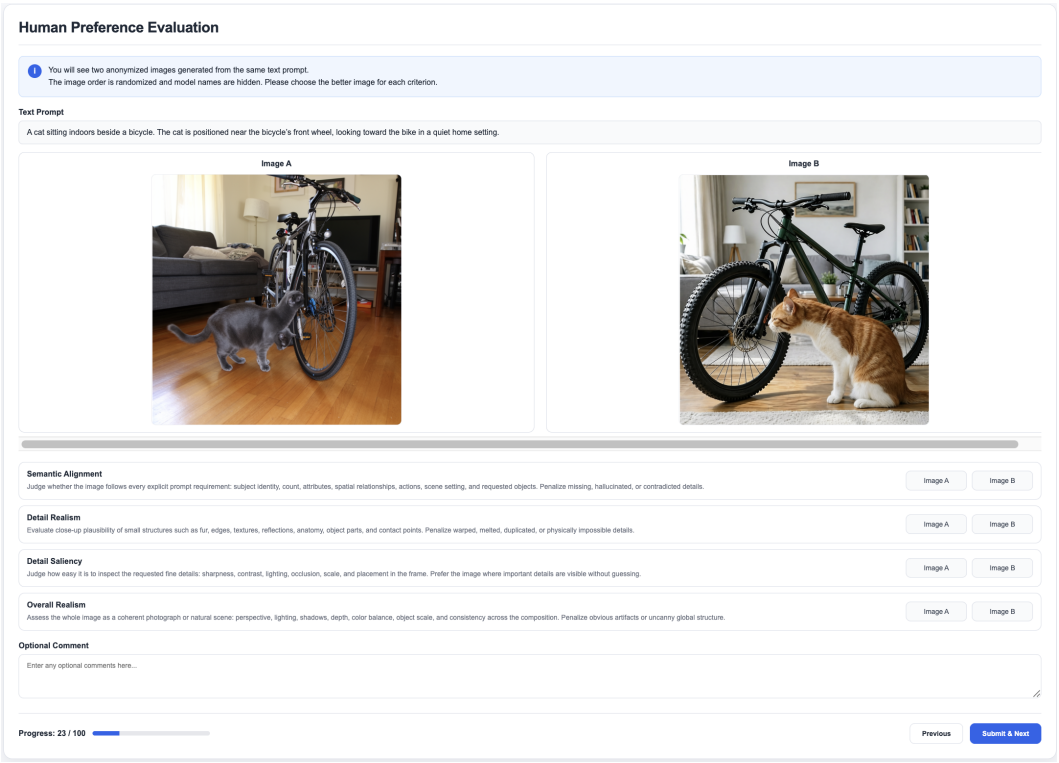

Figure 10: Screenshot of the pairwise rating interface presented to raters.

Compensation. Each rater was compensated USD 2,000 for one week of participation, which exceeded the applicable local minimum wage in the recruitment region.

I Limitations

GIEP has the following limitations. As quantified in Appendix G, GIEP is parameter-efficient rather than compute-efficient: cross-level weight sharing limits model size, but the K-level path axis increases per-forward FLOPs, memory, and latency. The path decomposition is theoretically required to be linear, invertible, and orthogonal (Def. 3.1, Prop. 3.3), but ablations show this is sufficient rather than necessary, with the approximately orthogonal Laplacian pyramid remaining competitive (Sec. 5.3). Finally, the level count K and band weights $\{\lambda_k\}$ are hyperparameters tuned at a fixed compute budget and were not exhaustively swept across model scales.

J Additional Qualitative Comparison

To complement the main-paper visual comparisons, Figure 11 provides additional side-by-side comparisons on challenging prompts. Each row uses the same prompt across methods, covering animal morphology, human-object interaction, multi-subject action, and parent-child pose composition.

These examples further show that GIEP more consistently preserves the requested global layout while refining high-frequency details such as tusks and skin folds, hand-held tools, dog fur, facial structure, and limb configuration.

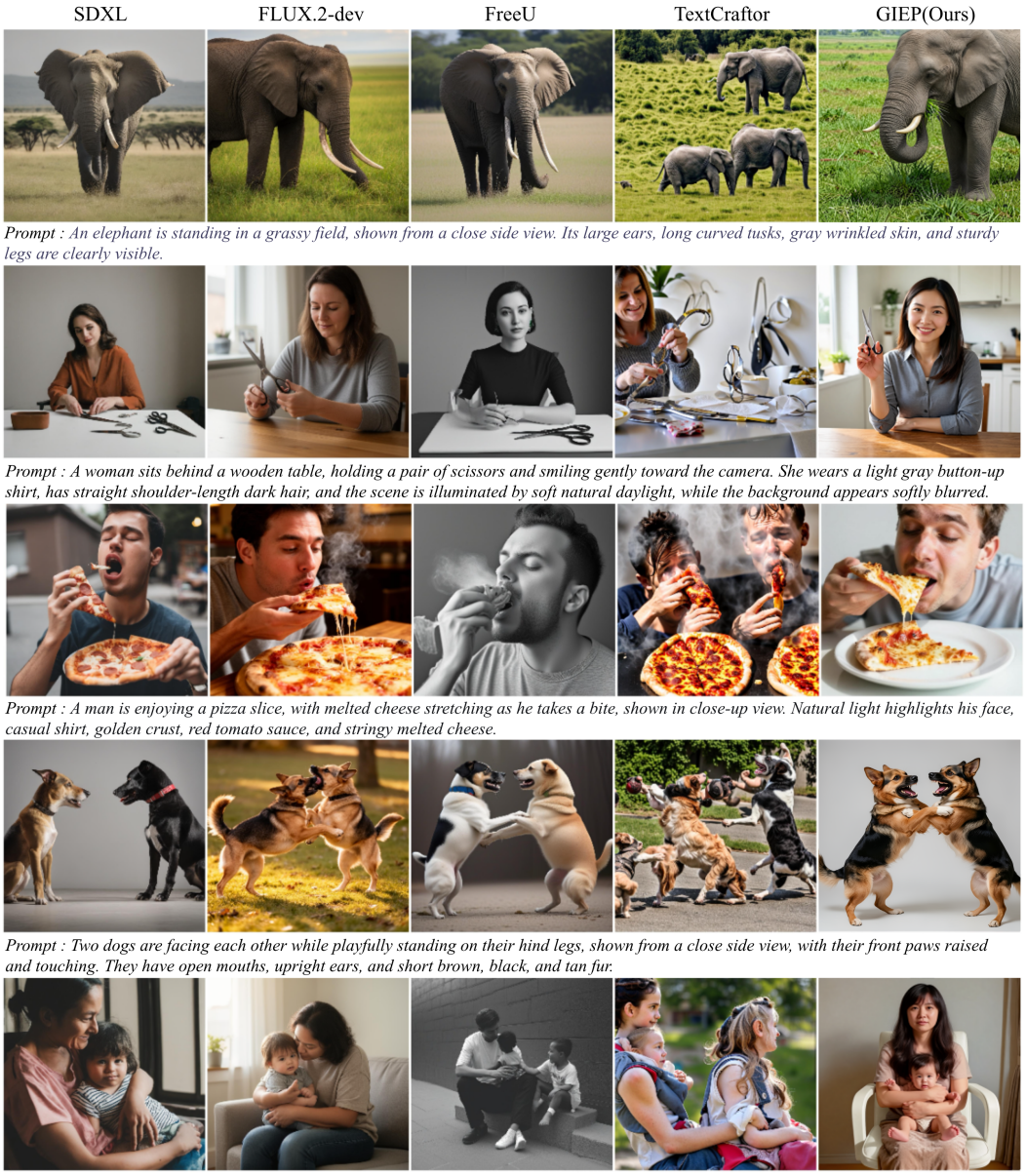

Figure 11: Additional qualitative comparison on diverse and compositional prompts. Each row compares different methods under the same prompt; GIEP (ours) is shown in the rightmost column. The examples highlight global layout preservation and fine-detail synthesis in scenes with object geometry, human-object interaction, and multiple subjects.

To complement the main-paper visual comparisons, we provide additional samples generated by GIEP on diverse prompts. These examples show that GIEP maintains coherent global structure while rendering fine-grained details across portraits, animals, vehicles, outdoor scenes, and indoor scenes.



Figure 12: Additional qualitative results of GIEP on diverse prompts. The samples further illustrate strong global composition and fine-detail synthesis across a wide range of subjects and scenes.