WAVEALIGN: WAVELET-BASED ALIGNMENT FOR TRAINING-FREE HIGH-RESOLUTION IMAGE GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: **Gallery of WAVEALIGN.** WAVEALIGN enables training-free high-resolution generation with pretrained text-to-image flow models, producing coherent and detailed images up to 4K.

ABSTRACT

High-resolution image generation aims to extend pretrained generator beyond their native training resolution while preserving coherent global structure and enabling finer details to emerge. Existing progressive methods transfer information across resolutions, yet parent and child stages differ in spatial scale and sampling phase, leading to mismatches in initialization, reference selection, and structural guidance. To tackle this problem, we formulate progressive high-resolution generation as a cross-resolution correspondence problem involving stochastic, temporal, and structural coherence. Based on this perspective, we introduce WAVEALIGN, a training-free framework that establishes these correspondences in a shared orthonormal wavelet representation. Specifically, WAVEALIGN couples sampled noise and constructs a mixed clean anchor for coherent initialization, derives scale-time conjugacy for phase-matched parent references, and selectively corrects shared coarse structures while progressively releasing guidance for high-resolution detail synthesis. Experiments on FLUX.1-dev demonstrate effective progressive 1K $\rightarrow$ 2K $\rightarrow$ 4K generation, with WAVEALIGN achieving the lowest FID, KID, pFID, and pKID at 4K among the evaluated FLUX-based pipelines.

1 INTRODUCTION

High-resolution text-to-image generation aims to extend pretrained models beyond their native resolutions while preserving global composition and enabling finer details to emerge. Although diffusion and rectified flow models (Rombach et al., 2022; Liu et al., 2022), including Transformer-based architectures (Peebles & Xie, 2023), produce compelling results near their training resolutions, directly increasing the generation resolution often leads to repeated objects, layout inconsistencies,

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.

and degraded textures. Meanwhile, native high-resolution sampling incurs substantial computational and memory costs, while retraining or finetuning at larger resolutions requires additional data and optimization. These limitations motivate *training-free high-resolution generation*, where pretrained generators are extended beyond their native resolutions without parameter updates.

Progressive generation provides a natural strategy for training-free high-resolution synthesis: global composition is first established at a lower resolution and then progressively refined at higher resolutions. Recent methods such as DemoFusion (Du et al., 2024a), I-Max (Du et al., 2024b), and HiFlow (Bu et al., 2025) demonstrate the effectiveness of transferring information across resolution stages. However, transferring information from a lower-resolution *parent* to a higher-resolution *child* is still nontrivial, since the two differ in stochastic realization, spatial scale, and sampling phase. Effective transfer therefore requires establishing explicit correspondences between the two stages: the child should inherit information that remains consistent across resolutions while retaining the freedom to synthesize resolution-specific details. This raises a central question: *How should the child be initialized, which parent phase should guide it, and which structural coordinates should be transferred?*

These questions motivate three complementary forms of cross-resolution alignment: *stochastic, temporal, and structural alignment*. Stochastic alignment concerns how parent and child stages share randomness at initialization while retaining resolution-specific variation, while temporal alignment determines which parent prediction corresponds to the current child sampling phase. And structural alignment controls which coarse information should be transferred across resolutions while leaving child-specific details free to emerge. Orthonormal wavelets provide a unified representation for these alignments through nested coarse subspaces and complementary detail coordinates (Mallat, 1989). This multiscale structure naturally separates information shared across resolutions from components that remain resolution-specific.

Building on this wavelet-based alignment perspective, we introduce WAVEALIGN, a training-free framework for frozen rectified flow models that aligns parent and child stages throughout progressive generation. At initialization, WAVEALIGN constructs a coupled Gaussian noise hierarchy for stochastic alignment and a mixed clean anchor that combines latent-path coarse information with complementary pixel-path components. During sampling, we derive *scale-time conjugacy* under an ideal cross-resolution clean-signal relation to retrieve phase-matched clean predictions from the completed parent cache, providing temporal alignment without additional parent evaluations. These references then guide the child only within shared coarse wavelet subspaces through an orthogonal minimum-intervention correction, while progressive guidance release preserves the flexibility to refine child-specific structures and details. Together, these designs instantiate wavelet-based initialization, phase matching, and structural correction within a unified cross-resolution alignment framework.

To evaluate our proposed WAVEALIGN, we use frozen FLUX.1-dev on the 1,000-pair LAION-Balanced-1K evaluation set for comparison. Among all the compared methods, WAVEALIGN achieves the lowest 4K pFID and pKID, while also obtaining the lowest point estimates for FID and KID, with smaller margins on the image-level metrics. Ablations further demonstrate the effectiveness of noise coupling, phase matching, mixed initialization, and guidance depth, highlighting their respective contributions to image-level and patch-level distribution quality.

Our contributions are threefold:

- **A cross-resolution alignment perspective for training-free high-resolution generation.** We identify parent–child coordination as a central challenge in progressive resolution extrapolation and formulate it through stochastic, temporal, and structural alignment.
- **A unified wavelet-based alignment framework.** We propose WAVEALIGN which uses a common orthonormal wavelet representation to support coupled initial noise, phase-matched parent references through scale-time conjugacy, and minimum-intervention correction in shared coarse coordinates, while preserving resolution-specific details.
- **Controlled empirical validation.** Experiments at 2K and 4K and component ablations evaluate the contributions of initialization, temporal alignment, and selective structural guidance, including the effects of guidance depth.

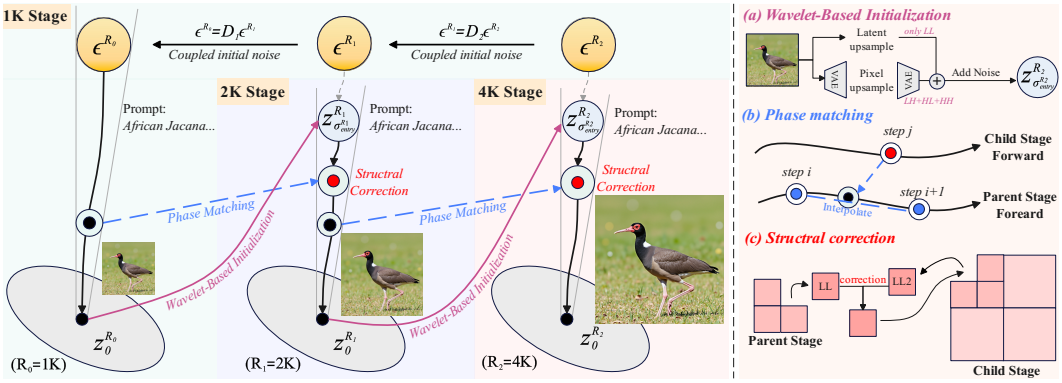

Figure 2: **WaveAlign overview.** The African jacana example at left shows the three resolution stages. Colored arrows correspond to the schematic panels at right: (a) wavelet-based initialization, (b) phase matching, and (c) structural correction.

2 METHOD

WAVEALIGN progressively generates images at $R_0 = 1\text{K}$, $R_1 = 2\text{K}$, and $R_2 = 4\text{K}$ using a frozen rectified flow model, where adjacent lower- and higher-resolution stages are referred to as the *parent* and *child*. Following the cross-resolution alignment perspective, WAVEALIGN coordinates adjacent parent and child stages through three components. Wavelet-based initialization first couples the parent and child through shared noise and a mixed clean anchor. During sampling, phase matching retrieves temporally aligned references from the completed parent trajectory, after which structural correction selectively aligns the shared coarse structure while leaving child-specific coordinates unconstrained. The noise hierarchy is constructed once, and each parent stage is completed before its child begins. The overview of WAVEALIGN is shown in Figure 2.

2.1 PRELIMINARIES

Rectified flow. All states and wavelet operations are defined in latent space. At resolution R, we denote the standard Gaussian initial noise by $\epsilon^R \sim \mathcal{N}(0, I)$. Throughout the paper, *initial noise* refers specifically to this random variable, whose realization is determined once by the noise hierarchy and kept fixed within the corresponding resolution stage. Under the ideal rectified flow interpolation (Liu et al., 2022; Lipman et al., 2022),

$$x_\sigma^R = (1 - \sigma)x_0^R + \sigma\epsilon^R,$$

where x_0^R is a clean sample independent of ϵ^R under the ideal interpolation model, and $\sigma \in [0, 1]$ is the *noise level*, or interpolation coordinate, decreasing toward zero during sampling. Importantly, σ is not the standard deviation of the noisy latent state. We distinguish the ideal interpolation state x_σ^R from the actual noisy latent state z_σ^R evolved by the numerical sampler. In a child stage, sampling instead begins from an initial noisy latent state formed by mixing a clean anchor with ϵ^R at the entry noise level; this state is generally not itself standard Gaussian.

At step i, the frozen model predicts $v_i^R = \mathcal{V}_\theta^{R,i}(z_{\sigma_i^R}^R, \sigma_i^R, c)$ given text c. Its clean estimate is

$$\hat{z}_{0,i}^R = z_{\sigma_i^R}^R - \sigma_i^R v_i^R. \tag{1}$$

The ideal interpolation states x are used only to derive the phase correspondence, whereas the model predictions $\hat{z}_0$ drive guidance along the actual sampled trajectory z.

Orthogonal wavelet transform. A two-dimensional discrete wavelet transform (Mallat, 1989) separates each latent channel into a low-frequency approximation (LL) and three detail subbands (LH, HL, HH), each with half the input height and width. LL captures coarse spatial variation, while the detail bands describe changes along different directions. For an orthogonal transform, the inverse is its transpose, allowing exact reconstruction from all four subbands.

Let D_k extract LL coefficients from R_k to R_{k-1}, with $D_k D_k^\top = I$. Its transpose $U_k = D_k^\top$ lifts LL coefficients back to the original resolution while setting detail coefficients to zero. The orthogonal projector $P_k^{(1)} = U_k D_k$ therefore separates a latent z into

$$z = P_k^{(1)} z + \left(I - P_k^{(1)}\right) z, \tag{2}$$

where the first term contains the reconstructed low-frequency component corresponding to coarse spatial information and the second contains the remaining details. Repeated LL analysis defines progressively coarser subspaces, e.g., LL1 after one level and LL2 after two levels. The filters are specified in Appendix A.1.

When applied to standard Gaussian initial noise, the row-orthonormal LL analysis preserves standard Gaussianity: $D_k \epsilon^{R_k} \sim \mathcal{N}(0, I)$, since $D_k D_k^\top = I$. Hence, the LL coefficients form a valid standard Gaussian initial noise at the lower resolution. Moreover, under the orthonormal wavelet transform, these LL coefficients are independent of the complementary detail coefficients. In the original child latent space, their reconstructed coarse and detail components are $P_k^{(1)} \epsilon^{R_k}$ and $(I - P_k^{(1)})\epsilon^{R_k}$, with covariances $P_k^{(1)}$ and $I - P_k^{(1)}$, respectively. These child-space reconstructions should be distinguished from $D_k \epsilon^{R_k}$, which denotes the lower-resolution LL coefficients themselves and follows $\mathcal{N}(0, I)$.

2.2 Wavelet-Based Initialization

A coherent child initialization must specify both the stochastic realization and the clean information inherited from its parent. We construct these two ingredients in the same orthonormal wavelet coordinates: coupled initial noise establishes cross-resolution stochastic alignment, while a mixed clean anchor combines complementary latent- and pixel-space information for child initialization.

Coupled initial noise. Independent Gaussian draws give valid marginals at each resolution but do not establish sample-wise correspondence between stages. We instead obtain the entire initial-noise hierarchy from one finest-resolution draw by successive LL analysis (Figure 2):

$$\epsilon^{R_2} \sim \mathcal{N}(0, I), \qquad \epsilon^{R_1} = D_2 \epsilon^{R_2}, \qquad \epsilon^{R_0} = D_1 \epsilon^{R_1}. \tag{3}$$

The resulting *coupled initial noise* satisfies $D_k \epsilon^{R_k} = \epsilon^{R_{k-1}}$, so the LL coefficients of the child initial noise exactly reproduce the parent initial noise. The complementary detail coefficients remain independent of these shared LL coefficients, retaining resolution-specific stochastic variation. By the Gaussian-preservation property of row-orthonormal LL analysis, every stage retains a standard Gaussian marginal, $\epsilon^{R_k} \sim \mathcal{N}(0, I)$ for $k \in \{0, 1, 2\}$.

Mixed clean anchor. The clean anchor should preserve coarse information inherited from the parent while supplying finer components required at the child resolution. Latent- and pixel-space upsampling provide complementary but incomplete solutions. ScaleDiff (Koh et al., 2025) reports that latent upsampling can reduce texture oversmoothing but introduce decoding artifacts, whereas image upsampling followed by VAE re-encoding supplies high-frequency latent components for more stable decoding but may promote oversmoothing. We therefore combine the two paths selectively in wavelet coordinates.

Starting from the parent's last cached clean prediction $\bar{z}_0^{R_{k-1}}$ (Figure 2(a)), we obtain $z_{\text{lat}}^{R_k}$ by bicubic interpolation in latent space and $z_{\text{pix}}^{R_k}$ by VAE decoding, bicubic image resizing, and re-encoding (Kingma & Welling, 2013; Rombach et al., 2022). Re-encoding uses the posterior mode without sampling, and both candidates are represented in the same normalized latent space. We then retain the LL1 component from the latent path and the complementary components from the pixel path:

$$z_{0,\text{anc}}^{R_k} = P_k^{(1)} z_{\text{lat}}^{R_k} + \left(I - P_k^{(1)}\right) z_{\text{pix}}^{R_k}. \tag{4}$$

Orthogonality ensures that Eq. (4) preserves the selected components exactly: $P_k^{(1)} z_{0,\text{anc}}^{R_k} = P_k^{(1)} z_{\text{lat}}^{R_k}$ and $(I - P_k^{(1)}) z_{0,\text{anc}}^{R_k} = (I - P_k^{(1)}) z_{\text{pix}}^{R_k}$. Equivalently, the anchor replaces only the pixel

path's LL1 with the latent path's LL1. The anchor is the Euclidean projection of $z_{\text{pix}}^{R_k}$ onto the affine set $\{z : P_k^{(1)} z = P_k^{(1)} z_{\text{lat}}^{R_k}\}$.

Noisy entry state. Following the perturb-then-denoise principle (Meng et al., 2021), we combine the anchor with coupled initial noise to form the child's initial noisy latent state:

$$z^{R_k}_{\sigma^{R_k}_{\text{entry}}} = (1 - \sigma^{R_k}_{\text{entry}}) z^{R_k}_{0,\text{anc}} + \sigma^{R_k}_{\text{entry}} \epsilon^{R_k}, \tag{5}$$

Here, $\sigma^{R_k}_{\text{entry}} \in (0, 1)$ is the child entry noise level, $z^{R_k}_{0,\text{anc}}$ is the anchor from Eq. (4), and ϵ^{R_k} is the initial noise from Eq. (3). We use $\sigma^{R_1}_{\text{entry}} = \sigma^{R_2}_{\text{entry}} = 0.90$. Initialization occurs once per transition; subsequent guidance uses the parent cache. Because the parent is generated from coupled initial noise, the anchor and child initial noise are not required to be independent.

2.3 Phase Matching via Scale-Time Conjugacy

Temporal alignment requires identifying which parent prediction corresponds to the current child sampling phase. Directly matching the same numerical noise level is insufficient, because LL projection scales the clean and noise components differently across resolutions. We therefore derive a resolution-aware phase correspondence by matching their relative interpolation coefficients under ideal cross-resolution consistency.

Under ideal cross-resolution consistency, we assume that the ideal parent clean sample is the child's 2×2 block average. Our normalized LL analysis is twice this average, giving $D_k x_0^{R_k} = \gamma x_0^{R_{k-1}}$ with $\gamma = 2$, whereas coupled initial noise satisfies $D_k \epsilon^{R_k} = \epsilon^{R_{k-1}}$. Applying D_k to the ideal child interpolation therefore yields clean and noise coefficients $\gamma(1 - \sigma)$ and σ. Matching their ratio to that of a parent interpolation, $(1 - \tau)/\tau = \gamma(1 - \sigma)/\sigma$, gives

$$\tau = \Phi_\gamma(\sigma) = \frac{\sigma}{\gamma - (\gamma - 1)\sigma}, \qquad \Phi_2(\sigma) = \frac{\sigma}{2 - \sigma}. \tag{6}$$

We call this ideal-path correspondence *scale-time conjugacy*. For $0 < \sigma < 1$, the mapping becomes a constant translation in logit noise coordinates: $\operatorname{logit} \Phi_\gamma(\sigma) = \operatorname{logit} \sigma - \log \gamma$, where $\operatorname{logit}(s) = \log(s/(1 - s))$. Thus, a change in spatial scale induces a deterministic shift in sampling phase. Although the same functional family appears in timestep shifting (Esser et al., 2024; Du et al., 2024b), here we derive the mapping from cross-resolution signal and noise scaling and use it only to determine parent-cache query phases. The mixed clean anchor is not required to satisfy the ideal clean-consistency assumption, so the phase map serves as a reference-alignment rule rather than requiring the sampled parent and child trajectories to coincide. Appendix A.4 characterizes the ideal correspondence and its approximation error.

Selecting the parent reference. The phase map determines where to query the parent trajectory, while the completed parent cache provides the corresponding clean prediction without additional model evaluations. Before child sampling, we therefore complete the parent stage and cache its clean predictions together with their noise levels. The 1K cache stores the original clean predictions, whereas higher-resolution caches store predictions after structural correction.

At guided child step i, we compute $\tau_i = \Phi_2(\sigma_i^{R_k})$. If τ_i matches a cached parent level, we use the corresponding clean prediction directly; otherwise, we linearly interpolate the neighboring cached predictions $h(a)$ and $h(b)$ for $a < \tau_i < b$, with weights defined in logit noise coordinates:

$$\begin{aligned} w_i &= \frac{\operatorname{logit}(\tau_i) - \operatorname{logit}(a)}{\operatorname{logit}(b) - \operatorname{logit}(a)}, \\ \mathcal{C}_{k-1}(\tau_i) &= (1 - w_i) h(a) + w_i h(b). \end{aligned} \tag{7}$$

We use $\mathcal{C}_{k-1}(\tau_i)$ to construct the phase-aligned structural target in Section 2.4. Under the specified sampling and release schedules, the cache brackets every active query (Appendix B.2), and this retrieval introduces no additional parent evaluations and leaves the child sampling grid unchanged.

2.4 STRUCTURAL CORRECTION

Transferring all parent coordinates can restrict child-resolution refinement. Following projected-flow and low-frequency guidance (Du et al., 2024b; Bu et al., 2025), we limit each correction to a shared coarse wavelet subspace (Figure 2(c)). In our default setting, parent LL1 and child LL2 reach the same spatial grid. This subspace is coarser than the LL1 used for initialization, leaving more detail coordinates outside the correction.

Let $\mathcal{L}_R^{(\ell)}$ apply LL analysis ℓ times, with $\mathcal{L}_R^{(0)} = I$. At guidance depth $\ell_k \geq 1$, define $A_k = \mathcal{L}_{R_k}^{(\ell_k)}$ and $B_k = \mathcal{L}_{R_{k-1}}^{(\ell_k-1)}$, so $A_k = B_k D_k$. The extra child LL analysis accounts for the adjacent resolution difference. We use $\ell_1 = \ell_2 = 2$. The child coordinates and parent target are

$$y_i^{(k)} = A_k \hat{z}_{0,i}^{R_k}, \qquad p_i^{(k)} = \gamma B_k \mathcal{C}_{k-1}\Big(\Phi_\gamma(\sigma_i^{R_k})\Big). \tag{8}$$

The ideal adjacent-scale amplitude factor $\gamma = 2$ accounts for the adjacent-scale clean-amplitude difference induced by the first LL projection and therefore remains unchanged with guidance depth. Rather than replacing the child prediction with the parent reference, we apply a minimum-intervention correction only within the shared structural subspace. Given a correction strength $\alpha_i^{R_k} \in [0, 1]$, we blend the child coarse coordinates toward the target by correcting the clean estimate and adjusting the velocity:

$$\begin{aligned} \Delta_i^{(k)} &= \alpha_i^{R_k} A_k^\top \Big(p_i^{(k)} - y_i^{(k)}\Big), \quad \hat{z}_{0,i}^{R_k,g} = \hat{z}_{0,i}^{R_k} + \Delta_i^{(k)}, \\ v_i^{R_k,g} &= v_i^{R_k} - \Delta_i^{(k)}/\sigma_i^{R_k}, \qquad\qquad \sigma_i^{R_k} > 0. \end{aligned} \tag{9}$$

Because $A_k A_k^\top = I$, the corrected coarse coordinates are $(1 - \alpha_i^{R_k}) y_i^{(k)} + \alpha_i^{R_k} p_i^{(k)}$. Among clean estimates realizing this blend, the update minimizes the Euclidean distance to the current estimate. This explicit correction preserves the complementary coordinates; subsequent model steps can change them. The solver advances using $v_i^{R_k,g}$.

We gradually reduce the guidance weight to zero and continue sampling without further parent correction, allowing later structural revision and detail synthesis. Appendix B gives the release settings, and Algorithm 1 summarizes the procedure.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

Comparison methods and baseline configuration. We use FLUX.1-dev (Black Forest Labs, 2024) as the pretrained generator for primary comparisons. We define a baseline, termed FLUX-Direct, that directly samples at the target resolution without resolution adaptation. We compare WAVEALIGN against training-free high-resolution generation methods, including I-Max (Du et al., 2024b), HiFlow (Bu et al., 2025), ScaleDiff (Koh et al., 2025), DyPE (Issachar et al., 2025), UltraImage (Zhao et al., 2025), SEGA (Rajabi et al., 2026), and HRDiT (Xue et al., 2026), all evaluated without LoRA or parameter updates. We additionally include BSRGAN (Zhang et al., 2021) and OSEDiff (Wu et al., 2024) as post-hoc super-resolution baselines applied to FLUX.1-dev outputs.

Implementation Details Unless otherwise specified, WAVEALIGN uses 30, 14, and 12 velocity evaluations at 1K, 2K, and 4K, respectively. Both child stages enter at $\sigma_{\text{entry}} = 0.90$, use guidance depth $\ell = 2$, and linearly decay the guidance strength to zero by $\sigma = 0.70$. To reduce the cost of early high-resolution sampling, the first six evaluations at 2K and the first five at 4K use overlapping tiles, after which sampling proceeds on the full latent canvas. Table 3 summarizes the stage-wise schedule, and more details are provided in Appendices A.8 and B.

3.2 EVALUATION METRICS

We evaluate each method separately at 2K and 4K on LAION-Balanced-1K, a curated evaluation set of 1,000 prompt–image pairs from RE-LAION-High-Resolution (LAION, 2024), with



Figure 3: **Qualitative comparison of high-resolution generation.** Each column group contains a full image and a magnified crop for one of four representative scenes.

100 pairs from each of ten semantic categories. We report image-level FID (Heusel et al., 2017), KID (Bińkowski et al., 2018), IS (Salimans et al., 2016), and CLIP Score (Radford et al., 2021; Hessel et al., 2021) computed with ViT-L/14, together with their patch-level counterparts pFID, pKID, and pIS. Patch metrics are computed from ten deterministic 256×256 native-pixel crops per image. FID, KID, pFID, and pKID are set-level estimates; IS and pIS are averaged over splits, while CLIP Score is averaged over images. Details of dataset construction, metric preprocessing, and statistical conventions are provided in Appendices D and E.

3.3 Experimental Comparison

3.3.1 Qualitative Results

Figure 3 compares four scenes that emphasize fine texture, local boundaries, repeated structures, and facial appearance. The enlarged crops expose characteristic differences among the methods, *e.g.,* I-Max smooths strawberry seeds and basket weave, HiFlow softens fruit boundaries in the market scene (1st column), BSRGAN reduces definition in snow-covered branches(2nd column), and OSEDiff produces coarser seed pits, more saturated lips, and flatter facial shading(1st and 4th column). In contrast, WaveAlign better preserves separated strawberry seeds and cream ridges, resolves snowy branches and woven strands more clearly, and maintains finer facial details with more gradual shading. The most visible improvements appear in densely repeated local structures, particularly the strawberry surface and basket weave.

3.3.2 Quantitative Results

Table 1 reports the quantitative comparison at 2K and 4K. The advantage of WaveAlign is most evident at 4K, where it achieves the lowest FID, KID, pFID, and pKID among the evaluated methods. In particular, WaveAlign achieves relative reductions of 5.1% in pFID and 26.5% in pKID at 4K. It also achieves the lowest image-level FID and KID, including a 7.0% relative reduction in KID compared with HiFlow. These results indicate that WaveAlign maintains superior quality at the more challenging 4K resolution, with particularly pronounced gains on patch-level metrics.

At 2K, WaveAlign also demonstrates strong and competitive performance across multiple metrics, achieving the lowest pFID and pKID as well as the highest IS and pIS. Together, the 2K and 4K

Table 1: **Quantitative comparison at 2K and 4K.** KID and pKID are scaled by 10^3, and CLIP Score is computed with ViT-L/14. Blue shading indicates 1st and 2nd based on the reported metric values at each resolution.

Method	FID ↓	KID ↓	pFID ↓	pKID ↓	IS ↑	pIS ↑	CLIP ↑
2K Resolution							
FLUX-Direct (Black Forest Labs, 2024)	70.097	7.184	42.516	20.199	19.582	8.413	26.864
I-Max (Du et al., 2024b)	64.593	5.397	29.907	11.267	19.850	11.427	27.864
HiFlow (Bu et al., 2025)	65.243	5.347	26.502	8.215	19.596	11.678	27.779
ScaleDiff (Koh et al., 2025)	65.428	5.651	26.574	6.478	20.070	11.359	27.843
DyPE (Issachar et al., 2025)	67.598	5.322	33.697	10.394	18.892	11.180	27.470
UltraImage (Zhao et al., 2025)	66.267	5.112	49.010	16.641	18.779	11.049	27.673
SEGA (Rajabi et al., 2026)	66.661	4.804	33.071	12.260	19.036	10.713	27.320
HRDiT (Xue et al., 2026)	64.510	5.415	27.032	7.278	19.812	12.052	27.734
FLUX+BSRGAN (Zhang et al., 2021)	65.331	5.751	32.573	12.673	20.195	10.895	27.601
FLUX+OSEDiff (Wu et al., 2024)	65.995	5.766	35.030	14.393	20.102	11.999	27.375
WAVEALIGN (ours)	64.561	5.346	23.776	6.351	20.420	12.172	27.573
4K Resolution							
FLUX-Direct (Black Forest Labs, 2024)	393.131	334.128	348.886	274.881	1.311	1.549	12.467
I-Max (Du et al., 2024b)	65.542	6.033	83.448	56.265	19.164	5.779	27.669
HiFlow (Bu et al., 2025)	65.258	4.741	49.252	24.943	19.341	7.159	27.667
ScaleDiff (Koh et al., 2025)	65.329	5.074	44.989	25.960	19.934	8.058	27.871
UltraImage (Zhao et al., 2025)	65.660	5.662	75.913	48.769	19.590	5.837	27.605
SEGA (Rajabi et al., 2026)	95.641	17.253	145.250	75.275	14.186	7.113	23.860
HRDiT (Xue et al., 2026)	64.721	5.061	54.660	26.700	19.581	7.937	27.715
FLUX+BSRGAN (Zhang et al., 2021)	65.482	5.936	60.094	35.587	20.274	6.824	27.587
FLUX+OSEDiff (Wu et al., 2024)	65.563	5.576	53.059	31.167	20.325	8.361	27.473
WAVEALIGN (ours)	64.364	4.409	42.679	18.324	20.269	7.787	27.473

results highlight the consistent advantage of WAVEALIGN on patch-level distribution quality while maintaining competitive image-level performance, with its distributional gains becoming more pronounced as the target resolution increases.

3.4 ABLATION STUDIES

We conduct controlled ablations by varying one component at a time while keeping the prompts, sampling schedules, evaluation budgets, and guidance-release settings fixed. For each configuration, the 4K stage uses the cached 2K predictions produced by that same configuration. Table 2 reports the primary 4K results, while the corresponding 2K results and extended sweeps are provided in Appendix C. Figure 4 provides a qualitative comparison on a representative golden-eagle example.

Noise coupling. Independent LL Noise replaces only the parent-aligned LL coefficients of the child initial noise with an independently sampled standard Gaussian realization, while retaining the same complementary detail coefficients. This preserves Gaussian marginals but removes sample-wise cross-scale correspondence, increasing 4K FID from 64.364 to 66.294 and pFID from 42.679 to 53.942. In the golden-eagle example, the pose remains recognizable, but the layered chest feathers become a more continuous dark texture, the wing feathers appear less differentiated, and the bark loses some of its pronounced cracks. The degradation persists with LL2 guidance retained, indicating that guidance does not fully compensate for removing noise coupling in this setting.

Phase matching. Same-σ Query, Half-Shift Query, and the full method use $\Phi_1(\sigma) = \sigma$, $\Phi_{\sqrt{2}}(\sigma)$, and $\Phi_2(\sigma)$, corresponding to zero, half, and full conjugate offsets in logit noise level. Child sampling grids remain fixed. Their 4K pFID values are 44.338, 43.164, and 42.679, respectively. In the golden-eagle example, both alternative queries preserve the pose, head, beak, and talons, with subtle differences in the shoulder feather fringe and chest texture. This consistent trend supports resolution-aware phase matching despite subtle qualitative differences.

Table 2: **Primary ablations at** 4K**.** Each variant changes one component relative to Full; LLj denotes guidance depth (j, j). KID and pKID are scaled by 10^3, and bold marks the best value among the configurations shown.

Configuration	FID ↓	KID ↓	pFID ↓	pKID ↓	IS ↑	pIS ↑	CLIP ↑
Independent LL Noise	66.294	5.549	53.942	30.857	19.243	6.955	27.314
Same-σ Query	64.694	4.703	44.338	20.205	19.400	7.781	27.481
Half-Shift Query	64.427	4.564	43.164	19.033	19.279	7.765	27.485
Pixel Bridge	64.958	4.826	49.915	24.751	18.892	7.089	**27.584**
LL1 Guidance	64.412	4.412	44.798	19.477	19.213	7.581	27.431
LL3 Guidance	64.764	4.467	**41.817**	**17.860**	19.506	**7.868**	27.499
WAVEALIGN (Full)	**64.364**	**4.409**	42.679	18.324	**20.269**	7.787	27.473

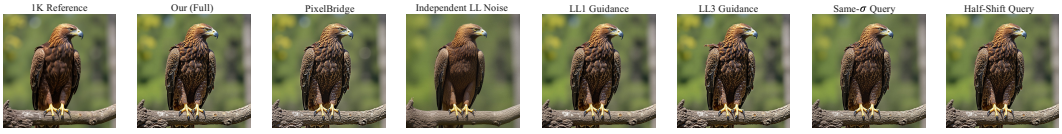

Figure 4: **Qualitative component ablations.** The top-left panel shows the 1K parent reference, and the remaining panels compare WAVEALIGN with representative ablation variants. Differences are most visible in feather organization, shoulder contours, and bark texture. Quantitative results are reported in Table 2.

Mixed initialization. Pixel Bridge uses only pixel-space upsampling followed by VAE re-encoding, while retaining the same subsequent LL2 guidance. Replacing its LL1 coefficients with those from latent interpolation reduces the 4K FID from 64.958 to 64.364 and pFID from 49.915 to 42.679. In the golden-eagle example, Pixel Bridge appears locally sharper around the wing feathers and bark, but this increased sharpness does not translate into better distributional scores. By combining the latent-space LL1 component with the complementary coefficients from the pixel-space candidate, the mixed anchor improves both image-level and patch-level distribution quality while preserving fine-resolution information from the pixel pathway.

Structural correction depth. We vary $(\ell_1, \ell_2) \in \{1, 2, 3\}^2$, fixing the bridge, $\gamma = 2$, and Φ_2. Deeper analysis confines correction to coarser subspaces. Among uniform depths, LL2 yields the lowest 4K FID and KID, whereas LL3 improves patch-level scores and CLIP. Relative to LL2, LL3 increases FID from 64.364 to 64.764 while reducing pFID from 42.679 to 41.817. In the golden-eagle example, LL1 remains visually close to Full, whereas LL3 produces an outward, feather-like protrusion at the left shoulder that appears less anatomically plausible. Considering the stability of structural transfer and the trade-off between image-level and patch-level distribution scores, we use $(2, 2)$ as the default guidance depth.

4 CONCLUSION

We presented WAVEALIGN, a training-free framework that formulates progressive high-resolution generation as a cross-resolution alignment problem. Through a unified wavelet representation, WAVEALIGN coordinates parent and child stages via coupled noise and mixed initialization, scale-time conjugacy for phase-matched parent references, and selective structural correction with progressive release. Experiments with frozen FLUX.1-dev on LAION-Balanced-1K demonstrate effective 1K → 2K → 4K generation, with WAVEALIGN achieving the lowest FID, KID, pFID, and pKID at 4K among the evaluated FLUX-based pipelines.

AI USE STATEMENT

AI tools were used only to assist with programming and to filter candidates for the evaluation dataset. For dataset construction, AI models helped assess candidate captions for semantic relevance and clarity and suggested category assignments. The authors manually checked the suggestions against the images, excluded mismatched or low-quality samples, and verified the final category counts.

REFERENCES

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1737–1752. PMLR, 2023. URL https://proceedings.mlr.press/v202/bar-tal23a.html.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1801.01401.

Black Forest Labs. Announcing black forest labs, 2024. URL https://blackforestlabs.ai/announcing-black-forest-labs/. Official FLUX.1 release; accessed 2026-09-25.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2504.06232.

Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. DemoFusion: Democratising high-resolution image generation with no $$$. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6159–6168. IEEE, 2024a. URL https://arxiv.org/abs/2311.16973.

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. *arXiv preprint arXiv:2410.07536*, 2024b. URL https://arxiv.org/abs/2410.07536.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

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. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

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. URL https://arxiv.org/abs/1706.08500.

Noam Issachar, Guy Yariv, Sagie Benaim, Yossi Adi, Dani Lischinski, and Raanan Fattal. DyPE: Dynamic position extrapolation for ultra high resolution diffusion. *arXiv preprint arXiv:2510.20766*, 2025. URL https://arxiv.org/abs/2510.20766.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013. URL https://arxiv.org/abs/1312.6114.

Sungho Koh, SeungJu Cha, Hyunwoo Oh, Kwanyoung Lee, and Dong-Jin Kim. ScaleDiff: Higher-resolution image synthesis via efficient and model-agnostic diffusion. *arXiv preprint arXiv:2510.25818*, 2025. URL https://arxiv.org/abs/2510.25818.

LAION. RE-LAION-High-Resolution, 2024. URL https://huggingface.co/datasets/laion/relaion-high-resolution. Dataset resource; accessed 2026-09-25.

Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. SyncDiffusion: Coherent montage via synchronized joint diffusions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.05178.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. *arXiv preprint arXiv:2210.02747*, 2022. URL https://arxiv.org/abs/2210.02747.

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. URL https://arxiv.org/abs/2209.03003.

Stéphane G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. doi: 10.1109/34.192463. URL https://doi.org/10.1109/34.192463.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. *arXiv preprint arXiv:2108.01073*, 2021. URL https://arxiv.org/abs/2108.01073.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022. URL https://arxiv.org/abs/2104.11222.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023. URL https://arxiv.org/abs/2212.09748.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Javad Rajabi, Kimia Shaban, Koorosh Roohi, David B. Lindell, and Babak Taati. SEGA: Spectral-energy guided attention for resolution extrapolation in diffusion transformers. *arXiv preprint arXiv:2605.22668*, 2026. URL https://arxiv.org/abs/2605.22668.

Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. UltraPixel: Advancing ultra-high-resolution image synthesis to new peaks. *arXiv preprint arXiv:2407.02158*, 2024. URL https://arxiv.org/abs/2407.02158.

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. URL https://arxiv.org/abs/2112.10752.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1606.03498.

Christoph Schuhmann. Improved aesthetic predictor, 2022. URL https://github.com/christophschuhmann/improved-aesthetic-predictor. CLIP+MLP model sac+logos+ava1-l14-linearMSE; accessed 2026-09-25.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. URL https://arxiv.org/abs/2210.08402.

Luigi Sigillo, Shengfeng He, and Danilo Comminiello. Latent wavelet diffusion for ultra high-resolution image synthesis. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10011421.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10007891.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-Step Effective Diffusion Network for Real-World Image Super-Resolution. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.08177.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. URL https://arxiv.org/abs/2410.10629.

Yu Xue, Haoxuan Qu, Zhuoling Li, Hongbin Xu, Jianxiong Yin, Simon See, Hossein Rahmani, and Jun Liu. HRDiT: Training-free high-resolution image generation with off-the-shelf diffusion transformer models. *arXiv preprint arXiv:2608.07003*, 2026. URL https://arxiv.org/abs/2608.07003.

Tian Ye, Song Fei, and Lei Zhu. UltraFlux: Data-model co-design for high-quality native 4K text-to-image generation across diverse aspect ratios. *arXiv preprint arXiv:2511.18050*, 2025. URL https://arxiv.org/abs/2511.18050.

Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23464–23473. IEEE, 2025. URL https://doi.org/10.1109/CVPR52734.2025.02185.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4791–4800, 2021. URL https://arxiv.org/abs/2103.14006.

Min Zhao, Bokai Yan, Xue Yang, Hongzhou Zhu, Jintao Zhang, Shilong Liu, Chongxuan Li, and Jun Zhu. UltraImage: Rethinking resolution extrapolation in image diffusion transformers. *arXiv preprint arXiv:2512.04504*, 2025. URL https://arxiv.org/abs/2512.04504.

Xingyu Zheng, Xianglong Liu, Yifu Ding, Weilun Feng, Junqing Lin, Jinyang Guo, and Haotong Qin. Multi-resolution flow matching: Training-free diffusion acceleration via staged sampling. *arXiv preprint arXiv:2607.01642*, 2026. URL https://arxiv.org/abs/2607.01642.

A Theoretical Foundations and Design Properties

We first establish the wavelet geometry and stochastic hierarchy, then derive the phase correspondence and projection properties used by mixed initialization and structural correction.

A.1 Wavelet Operators and Recursive Subspaces

We distinguish interpolation states x_σ^R, sampled states z_σ^R, and step-wise clean estimates $\hat{z}_{0,i}^R$. All orthogonal projections refer to the Euclidean inner product on latent coefficients. Adjacent wavelet analysis and synthesis operators satisfy $U_k = D_k^\top$ and $D_k U_k = I$. Recursive analysis uses compatible grids and the same normalization at every depth. Let $\mathcal{L}_R^{(\ell)}$ denote ℓ successive LL analyses, with $\mathcal{L}_R^{(0)} = I$. Let $\mathcal{H}_R$ denote the latent coefficient space at resolution R.

We use the separable two-dimensional orthonormal Haar transform (Mallat, 1989), with one-dimensional low-pass and high-pass filters $h = (1,1)/\sqrt{2}$ and $g = (1,-1)/\sqrt{2}$. Applying these filters along both spatial axes with downsampling by two yields the LL, LH, HL, and HH subbands. In particular, the LL kernel is

$$H_{\mathrm{LL}} = hh^\top = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Thus, for a 2×2 block,

$$a_{\mathrm{LL}} = \frac{a_1 + a_2 + a_3 + a_4}{2} = 2a_{\mathrm{avg}}. \tag{10}$$

Here, $a_{\mathrm{avg}} = (a_1 + a_2 + a_3 + a_4)/4$. The LL analysis therefore returns an orthonormal wavelet coefficient equal to twice the block average. Each analysis row has squared norm $4(1/2)^2 = 1$, and distinct rows act on disjoint blocks, giving $D_k D_k^\top = I$. Consequently, standard Gaussian input noise retains identity covariance after LL analysis. Under the ideal assumption that the parent clean sample equals the child's block average, the clean coefficients satisfy $D_k x_0^{R_k} = 2x_0^{R_{k-1}}$, yielding the adjacent-scale factor $\gamma = 2$ used in the phase map.

Define $P_k^{(1)} = U_k D_k$ and $Q_k = I - P_k^{(1)}$, consistently with the main text. Row orthonormality gives

$$(P_k^{(1)})^\top = P_k^{(1)}, \qquad (P_k^{(1)})^2 = P_k^{(1)}, \qquad Q_k^2 = Q_k, \qquad P_k^{(1)} Q_k = 0. \tag{11}$$

For $\ell \geq 1$, write $B = \mathcal{L}_{R_{k-1}}^{(\ell-1)}$ and $A = BD_k$. Each recursive analysis is row-orthonormal, as follows inductively:

$$AA^\top = BD_k D_k^\top B^\top = BB^\top = I. \tag{12}$$

Define

$$\Pi_R^{(\ell)} = (\mathcal{L}_R^{(\ell)})^\top \mathcal{L}_R^{(\ell)}, \qquad \mathcal{S}_R^{(\ell)} = \mathrm{Range}((\mathcal{L}_R^{(\ell)})^\top). \tag{13}$$

Then $\Pi_R^{(\ell)}$ is the orthogonal projector onto $\mathcal{S}_R^{(\ell)}$. The factorization of deeper analysis through shallower analysis implies

$$\mathcal{S}_R^{(\ell+1)} \subseteq \mathcal{S}_R^{(\ell)}. \tag{14}$$

For $m \geq 1$, define the excluded detail band and its projector by

$$\mathcal{B}_R^{(m)} = \mathcal{S}_R^{(m-1)} \ominus \mathcal{S}_R^{(m)}, \qquad \Pi_{R,\mathrm{band}}^{(m)} = \Pi_R^{(m-1)} - \Pi_R^{(m)}. \tag{15}$$

Consequently,

$$\mathcal{H}_R = \mathcal{S}_R^{(\ell)} \oplus \bigoplus_{m=1}^{\ell} \mathcal{B}_R^{(m)}. \tag{16}$$

At 4K with depth two, $\mathcal{S}_{R_2}^{(2)}$ contains LL-of-LL coordinates. The two excluded bands contain the first-level detail and the intermediate detail within first-level LL. The decomposition organizes latent coefficients by wavelet decomposition level.

A.2 Gaussian Preservation and Independent Initial-Noise Components

For $\epsilon^{R_k} \sim \mathcal{N}(0, I)$, linear Gaussian transformation yields

$$D_k \epsilon^{R_k} \sim \mathcal{N}(0, D_k D_k^\top) = \mathcal{N}(0, I). \tag{17}$$

The reconstructed LL component and complementary detail component of the initial noise have covariances $P_k^{(1)}$ and Q_k, respectively, and

$$\mathrm{Cov}(P_k^{(1)} \epsilon^{R_k}, Q_k \epsilon^{R_k}) = P_k^{(1)} Q_k = 0. \tag{18}$$

Joint Gaussianity and zero cross-covariance establish independence. The complementary component has covariance Q_k in the ambient space and identity covariance when expressed in an orthonormal basis of $\mathrm{Range}(Q_k)$.

Since $P_k^{(1)} \epsilon^{R_k} = U_k \epsilon^{R_{k-1}}$, the child initial noise decomposes as

$$\epsilon^{R_k} = U_k \epsilon^{R_{k-1}} + Q_k \epsilon^{R_k}. \tag{19}$$

Applying further recursive analysis gives

$$\mathcal{L}_{R_k}^{(\ell)} \epsilon^{R_k} = \mathcal{L}_{R_{k-1}}^{(\ell-1)} \epsilon^{R_{k-1}}. \tag{20}$$

The identities characterize the stochastic initialization of the hierarchy. Subsequent generation evolves these coupled variables through the pretrained nonlinear field.

A.3 Clean-Signal Consistency and Amplitude Scaling

The LL normalization above motivates the clean-signal consistency assumption used in the phase derivation.

Assumption 1 (Cross-resolution clean consistency)**.** For the ideal coupled interpolation model, the parent endpoint equals the block average of the child endpoint, so that $D_k x_0^{R_k} = \gamma x_0^{R_{k-1}}$ with $\gamma = 2$.

For general endpoint pairs, define

$$e_k = D_k x_0^{R_k} - \gamma x_0^{R_{k-1}}, \qquad \gamma = 2. \tag{21}$$

The residual quantifies departures from Assumption 1, including those introduced by pixel-space resizing and VAE encoding. Its magnitude determines the approximation error in the normalized interpolation correspondence.

For transport depth ℓ, set $B = \mathcal{L}_{R_{k-1}}^{(\ell-1)}$ and $A = BD_k$. Then

$$A x_0^{R_k} = \gamma B x_0^{R_{k-1}} + B e_k. \tag{22}$$

The compared representations differ by one adjacent resolution transition, giving amplitude factor γ at every depth. Since $BB^\top = I$, the analyzed residual satisfies the absolute bound $\|Be_k\|_2 \leq \|e_k\|_2$.

A.4 Derivation of Scale-Time Conjugacy

For ideal clean consistency, write the child noise level as σ and the matched parent query level as τ_q, with $0 < \sigma, \tau_q < 1$. Matching their relative coefficients requires

$$\frac{1 - \tau_q}{\tau_q} = \gamma \frac{1 - \sigma}{\sigma}. \tag{23}$$

Define the positive amplitude factor $\kappa_\gamma(\sigma) := \gamma - (\gamma - 1)\sigma$.

Proposition 1 (Scale-time correspondence)**.** *Let paired interpolation paths have coupled initial noise $D_k \epsilon^{R_k} = \epsilon^{R_{k-1}}$ and clean residual e_k defined by Eq.* (21)*. Their projected states satisfy Eq.* (26)*. Under Assumption 1, the correspondence is exact after normalization by $\kappa_\gamma(\sigma)$.*

Proof. Projection of the child interpolation path gives

$$D_k x_\sigma^{R_k} = \gamma(1-\sigma)x_0^{R_{k-1}} + \sigma\epsilon^{R_{k-1}} + (1-\sigma)e_k. \tag{24}$$

To represent the first two terms as a scaled parent interpolation state, require

$$\kappa(1-\tau_q) = \gamma(1-\sigma), \qquad \kappa\tau_q = \sigma. \tag{25}$$

Adding the equalities gives $\kappa = \kappa_\gamma(\sigma)$, and solving for τ_q yields $\tau_q = \Phi_\gamma(\sigma)$ in Eq. (6). Consequently,

$$D_k x_\sigma^{R_k} = \kappa_\gamma(\sigma) x_{\Phi_\gamma(\sigma)}^{R_{k-1}} + (1-\sigma)e_k. \tag{26}$$

For $\gamma > 0$, the amplitude factor is positive on $[0,1]$. The normalized discrepancy therefore satisfies

$$\left\| D_k x_\sigma^{R_k}/\kappa_\gamma(\sigma) - x_{\Phi_\gamma(\sigma)}^{R_{k-1}} \right\|_2 = \frac{1-\sigma}{\kappa_\gamma(\sigma)} \left\| e_k \right\|_2. \tag{27}$$

For $e_k = 0$, division by $\kappa_\gamma(\sigma)$ yields exact state correspondence at the matched phase. □

Logit-time form. For $0 < \sigma < 1$,

$$\frac{\Phi_\gamma(\sigma)}{1-\Phi_\gamma(\sigma)} = \frac{\sigma}{\gamma(1-\sigma)}. \tag{28}$$

Taking logarithms gives

$$\operatorname{logit}\Phi_\gamma(\sigma) = \operatorname{logit}\sigma - \log\gamma, \qquad 0 < \sigma < 1. \tag{29}$$

The rational map extends continuously to $\Phi_\gamma(0) = 0$ and $\Phi_\gamma(1) = 1$, while the logit representation is restricted to the open interval.

Recursive coordinates. Applying $B = \mathcal{L}_{R_{k-1}}^{(\ell-1)}$ to Eq. (26) yields

$$Ax_\sigma^{R_k} = \kappa_\gamma(\sigma) B x_{\Phi_\gamma(\sigma)}^{R_{k-1}} + (1-\sigma)Be_k. \tag{30}$$

Thus, the same coefficient-based phase map applies at every supported transport depth. Depth selects spatial coordinates; it does not change the adjacent resolution factor in this interpolation model.

Extension to sampled trajectories. For a sampled trajectory with generated endpoint z_0^R, construct $x_\sigma^R = (1-\sigma)z_0^R + \sigma\epsilon^R$ and define its interpolation deviation $q_\sigma^R = z_\sigma^R - x_\sigma^R$. For compactness, write $H = R_k$, $L = R_{k-1}$, $D = D_k$, $\tau_q = \Phi_\gamma(\sigma)$, and $\kappa = \kappa_\gamma(\sigma)$. With $e = Dz_0^H - \gamma z_0^L$, the sampled-state discrepancy is

$$Dz_\sigma^H - \kappa z_{\tau_q}^L = (1-\sigma)e + Dq_\sigma^H - \kappa q_{\tau_q}^L. \tag{31}$$

Since D has operator norm one, the discrepancy satisfies

$$\left\| Dz_\sigma^H - \kappa z_{\tau_q}^L \right\|_2 \le (1-\sigma)\left\|e\right\|_2 + \left\|q_\sigma^H\right\|_2 + \kappa \left\| q_{\tau_q}^L \right\|_2. \tag{32}$$

This decomposition separates endpoint inconsistency from trajectory deviation. When the right-hand side of the bound is sufficiently small, the phase map approximates cross-scale correspondence between sampled states, but the bound does not directly guarantee agreement between parent and child clean predictions. We therefore use Φ_γ as a parent query rule derived from interpolation coefficient matching; the accuracy of the reference also depends on model prediction and cache interpolation errors.

A.5 Phase-Map Composition and Scheduler Compatibility

The phase maps compose according to

$$\Phi_{\gamma_2}(\Phi_{\gamma_1}(\sigma)) = \Phi_{\gamma_1\gamma_2}(\sigma). \tag{33}$$

This follows by adding the logit translations, with endpoints obtained by continuity. In particular, two adjacent factor-two maps compose to Φ_4.

For a multiplicatively shifted RF scheduler with raw coordinate ρ and positive shift η, define

$$\sigma_\eta(\rho) = \frac{\eta\rho}{1 + (\eta - 1)\rho}. \tag{34}$$

On the open interval, $\operatorname{logit}\sigma_\eta(\rho) = \operatorname{logit}\rho + \log\eta$. Hence,

$$\eta_{R_k} = \gamma\eta_{R_{k-1}} \quad \Longrightarrow \quad \sigma_{\eta_{R_{k-1}}}(\rho) = \Phi_\gamma(\sigma_{\eta_{R_k}}(\rho)). \tag{35}$$

The ratio $1 : 2 : 4$ therefore aligns the three resolutions at a common raw scheduler coordinate under this scheduler family. This sufficient compatibility condition applies to the scheduler family in Eq. (34). Parent-reference interpolation handles distinct discrete raw grids.

A.6 Minimum-Intervention Formulation of Subspace Correction

Fix $z \in \mathbb{R}^n$, $p \in \mathbb{R}^m$, and a row-orthonormal analysis $A \in \mathbb{R}^{m\times n}$ with $AA^\top = I_m$. For finite $\lambda \geq 0$, consider

$$x^\star = \arg\min_x \left\{ \frac{1}{2}\|x - z\|_2^2 + \frac{\lambda}{2}\|Ax - p\|_2^2 \right\}. \tag{36}$$

The Hessian $I + \lambda A^\top A$ is positive definite, so the minimizer is unique. Stationarity yields

$$(I + \lambda A^\top A)x^\star = z + \lambda A^\top p. \tag{37}$$

Since $A^\top A$ projects orthogonally onto $\operatorname{Range}(A^\top) \subseteq \mathbb{R}^n$,

$$(I + \lambda A^\top A)^{-1} = I - \frac{\lambda}{1 + \lambda}A^\top A. \tag{38}$$

Substitution gives

$$x^\star = z + \alpha A^\top(p - Az), \qquad \alpha = \frac{\lambda}{1 + \lambda}. \tag{39}$$

For $r = p - Az$, applying A yields

$$Ax^\star = (1 - \alpha)Az + \alpha p, \qquad p - Ax^\star = (1 - \alpha)r. \tag{40}$$

Moreover, $\left\|A^\top r\right\|_2^2 = r^\top AA^\top r = \|r\|_2^2$, so

$$\|x^\star - z\|_2 = \alpha\|r\|_2. \tag{41}$$

The identities characterize the correction magnitude and structural residual at a fixed sampling step.

Hard projection. Finite λ corresponds to $0 \leq \alpha < 1$. At $\alpha = 1$, the limiting update solves the constrained problem

$$x^\star = \arg\min_{Ax=p} \frac{1}{2}\|x - z\|_2^2 = z + A^\top(p - Az). \tag{42}$$

For every p, $x = A^\top p$ is feasible. The update projects z orthogonally onto the affine set $\{x : Ax = p\}$, defining full guidance at $\alpha = 1$.

A.7 Complementary-Component Invariance under Explicit Subspace Correction

Proposition 2 (Pointwise complementary invariance)**.** *Let $AA^\top = I$, $\Pi = A^\top A$, and $\Delta = \alpha A^\top r$. The update $z \mapsto z + \Delta$ preserves $(I - \Pi)z$ exactly for any fixed z, r, and scalar α.*

Proof. Using $AA^\top = I$,

$$(I - \Pi)\Delta = \alpha(A^\top - A^\top AA^\top)r = 0. \tag{43}$$

Thus, the corrected and uncorrected clean estimates have identical complementary coordinates at the same step:

$$(I - \Pi)(z + \Delta) = (I - \Pi)z. \tag{44}$$

For the nested detail bands excluded by depth ℓ_k,

$$\Pi^{(m)}_{R_k,\text{band}}\Delta_i^{(k)} = 0, \qquad 1 \leq m \leq \ell_k. \tag{45}$$

The same property holds for the instantaneous velocity correction because it is a scalar multiple of $\Delta_i^{(k)}$. □

The invariant in Eq. (44) applies to the current explicit correction. Subsequent cross-subspace interaction is governed by the model field. For a differentiable field, its first-order complementary response to a structural perturbation is $(I - \Pi)J_v\delta z$. This term characterizes the indirect influence of guidance on later detail evolution.

A.8 Wavelet-Mixed Initialization and Projection Identities

The mixed anchor is used only for initialization. The dynamic target remains the phase-matched parent prediction $p_i^{(k)}$ in Eq. (8); we do not replace the parent cache with the mixed anchor or add an anchor-dependent offset to the guidance target.

Complementary lifting paths. Both candidates use the parent's last cached clean prediction $\bar{z}_0^{R_{k-1}}$. The root cache stores uncorrected estimates; child caches store estimates after shared-subspace correction. This terminal prediction need not equal the completed solver endpoint $z_0^{R_{k-1}}$. The two paths are

$$z_{\text{lat}}^{R_k} = \text{Bicubic}_{2\times}^{\text{lat}}\left(\bar{z}_0^{R_{k-1}}\right), \tag{46}$$

$$z_{\text{pix}}^{R_k} = \text{Enc}_{\text{VAE}}\left(\text{Bicubic}_{2\times}^{\text{pix}}\left(\text{Dec}_{\text{VAE}}\left(\bar{z}_0^{R_{k-1}}\right)\right)\right). \tag{47}$$

Orthogonal mixing. The one-level LL projector is

$$P_k^{(1)} = U_k D_k. \tag{48}$$

With $U_k = D_k^\top$ and $D_k U_k = I$, $P_k^{(1)}$ is symmetric and idempotent, hence an orthogonal projector. The anchor in Eq. (4) equivalently equals $z_{\text{pix}}^{R_k} + P_k^{(1)}(z_{\text{lat}}^{R_k} - z_{\text{pix}}^{R_k})$, and satisfies

$$P_k^{(1)} z_{0,\text{anc}}^{R_k} = P_k^{(1)} z_{\text{lat}}^{R_k}, \qquad \left(I - P_k^{(1)}\right) z_{0,\text{anc}}^{R_k} = \left(I - P_k^{(1)}\right) z_{\text{pix}}^{R_k}. \tag{49}$$

These identities separate the contributions of the two paths; they do not imply that either candidate supplies ground-truth high-resolution structure or detail.

Projected entry. The initial noisy latent state in Eq. (5) can be written as a displacement from the coupled standard Gaussian initial noise:

$$z_{\sigma_{\text{entry}}^{R_k}}^{R_k} = \epsilon^{R_k} + (1 - \sigma_{\text{entry}}^{R_k})(z_{0,\text{anc}}^{R_k} - \epsilon^{R_k}). \tag{50}$$

For a child anchor, define its anchor-interpolated noisy latent state at noise level σ by

$$z_{\sigma,\text{anc}}^{R_k} := (1 - \sigma) z_{0,\text{anc}}^{R_k} + \sigma \epsilon^{R_k}. \tag{51}$$

At $\sigma = \sigma_{\text{entry}}^{R_k}$, this equals the initial noisy latent state in Eq. (5); at other noise levels it denotes an anchor interpolation, not the sampled trajectory. Projection gives

$$D_k z_{\sigma,\text{anc}}^{R_k} = (1 - \sigma) D_k z_{0,\text{anc}}^{R_k} + \sigma \epsilon^{R_{k-1}}. \tag{52}$$

The stochastic term is therefore exactly coupled to the parent. More generally, with $A = BD_k$,

$$A z_{\sigma,\text{anc}}^{R_k} = (1 - \sigma) A z_{0,\text{anc}}^{R_k} + \sigma B \epsilon^{R_{k-1}}. \tag{53}$$

These identities establish exact cross-scale consistency of the initial-noise terms for an arbitrary anchor. For the mixed anchor, $D_k P_k^{(1)} = D_k$ implies $D_k z_{0,\text{anc}}^{R_k} = D_k z_{\text{lat}}^{R_k}$, so at entry

$$D_k z_{\sigma_{\text{entry}}^{R_k}}^{R_k} = (1 - \sigma_{\text{entry}}^{R_k}) D_k z_{\text{lat}}^{R_k} + \sigma_{\text{entry}}^{R_k} \epsilon^{R_{k-1}}. \tag{54}$$

The clean term need not satisfy $D_k z_{\text{lat}}^{R_k} = \gamma \bar{z}_0^{R_{k-1}}$; the anchor-dependent clean-consistency residual is therefore retained.

The induced joint distribution couples the anchor to child initial noise through the generated parent's terminal clean prediction. The mixed bridge does not restore the independence between the clean sample and initial noise used during training. The exact guarantee established here is the projected initial-noise identity in Eqs. (52) and (53).

A.9 Nested Initialization and Structural-Correction Subspaces

The bridge and the dynamic correction operate at different depths of the same wavelet hierarchy. The bridge uses the one-level projector $P_k^{(1)}$, whereas the baseline guidance uses $\Pi_k = A_k^\top A_k$ with $\ell_k = 2$. Their subspaces are nested:

$$\Pi_k P_k^{(1)} = P_k^{(1)} \Pi_k = \Pi_k. \tag{55}$$

The anchor can therefore be decomposed into three orthogonal components:

$$z_{0,\mathrm{anc}}^{R_k} = \Pi_k z_{\mathrm{lat}}^{R_k} + \left(P_k^{(1)} - \Pi_k\right) z_{\mathrm{lat}}^{R_k} + \left(I - P_k^{(1)}\right) z_{\mathrm{pix}}^{R_k}. \tag{56}$$

The coarsest component is subsequently regulated by phase-matched parent guidance. The intermediate component within the outer LL subspace and the complementary detail components are not directly modified by this explicit correction:

$$\left(P_k^{(1)} - \Pi_k\right) \Delta_i^{(k)} = 0, \qquad \left(I - P_k^{(1)}\right) \Delta_i^{(k)} = 0. \tag{57}$$

They nevertheless remain free to evolve under the child's pretrained velocity field.

B Inference Procedure and Implementation Details

This section instantiates the design as a discrete progressive sampler. Sampling grids and reference queries are specified before the correction operators, spatial evaluation, and complete inference algorithm.

B.1 Sampling Schedule and Guidance Release

Manual scheduler shifts are $(\eta_{R_0}, \eta_{R_1}, \eta_{R_2}) = (3, 6, 12)$. Each child uses an exact-entry tail uniformly spaced in the unshifted schedule coordinate and ending at $\rho = 1/30$, followed by the scheduler endpoint update. The last 4K model evaluation is at $\sigma = 12/41 \approx 0.292683$. Parent references are queried at $\Phi_2(\sigma)$ through linear interpolation in logit-σ. The discrete grids and cache queries are detailed in Sections B.1.1 and B.2.

Table 3: Frozen baseline schedule. Tiled counts refer to the initial stage-level evaluations; each tiled evaluation aggregates multiple tile-level Transformer calls.

Resolution	Shift	Evaluations	Entry	Tiled evaluations
1024	3	30	Gaussian	0
2048	6	14	0.90	first 6
4096	12	12	0.90	first 5

B.1.1 Discrete Sampling with a Specified Entry Noise Level

Let $N_1 = 14$, $N_2 = 12$, and $\sigma_{\mathrm{entry}}^{R_1} = \sigma_{\mathrm{entry}}^{R_2} = 0.90$. Inverting the shifted scheduler gives each child entry in the unshifted schedule coordinate:

$$\rho_{\mathrm{entry}}^{R_k} = \frac{\sigma_{\mathrm{entry}}^{R_k}}{\eta_{R_k} - (\eta_{R_k} - 1)\sigma_{\mathrm{entry}}^{R_k}}. \tag{58}$$

For the shifts 6 and 12, these values are $3/5$ and $3/7$. Indexing child-local evaluations by $i = 0, \ldots, N_k - 1$, the exact-entry tail in this coordinate is

$$\rho_i^{R_k} = \rho_{\mathrm{entry}}^{R_k} + \frac{i}{N_k - 1}\left(\frac{1}{30} - \rho_{\mathrm{entry}}^{R_k}\right). \tag{59}$$

The model evaluation levels and terminal scheduler node are

$$\sigma_i^{R_k} = \sigma_{\eta_{R_k}}(\rho_i^{R_k}), \qquad \sigma_{N_k}^{R_k} = 0. \tag{60}$$

The positive-noise evaluation grid contains N_k nodes and begins exactly at $\sigma_{\text{entry}}^{R_k}$. Its final evaluation is

$$\sigma_{N_k-1}^{R_k} = \frac{\eta_{R_k}}{29+\eta_{R_k}}, \qquad \sigma_{13}^{R_1} = \frac{6}{35}, \qquad \sigma_{11}^{R_2} = \frac{12}{41}. \tag{61}$$

The endpoint node is reached by the scheduler update following the last evaluation. It is distinct from an additional model evaluation at zero noise. The exact-entry construction determines all guidance weights and cache query phases on the executed child grids.

B.1.2 Guidance Release

Both child stages use the same linear release schedule:

$$\alpha_i^{R_k} = \operatorname{clip}\left(\frac{\sigma_i^{R_k} - 0.70}{0.20}, 0, 1\right), \tag{62}$$

where $\operatorname{clip}(\cdot, 0, 1)$ restricts the weight to $[0, 1]$. Guidance is fully active at the entry noise level 0.90, decreases linearly to zero at 0.70, and remains off thereafter. Sampling continues from the current state using the uncorrected model velocity. Once $\alpha_i^{R_k} = 0$, no parent-cache query is needed.

B.2 Parent Prediction Cache and Logit-Time Interpolation

The root cache records uncorrected clean estimates, while each child cache records the estimates after its shared-subspace correction. The last stored value defines the terminal clean prediction $\bar{z}_0^{R_k}$ used by both lifting paths for the next stage, not the completed solver endpoint $z_0^{R_k}$. The cache remains a record of model clean predictions; it is not replaced by the mixed anchor.

Let $h_j^{R_k}$ be a clean prediction recorded by the parent stage at positive noise level $s_j^{R_k}$, and define $h^{R_k}(s_j^{R_k}) := h_j^{R_k}$ on the recorded nodes. The cache query $\mathcal{C}_k(\tau_q)$ is indexed by resolution stage, and τ_q denotes the parent query noise level. The baseline phase relation is

$$\operatorname{logit}\tau_q = \operatorname{logit}\sigma - \log 2, \qquad \tau_q = \Phi_2(\sigma). \tag{63}$$

For neighboring cached levels $0 < a \le \tau_q \le b < 1$ with $a < b$, define

$$w(\tau_q; a, b) = \frac{\operatorname{logit}\tau_q - \operatorname{logit} a}{\operatorname{logit} b - \operatorname{logit} a}. \tag{64}$$

The reference is interpolated in logit noise level:

$$\mathcal{C}_k(\tau_q) = (1 - w(\tau_q; a, b))h^{R_k}(a) + w(\tau_q; a, b)h^{R_k}(b). \tag{65}$$

An exact-node query returns the stored prediction. This construction matches the scale-dependent clock independently of the raw sampling index.

For active guidance, $0.70 < \sigma \le 0.90$, the query interval is

$$\frac{7}{13} < \tau_q \le \frac{9}{11}. \tag{66}$$

The parent cache must bracket each active query:

$$\Phi_2(\sigma_i^{R_k}) \in [s_{\min}^{R_{k-1}}, s_{\max}^{R_{k-1}}] \quad \text{for } \alpha_i^{R_k} > 0. \tag{67}$$

The 1K and 2K parent grids span $[3/32, 1]$ and $[6/35, 0.90]$, respectively; both bracket every active child query in $(7/13, 9/11]$. Active queries are interior noise levels, so they avoid the singular logit endpoints. Once $\alpha = 0$, the update is independent of the reference.

The completed parent cache supplies guidance without additional parent evaluations. Since B_k is linear, analyzing the cached predictions before interpolation gives the same result as analyzing $\mathcal{C}_{k-1}(\tau_q)$ afterward.

B.3 Mixed-Initialization Implementation

The two lifting paths are defined in Eqs. (46) and (47).

The encoder and decoder use the same pretrained VAE and the backbone's latent normalization convention.

Latent interpolation is bicubic with `align_corners=False`, and wavelet mixing is evaluated in FP32 in the backbone's normalized latent coordinates. For the pixel path, decoded images and resized images are clamped to $[-1, 1]$. VAE re-encoding takes the posterior mode, yielding a deterministic latent representation rather than a random sample. No additional amplitude factor, variance normalization, or random draw is introduced by the latent mixing operation.

B.4 Structural Correction and Solver Updates

B.4.1 Hierarchical Wavelet Implementation

The frozen baseline fixes the same recursive depth at both child stages:

$$\ell_1 = \ell_2 = 2, \qquad A_k = \mathcal{L}^{(2)}_{R_k}, \qquad B_k = \mathcal{L}^{(1)}_{R_{k-1}}. \tag{68}$$

Thus, two child LL analyses and one parent LL analysis yield a transport grid with one quarter of the child latent height and width, without introducing an additional generation stage.

To express the runtime operation, define the outer and inner analyses by $D_{\mathrm{o},k} = D_k$ and $D_{\mathrm{i},k} = B_k$, with adjoints $U_{\mathrm{o},k}$ and $U_{\mathrm{i},k}$. The outer-LL discrepancy is

$$r^{(k)}_{\mathrm{o},i} = 2\mathcal{C}_{k-1}(\Phi_2(\sigma_i^{R_k})) - D_{\mathrm{o},k}\hat{z}^{R_k}_{0,i}. \tag{69}$$

Its inner LL component defines the transported residual, $r^{(k)}_i = D_{\mathrm{i},k} r^{(k)}_{\mathrm{o},i}$. The resulting correction acts only in the two-level LL subspace.

Using $A_k = D_{\mathrm{i},k} D_{\mathrm{o},k}$, the same update has the runtime form

$$\Delta^{(k)}_i = \alpha_i^{R_k} U_{\mathrm{o},k} U_{\mathrm{i},k} D_{\mathrm{i},k} r^{(k)}_{\mathrm{o},i}. \tag{70}$$

B.4.2 Instantiation of Two-Level Low-Frequency Subspace Correction

Both child stages use depth two. At the first transition, let $F = \mathcal{L}^{(1)}_{R_0}$ analyze the 1K parent once. Then

$$A_1 = FD_1, \qquad B_1 = F, \qquad p^{(1)}_i = 2F\mathcal{C}_0(\Phi_2(\sigma_i^{R_1})). \tag{71}$$

The resulting 2K correction is

$$\Delta^{(1)}_i = \alpha_i^{R_1} U_1 F^\top (p^{(1)}_i - FD_1 \hat{z}^{R_1}_{0,i}). \tag{72}$$

The analysis F introduces a coarser coordinate grid without introducing a separate generated stage.

For the 2K $\to$ 4K transition,

$$A_2 = D_1 D_2, \qquad B_2 = D_1, \qquad p^{(2)}_i = 2D_1\mathcal{C}_1(\Phi_2(\sigma_i^{R_2})). \tag{73}$$

Thus,

$$\Delta^{(2)}_i = \alpha_i^{R_2} U_2 U_1 (p^{(2)}_i - D_1 D_2 \hat{z}^{R_2}_{0,i}). \tag{74}$$

The adjacent clean-amplitude factor is two at both stages. Combining these instantiations gives $\Delta^{(k)}_i = \alpha_i^{R_k} U_{\mathrm{o},k} U_{\mathrm{i},k} D_{\mathrm{i},k} r^{(k)}_{\mathrm{o},i}$, exactly the minimum-band operation of the baseline.

Preserved complementary components. Define the child-only and outer-LL detail projectors by

$$Q_{\mathrm{o},k} = I - U_{\mathrm{o},k} D_{\mathrm{o},k}, \qquad Q_{\mathrm{i},k} = I - U_{\mathrm{i},k} D_{\mathrm{i},k}. \tag{75}$$

The two complementary coordinates obey

$$Q_{\mathrm{o},k}\hat{z}^{R_k,g}_{0,i} = Q_{\mathrm{o},k}\hat{z}^{R_k}_{0,i}, \qquad Q_{\mathrm{i},k} D_{\mathrm{o},k}\hat{z}^{R_k,g}_{0,i} = Q_{\mathrm{i},k} D_{\mathrm{o},k}\hat{z}^{R_k}_{0,i}. \tag{76}$$

These equalities preserve both child-only wavelet detail and the inner detail of outer LL in the current prediction. The second component is the parent-resolution detail excluded by the inner LL analysis. The invariants apply at each correction step, by Proposition 2.

B.4.3 Velocity Update and Solver Continuation

From $\hat{z}_{0,i}^{R_k} = z_{\sigma_i^{R_k}}^{R_k} - \sigma_i^{R_k} v_i^{R_k}$, the corrected clean estimate yields

$$v_i^{R_k,g} = \frac{z_{\sigma_i^{R_k}}^{R_k} - \hat{z}_{0,i}^{R_k,g}}{\sigma_i^{R_k}} = v_i^{R_k} - \frac{\Delta_i^{(k)}}{\sigma_i^{R_k}}. \tag{77}$$

The identity holds at positive noise levels. The implementation uses the release schedule in Eq. (62). Once the weight reaches zero, the uncorrected field supplies the remaining updates.

For a decreasing-noise Euler step, the update is

$$z_{\sigma_{i+1}^{R_k}}^{R_k} = z_{\sigma_i^{R_k}}^{R_k} + (\sigma_{i+1}^{R_k} - \sigma_i^{R_k}) v_i^{R_k,g}. \tag{78}$$

This example fixes the sign convention; the corrected field can instead be supplied to the chosen solver. Once $\alpha_i^{R_k} = 0$, the solver uses the uncorrected field at the current state. The continuation starts from the state reached by the preceding guided integration.

B.5 Spatial Evaluation and Attention

B.5.1 Stage-Local Tile and Patchification Schedule

For zero-based child-local step i, define tile-window lengths $T_1 = 6$ and $T_2 = 5$. Tiled evaluation is used for $i < T_k$, and full-canvas evaluation for $i \geq T_k$. The resulting full-canvas counts are eight at 2K and seven at 4K. The root uses 30 full-canvas evaluations.

Even child-local steps use a one-cell spatial padding on each side, and odd steps use the native patchification phase. In tiled execution, padding is applied to the full latent before extracting tiles, and the assembled field is cropped back after blending. In full-canvas execution, even steps use zero padding before the Transformer and cropping after its forward pass. Step parity continues across the tile-to-full transition. Thus, the first full 2K evaluation occurs at padded step $i = 6$, whereas the first full 4K evaluation occurs at native-phase step $i = 5$.

B.5.2 Tile Aggregation and Attention Configuration

Tiles have spatial latent size 128×128 and stride 96, giving 25% overlap along each axis. Each tile uses its global packed-token coordinates. Let $\mathcal{P}_i$ denote the latent phase-adjustment operator, E_t extract tile t, W_t denote its Tukey weights, and $\mathcal{K}_i$ crop back to the native extent. The tile velocity $v_{i,t}^{R_k}$ is evaluated on $E_t \mathcal{P}_i z_{\sigma_i^{R_k}}^{R_k}$ at noise level $\sigma_i^{R_k}$ with condition c. The assembled velocity is

$$v_i^{R_k} = \mathcal{K}_i \left[\frac{\sum_t E_t^\top (W_t \odot v_{i,t}^{R_k})}{\sum_t E_t^\top W_t} \right]. \tag{79}$$

The weights are nonnegative, and the total weight must satisfy $\sum_t (E_t^\top W_t)(u) > 0$ at every spatial position u retained after cropping so that normalized blending is well-defined. The ratio is computed elementwise at these positions, with weights broadcast over channels. Guidance is applied to the clean estimate recovered from the assembled velocity.

For each spatial RoPE axis, the effective NTK factor is one when the input sequence length satisfies $L \leq L_{\mathrm{th}}$, and equals the configured stage factor otherwise. Full-canvas calls use $L_{\mathrm{th}} = 64^2 + 512 = 4608$, with stage factors one at 1K and ten at the child resolutions. Tiled child calls set $L_{\mathrm{th}} = 0$. Since every tile input has $L > 0$, the factor-one fallback is never activated: both tiled and full-canvas child evaluations use spatial NTK factor ten. The text RoPE axis uses factor one throughout. Tile attention uses temperature $1/\sqrt{d}$, where d is the attention-head dimension. Full-canvas child calls retain proportional attention. These stage-dependent settings are included in the evaluation operator $\mathcal{V}_\theta^{R,i}$.

B.6 Inference Algorithm and Runtime Accounting

Algorithm 1 WaveAlign: coarse-to-fine generation with shared-subspace transport

Require: Prompt c; frozen model; resolution hierarchy (R_0, R_1, R_2)
Ensure: Images $I^{R_0}, I^{R_1}, I^{R_2}$

```
Sample finest-scale standard Gaussian initial noise and apply LL analysis using Eq. (3)
Generate z_0^{R_0} from ε^{R_0} and cache intermediate clean predictions in C_0
for k = 1, 2 do
    Read the last cached clean prediction z̄_0^{R_{k-1}} from C_{k-1}
    Construct z_lat^{R_k} and z_pix^{R_k} using Eqs. (46) and (47)
    Mix the candidates into z_{0,anc}^{R_k} using Eq. (4)
    Initialize z with the anchor and coupled initial noise ε^{R_k} using Eq. (5)
    C_k ← [ ]
    for each denoising step at noise level σ > 0 do
        v ← V_θ^{R_k,i}(z, σ, c); h ← z − σv
        Reduce guidance strength α as denoising proceeds; Δ ← 0
        if α > 0 then
            Query parent prediction h_p from C_{k-1} at q = Φ_γ(σ)
            Δ ← α A_k^⊤ (γ B_k h_p − A_k h)
        end if
        Cache (σ, h + Δ) in C_k
        Advance z with corrected velocity v^g = v − Δ/σ
    end for
    Store the completed endpoint z_0^{R_k} ← z
end for
Decode (z_0^{R_0}, z_0^{R_1}, z_0^{R_2}) and return (I^{R_0}, I^{R_1}, I^{R_2})
```

For the reported square-resolution configuration, the complete 1K $\rightarrow$ 2K $\rightarrow$ 4K cascade uses $30 + 14 + 12 = 56$ stage-level velocity evaluations. Each tiled 2K evaluation uses nine tiles. At 4K, padded steps $i = 0, 2, 4$ use 36 tiles each, whereas unpadded steps $i = 1, 3$ use 25. The total number of Transformer forward calls is therefore

$$30 + 6 \times 9 + 8 + (3 \times 36 + 2 \times 25) + 7 = 257.$$

Stopping after 2K requires $30 + 6 \times 9 + 8 = 92$ calls. These calls have different sequence lengths, so their counts alone do not quantify relative computational cost. End-to-end latency also includes overlap aggregation, VAE bridges, and final decoding. Each run records the evaluated noise levels, release weights, parent query levels, and tile/full and phase decisions, supporting reproduction of the executed schedule.

The production path uses a fixed coupled initial-noise hierarchy, one-level wavelet-mixed entry from the last cached parent clean prediction, two-level minimum-band guidance, and the specified attention schedule. Direct pixel-candidate entry is used only in the pixel-bridge ablation. The quadratic objective in Appendix A.6 characterizes the closed-form clean correction; it introduces no additional trajectory optimization or network evaluation.

B.7 Alternative Entry Rule via Tail-Node Retention

The fixed baseline entry $\sigma_{\text{entry}}^{R_k} = 0.90$ is separate from an optional retained-tail entry rule. Let the unshifted schedule contain M positive-noise nodes $\rho_j = 1 - j/M$, $j = 0, \ldots, M-1$, with the zero endpoint used separately to terminate integration. For an integer $1 \le K \le M$, retaining the final K nodes gives the entry $\rho_{M-K} = K/M$, hence

$$\rho_{\text{entry}}^{\text{auto}} = \frac{K}{M}, \qquad \sigma_{\text{entry}}^{\text{auto}} = \frac{\eta K/M}{1 + (\eta - 1)K/M}. \tag{80}$$

For $M = 30$, the configurations $(\eta, K) = (6, 14)$ and $(12, 12)$ yield $\sigma_{\text{entry}}^{\text{auto}} = 0.84$ and $8/9$, respectively. An additional entry-relative lock can couple the guidance lock level to the resolved entry. These settings define ablations and leave the production configuration in Table 3 unchanged.

C Additional Component Ablation Results

C.1 Controlled Protocol and Interpretation

All ablations use the same 1,000 prompts, balanced across ten semantic categories, the same 1,000 real reference images, and a fixed generation seed of 42. The pretrained backbone, output resolutions, stage-level evaluation counts ($30/14/12$), time shifts ($3/6/12$), and tile-to-full evaluation settings are held fixed. Both child stages enter at $\sigma = 0.90$, with the common release schedule

$$\alpha(\sigma) = \mathrm{clip}\left(\frac{\sigma - 0.70}{0.20}, 0, 1\right). \tag{81}$$

Full uses orthogonally coupled initial noise, the Φ_2 query, one-level direct wavelet mixing, and LL2 guidance at both child stages. Unless explicitly varied, these settings remain fixed.

The intervention is applied at both child stages for the primary comparisons. Each $4\mathrm{K}$ stage uses the $2\mathrm{K}$ trajectory generated by the same configuration. Its result therefore includes effects inherited from the preceding stage; endpoint metrics alone do not identify a mechanism of error propagation. In the stage-wise depth sweep, varying ℓ_2 at fixed ℓ_1 holds the $2\mathrm{K}$ parent trajectory fixed.

All results use the data alignment, metric definitions, and reporting protocol in Appendix E.

Each configuration is evaluated under one fixed generation seed. The paired bootstrap in Appendix E.4 quantifies variation over the evaluated prompt–image units, not across repeated generation seeds. Identical $2\mathrm{K}$ outputs are reused where only the $4\mathrm{K}$ setting changes, and are not treated as independent replicates.

C.2 Noise Coupling and Phase Matching

Independent LL Noise replaces only the parent-aligned LL coefficients of the child initial noise with an independently sampled standard Gaussian realization, while retaining the same complementary detail coefficients. It therefore removes stochastic correspondence while preserving Gaussian marginals and retaining the same structural-guidance rule. Since initial-noise correspondence is a premise of the ideal conjugacy, this experiment tests the contribution of that premise within the complete pipeline.

The temporal variants change only the parent-cache query. With $\beta \in \{0, \frac{1}{2}, 1\}$, their queries can be written as

$$\mathrm{logit}\, q_\beta(\sigma) = \mathrm{logit}\, \sigma - \beta \log 2. \tag{82}$$

Thus, $\beta = 0$, $\frac{1}{2}$, and 1 correspond to Same-σ Query, Half-Shift Query, and Full. Half-Shift Query halves the offset in logit noise level; it neither halves the child noise level nor changes the child time shifts, number of evaluations, or sampling nodes. The clean-coordinate amplitude factor remains $\gamma = 2$ in all temporal variants.

Table 4 gives the corresponding $2\mathrm{K}$ results, and Table 5 reports the auxiliary metrics at both resolutions. Full has lower FID, KID, pFID, and pKID point estimates than both temporal alternatives at each resolution. This ordering does not extend to every metric: Same-σ Query and Half-Shift Query have slightly higher CLIP Scores, and Same-σ Query has lower Repeat and Spectral values. These comparisons isolate the distributional effects of stochastic and temporal correspondence across the reported criteria.

Table 4: Primary ablations at 2K under the same protocol as Table 2. KID and pKID are multiplied by 10^3; CLIP denotes the CLIP Score computed with ViT-L/14. Bold indicates the best value among the configurations shown, using unrounded scores.

Configuration	FID ↓	KID ↓	pFID ↓	pKID ↓	IS ↑	pIS ↑	CLIP ↑
Independent LL Noise	65.013	5.894	26.478	9.536	19.950	11.791	27.519
Same-σ Query	64.617	5.574	23.908	6.559	19.792	12.141	27.586
Half-Shift Query	64.616	5.511	23.836	6.457	19.779	12.174	27.576
Pixel Bridge	64.840	5.845	25.771	8.288	19.662	11.560	**27.663**
LL1 Guidance	**64.474**	**5.316**	24.247	6.499	19.777	12.211	27.566
LL3 Guidance	64.615	5.539	**23.621**	**6.180**	19.789	**12.235**	27.590
WAVEALIGN (Full)	64.561	5.346	23.776	6.351	**20.420**	12.172	27.573

Table 5: Auxiliary metrics for the primary ablations. CLIP-B denotes the CLIP Score computed with ViT-B/16. Repeat and Spectral are reported in the decreasing direction.

Configuration	CLIP-B ↑	Aesth. ↑	Repeat ↓	Spectral ↓
		2K		
Independent LL Noise	32.486	5.472	0.1335	0.0653
Same-σ Query	32.528	5.495	0.1216	0.0585
Half-Shift Query	32.517	5.493	0.1229	0.0589
Pixel Bridge	32.549	5.466	0.1267	0.0631
LL1 Guidance	32.487	5.493	0.1272	0.0618
LL3 Guidance	32.535	5.489	0.1229	0.0590
WAVEALIGN (Full)	32.507	5.492	0.1235	0.0591
		4K		
Independent LL Noise	32.164	5.539	0.1237	0.0598
Same-σ Query	32.379	5.561	0.0949	0.0453
Half-Shift Query	32.393	5.556	0.0968	0.0461
Pixel Bridge	32.432	5.526	0.0982	0.0488
LL1 Guidance	32.339	5.555	0.1045	0.0520
LL3 Guidance	32.388	5.547	0.0956	0.0459
WAVEALIGN (Full)	32.352	5.555	0.0973	0.0464

C.3 MIXED INITIALIZATION

Bridge depth and guidance depth are distinct parameters. Let $P_k^{(b)} = (\mathcal{L}_{R_k}^{(b)})^\top \mathcal{L}_{R_k}^{(b)}$ denote the b-level orthogonal low-frequency (LL) subband projector in the child latent space. The bridge variants use

$$z_{0,\mathrm{anc}}^{R_k} = z_{\mathrm{pix}}^{R_k} + \lambda P_k^{(b_k)} \left(z_{\mathrm{lat}}^{R_k} - z_{\mathrm{pix}}^{R_k} \right). \tag{83}$$

Here $z_{\mathrm{lat}}^{R_k}$ is obtained by latent bicubic interpolation, whereas $z_{\mathrm{pix}}^{R_k}$ is obtained by VAE decoding, pixel-space bicubic resizing, and re-encoding. Direct mixing uses $\lambda = 1$, and Pixel Bridge uses $\lambda = 0$. Bridge decomposition depths $(b_1, b_2) = (j, j)$ apply j levels at both child stages, whereas the one-level-to-two-level bridge uses $(b_1, b_2) = (1, 2)$.

All bridge variants retain $(\ell_1, \ell_2) = (2, 2)$ guidance, coupled initial noise, the full conjugate parent query, and the common entry noise level. The bridge and guidance studies are separate controlled sweeps, not a factorial evaluation of all depth combinations.

Tables 6 and 7 report the primary and auxiliary results. Full improves all four distributional distances over Pixel Bridge at both resolutions, although Pixel Bridge has higher CLIP Scores.

Changing the bridge depth also produces a trade-off. Direct two-level mixing reduces 4K pFID to 42.429 and pKID to 17.428 but increases FID to 64.710. Using the one-level-to-two-level bridge gives pFID 42.237 and FID 64.537, compared with 42.679 and 64.364 for Full. The staged bridge shares exactly the same 2K output as Full, so this comparison isolates the 4K bridge change. These results show that the chosen mixture influences the final distribution even when subsequent guidance is fixed; they do not establish a universal best bridge depth.

Table 6: Wavelet-based initialization ablations with LL2 guidance fixed at both child stages. Direct mixing uses $\lambda = 1$; Pixel Bridge uses $\lambda = 0$. The $(b_1, b_2) = (1, 2)$ bridge shares its 2K output with Full. KID and pKID are multiplied by 10^3; CLIP denotes the CLIP Score computed with ViT-L/14. Full is listed last in each resolution block.

Bridge	(b_1, b_2)	λ	FID ↓	KID ↓	pFID ↓	pKID ↓	IS ↑	pIS ↑	CLIP ↑
					2K				
Pixel	–	0	64.840	5.845	25.771	8.288	19.662	11.560	27.663
Wavelet	$(2, 2)$	1	64.741	5.621	23.966	6.244	19.707	12.168	27.591
Wavelet	$(3, 3)$	1	64.775	5.628	23.978	6.285	19.658	12.151	27.590
Wavelet	$(1, 2)$	1	64.561	5.346	23.776	6.351	20.420	12.172	27.573
Wavelet (Full)	$(1, 1)$	1	64.561	5.346	23.776	6.351	20.420	12.172	27.573
					4K				
Pixel	–	0	64.958	4.826	49.915	24.751	18.892	7.089	27.584
Wavelet	$(2, 2)$	1	64.710	4.564	42.429	17.428	19.160	7.955	27.493
Wavelet	$(3, 3)$	1	64.890	4.647	42.895	17.535	18.974	7.951	27.507
Wavelet	$(1, 2)$	1	64.537	4.443	42.237	17.687	19.270	7.872	27.487
Wavelet (Full)	$(1, 1)$	1	64.364	4.409	42.679	18.324	20.269	7.787	27.473

Table 7: Auxiliary metrics for the bridge-construction ablations. CLIP-B denotes the CLIP Score computed with ViT-B/16. All values use the same evaluation protocol as the primary ablations; Full is listed last in each resolution block.

Bridge	(b_1, b_2)	λ	CLIP-B ↑	Aesth. ↑	Repeat ↓	Spectral ↓
			2K			
Pixel	–	0	32.549	5.466	0.1267	0.0631
Wavelet	$(2, 2)$	1	32.516	5.491	0.1249	0.0615
Wavelet	$(3, 3)$	1	32.524	5.489	0.1246	0.0620
Wavelet	$(1, 2)$	1	32.507	5.492	0.1235	0.0591
Wavelet (Full)	$(1, 1)$	1	32.507	5.492	0.1235	0.0591
			4K			
Pixel	–	0	32.432	5.526	0.0982	0.0488
Wavelet	$(2, 2)$	1	32.401	5.555	0.0975	0.0483
Wavelet	$(3, 3)$	1	32.426	5.552	0.0973	0.0487
Wavelet	$(1, 2)$	1	32.378	5.559	0.0961	0.0462
Wavelet (Full)	$(1, 1)$	1	32.352	5.555	0.0973	0.0464

C.4 Structural Correction: Stage-Wise Guidance Depth

Here (ℓ_1, ℓ_2) denotes the recursive guidance depth at the 2K and 4K child stages, respectively. The child and parent undergo ℓ_k and $\ell_k - 1$ LL analyses to obtain compatible structural coordinates. Changing this depth changes the subspace of the explicit correction; it does not add generation stages or alter the amplitude factor $\gamma = 2$. All depth variants retain the one-level wavelet bridge, orthogonal initial-noise coupling, and the parent query Φ_2.

Tables 8 and 9 report all nine combinations in $\{1, 2, 3\}^2$, forming a complete 3×3 stage-wise sweep.

At each fixed $\ell_1 \in \{1, 2, 3\}$, increasing ℓ_2 reduces 4K pFID and pKID. For example, with $\ell_1 = 2$, pFID decreases from 43.685 to 42.679 to 42.191 as ℓ_2 increases from one to three, while FID increases from 64.317 to 64.364 to 64.621. Because the parent trajectory is shared in this comparison, the trend directly characterizes the effect of restricting the 4K correction to coarser shared coordinates. It is consistent with leaving more resolution-specific refinement to the child field.

Within the 3×3 sweep, the lowest FID is obtained by $(2, 1)$, the lowest KID by $(1, 2)$, and the lowest pFID and pKID by $(3, 3)$. Among uniform depths, LL2 has the lowest 4K FID and KID; compared with LL3, it also has lower 2K FID and KID, but worse patch-level scores at both resolutions. The LL2 default favors image-level distribution fidelity within this trade-off, rather than minimizing every metric over all stage-wise settings.

The 2K output depends only on ℓ_1. Consequently, all configurations sharing $\ell_1 = 1$, 2, or 3 reuse the corresponding LL1, Full, or LL3 2K scores in Tables 4 and 5. These depth comparisons assess the extent of guidance, not the effect of removing guidance altogether.

Table 8: Stage-wise guidance-depth ablations at 4K. The configurations in $\{1, 2, 3\}^2$ form the complete 3×3 sweep, with Full listed last. Configurations are written as (ℓ_1, ℓ_2). KID and pKID are multiplied by 10^3; CLIP denotes the CLIP Score computed with ViT-L/14.

Configuration	FID ↓	KID ↓	pFID ↓	pKID ↓	IS ↑	pIS ↑	CLIP ↑
$(1, 1)$	64.412	4.412	44.798	19.477	19.213	7.581	27.431
$(1, 2)$	64.415	4.311	43.897	18.980	19.195	7.745	27.444
$(1, 3)$	64.574	4.435	42.794	18.463	19.372	7.841	27.443
$(2, 1)$	64.317	4.465	43.685	18.977	19.328	7.592	27.463
$(2, 3)$	64.621	4.493	42.191	18.067	19.297	7.852	27.468
$(3, 1)$	64.321	4.486	43.015	18.618	19.411	7.663	27.496
$(3, 2)$	64.567	4.502	42.275	18.159	19.458	7.745	27.494
$(3, 3)$	64.764	4.467	41.817	17.860	19.506	7.868	27.499
$(2, 2)$ (Full)	64.364	4.409	42.679	18.324	20.269	7.787	27.473

Table 9: Auxiliary 4K metrics for the guidance-depth variants. The same parent trajectory is used when ℓ_1 is fixed. CLIP-B denotes the CLIP Score computed with ViT-B/16; Full is listed last.

Configuration	CLIP-B ↑	Aesth. ↑	Repeat ↓	Spectral ↓
$(1, 1)$	32.339	5.555	0.1045	0.0520
$(1, 2)$	32.342	5.556	0.1012	0.0497
$(1, 3)$	32.325	5.552	0.0990	0.0483
$(2, 1)$	32.358	5.555	0.1003	0.0483
$(2, 3)$	32.344	5.551	0.0960	0.0459
$(3, 1)$	32.387	5.551	0.1000	0.0482
$(3, 2)$	32.391	5.549	0.0970	0.0464
$(3, 3)$	32.388	5.547	0.0956	0.0459
$(2, 2)$ (Full)	32.352	5.555	0.0973	0.0464

D Test-Set Construction and Quality Control

We select 1,000 prompt–image pairs from RE-LAION-High-Resolution (LAION, 2024), a resource in the LAION dataset family (Schuhmann et al., 2022), balancing semantic diversity, category coverage, and image–text quality. The test set spans ten categories—Animals, People, Nature, Architecture, Food, Vehicles, Fashion, Objects, Arts, and Travel—with 100 pairs per category. Each English caption serves as a generation prompt, and the corresponding source images form the fixed reference set for FID.

Selection proceeds in five stages: **(1) Basic filtering** screens candidates using metadata on language, resolution, image–text similarity, and safety, excluding obvious advertising and web noise; **(2) Semantic filtering and ranking** checks caption relevance using category keywords and negative keyword lists, with AI models assisting prompt screening and category assignment while prioritizing clear and concise descriptions; **(3) Image-quality checks** assess file integrity, sharpness, aspect ratio, watermarks, and potential collages; **(4) Manual visual review** excludes image–text mismatches, poor composition, product displays, and other low-quality content; and **(5) Deduplication and final validation** combine file hashes with manual inspection to remove duplicates and near-duplicates and verify category counts and quality requirements.

AI-assisted prompt screening and category assignments are checked during manual review and final validation; quality standards are not relaxed to meet category quotas. All retained source images have both width and height of at least 1,024 pixels and are preserved in their original file form to avoid additional resizing or compression errors during dataset construction.

E Evaluation Protocol and Metric Definitions

We specify evaluation inputs and patch sampling before defining the distributional, semantic, aesthetic, and artifact metrics. Supplementary comparisons using these definitions follow in Appendix F.

E.1 Data Alignment and Reporting

For every method and resolution, evaluation uses the complete set of 1,000 generated images and their ID-aligned prompts. The real distribution for FID/KID consists of the fixed 1,000 original LAION-Balanced-1K images described in Appendix D. Resolutions are evaluated independently, without pooling $1\mathrm{K}$, $2\mathrm{K}$, and $4\mathrm{K}$ outputs. Sample counts, records, and image hashes are checked; incomplete sets are not substituted for the benchmark. Images are converted to RGB with PIL without automatic EXIF orientation correction. The main tables report means or set-level point estimates without $\pm$ split standard deviations; selected paired FID/pFID intervals appear in Appendix E.4. All reported results are computed under this protocol rather than copied from the original method papers.

E.2 Shared Patch Sampling Protocol

pFID, pKID, and pIS use the same crop definition: ten 256×256 native-pixel patches per image, with crop seed 42. Images are ordered by file path. The tuple consisting of the seed, image index, and crop index is hashed with SHA256 to drive deterministic uniform sampling over valid integer crop coordinates. Crops are taken from the original image before metric-specific input resizing; the full high-resolution image is not first downsampled. Patches may overlap or repeat and need not cover the image. The same rule is applied at each resolution without enforcing semantic correspondence between crops. Thus, a patch covers different relative areas at $2\mathrm{K}$ and $4\mathrm{K}$, measuring local quality at a fixed pixel scale.

pFID and pKID use 10,000 real patches and 10,000 generated patches under the same Clean-FID feature protocol as their image-level counterparts. pIS uses the 10,000 generated patches only. Multiple patches from one image do not increase the number of independent source images.

E.3 FID AND KID

FID (Heusel et al., 2017) uses Clean-FID (Parmar et al., 2022) 0.1.35 in clean mode with 2048-dimensional Inception-v3 features. For real and generated feature means μ_r, μ_g and covariances Σ_r, Σ_g, it is

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r^{1/2}\Sigma_g\Sigma_r^{1/2}\right)^{1/2}\right). \tag{84}$$

KID (Bińkowski et al., 2018) uses the same features and the unbiased squared maximum mean discrepancy with kernel

$$k(\mathbf{x}, \mathbf{y}) = \left(\frac{\mathbf{x}^\top \mathbf{y}}{2048} + 1\right)^3. \tag{85}$$

The estimator averages 100 subset evaluations, sampling without replacement up to 1,000 features from each distribution with seed 42. For image-level KID, both distributions contain exactly 1,000 features, so every subset is the full set: the repetitions only permute the samples and do not constitute 100 independent smaller-subset evaluations. For pKID, each distribution contains 10,000 patch features, and each evaluation selects 1,000 from each side. Raw summary-file KID scores are unscaled; all KID and pKID values in the paper table are multiplied by 10^3. Slightly negative unbiased estimates are retained, not clipped to zero. FID and KID compare feature distributions, not averages of paired image distances, even though each reference image has a corresponding prompt.

E.4 PAIRED BOOTSTRAP FOR TEMPORAL-QUERY ABLATIONS

We assess benchmark-sample variation in the 4K temporal-query ablations by resampling the 1,000 prompt–image IDs with replacement within each of the ten semantic categories (100 IDs per category). The same ID multiplicities are applied to the real reference and all three variants; all ten patches from an image move together for pFID. We recompute each score from its saved Clean-FID features over 200 paired replicates (seed 20260926). Table 10 reports percentile 95% intervals for $\Delta = \mathrm{metric}(\mathrm{alternative}) - \mathrm{metric}(\mathrm{Full})$, so positive values favor Full. Full's original scores are 64.364 for FID and 42.679 for pFID.

Table 10: 4K paired metric differences for the temporal-query ablations under the fixed-seed LAION-Balanced-1K protocol. Intervals are percentile bootstrap intervals over source-image clusters, not over independently generated seeds.

Metric	Alternative	Point Δ	95% CI for Δ
FID	Same-σ Query	$+0.330$	$[-0.097, +0.655]$
FID	Half-Shift Query	$+0.063$	$[-0.206, +0.298]$
pFID	Same-σ Query	$+1.660$	$[+1.237, +2.122]$
pFID	Half-Shift Query	$+0.485$	$[+0.227, +0.769]$

For both alternatives, paired bootstrap intervals for the pFID differences to Full are positive, whereas the FID intervals include zero. These intervals quantify variation over the selected prompt–image units conditional on one generation seed. They do not quantify variation across repeated generation seeds or address possible bias from the reference-set construction. We report paired differences rather than separate percentile intervals for individual FID scores: with 1,000 images and 2,048-dimensional features, duplicate-heavy resamples produce substantial finite-sample upward bias in the individual FID estimates.

E.5 IS AND PATCH-IS

IS (Salimans et al., 2016) and pIS use the official-compatible Inception implementation in `torch-fidelity==0.3.0`, with 1008-class `logits_unbiased` converted to class probabilities $p(y \mid x)$. The model internally applies TF1-compatible bilinear resizing to 299×299. Samples are shuffled with seed 42 and divided into ten splits: 100 images per split for IS and 1,000 patches

per split for pIS. For a split S, define

$$p_S(y) = \frac{1}{|S|} \sum_{x \in S} p(y \mid x),$$
$$\mathrm{IS}(S) = \exp\left[\frac{1}{|S|} \sum_{x \in S} D_{\mathrm{KL}}(p(y \mid x) \,\|\, p_S(y)) \right]. \tag{86}$$

The reported score is the mean of the ten split scores, without their standard deviation. Neither metric uses real reference images. Patches from one image can fall into different splits, so these patches must not be treated as independent source samples. IS and pIS magnitudes should not be compared directly.

E.6 CLIP Scores and Aesthetics

We compute a rescaled CLIPScore-style similarity (Hessel et al., 2021) using OpenAI CLIP (Radford et al., 2021) ViT-L/14 and ViT-B/16, denoted CLIP-L/14 and CLIP-B/16, respectively. Each model uses its corresponding `CLIPProcessor`: RGB input, bicubic resizing, a 224×224 center crop, and CLIP normalization. Text inputs are the original ID-aligned prompts; sequences longer than 77 tokens are truncated, and the number of truncated prompts is recorded. For the image and text encoders f_{img} and f_{text} of the chosen model,

$$\mathrm{CLIPScore} = \frac{100}{N} \sum_{i=1}^{N} \max\left(0, \frac{f_{\mathrm{img}}(I_i)^\top f_{\mathrm{text}}(c_i)}{\|f_{\mathrm{img}}(I_i)\|_2 \|f_{\mathrm{text}}(c_i)\|_2}\right), \qquad N = 1000. \tag{87}$$

We use the positive-part cosine construction of CLIPScore with a scale factor of 100, rather than the original 2.5, and the encoders specified above rather than the original ViT-B/32. The original prompts are used without the caption prefix in the original CLIPScore protocol. The learned `logit_scale` is not used. The main comparison reports the ViT-L/14-based CLIP Score only; Table 12 reports scores from both encoders in separate columns. Scores from the two encoders are not pooled or substituted for one another.

Aesthetics feeds the normalized 768-dimensional ViT-L/14 image feature into the fixed `sac+logos+ava1-l14-linearMSE` aesthetic MLP (Schuhmann, 2022) and averages its per-image predictions. It uses the same image preprocessing, does not use prompts, and does not clamp outputs to a prescribed score range.

E.7 Repeat and Spectral Diagnostics

For both diagnostics, the full image is converted to grayscale and resized with Lanczos to 512×512. A separable smoothing kernel $[1, 2, 3, 2, 1]/9$ is applied along both axes; subtracting the resulting low-frequency image gives the high-pass image H.

For Repeat, subtract the mean of H and compute its circular spatial autocorrelation C via FFT. After centering the zero displacement, the per-image diagnostic is

$$\mathrm{Repeat}(I) = \frac{Q_{0.999}(\{C(\delta) : 4 \le \|\delta\|_2 \le 64\})}{C(0)}, \tag{88}$$

where $Q_{0.999}$ is the 99.9th percentile and displacement distances are measured in pixels of the resized grayscale image. The ratio is defined for positive zero-displacement autocorrelation.

For Spectral, compute the two-dimensional Fourier power spectrum $P(\omega) = |\mathcal{F}H(\omega)|^2$ and center its zero frequency. Let $\Omega = \{\omega : \|\omega\|_2 \ge 8\}$ exclude the central disk of radius eight frequency bins, and let $T \subset \Omega$ contain the top 0.1% of the remaining bins by power. Then

$$\mathrm{Spectral}(I) = \frac{\sum_{\omega \in T} P(\omega)}{\sum_{\omega \in \Omega} P(\omega)}. \tag{89}$$

This ratio is defined when the remaining total power is positive. Each diagnostic is averaged over the 1,000 generated images. These scores measure periodic responses and spectral concentration, respectively, using 512×512 full-image preprocessing distinct from the native-pixel patch protocol.

F Supplementary Quantitative Comparisons

F.1 Pretrained High-Resolution Generator References

Diffusion-4K (Zhang et al., 2025), UltraFlux (Ye et al., 2025), SANA-4K (Xie et al., 2024), and UltraPixel (Ren et al., 2024) use separately pretrained high-resolution systems released by their respective authors. They do not share the FLUX.1-dev comparison pipeline used in the main text and are therefore reported only as unranked appendix references. Table 11 gives their primary metrics; their CLIP-B/16, Aesthetics, Repeat, and Spectral results are included in Tables 12 and 13.

Table 11: Primary metrics for separately pretrained high-resolution generators at 2K and 4K. KID and pKID are multiplied by 10^3; CLIP denotes the CLIP Score computed with ViT-L/14. These systems are appendix references and are not ranked against the FLUX.1-dev-based primary comparison.

Method	FID ↓	KID ↓	pFID ↓	pKID ↓	IS ↑	pIS ↑	CLIP ↑
		2K					
Diffusion-4K	65.719	4.619	31.514	14.186	19.127	10.364	27.178
UltraFlux	66.982	5.341	35.737	15.615	19.340	11.242	27.669
SANA-4K	70.228	5.297	36.013	12.935	19.676	13.418	25.534
UltraPixel	67.233	6.878	32.393	10.523	19.906	12.987	28.064
		4K					
Diffusion-4K	68.248	4.323	77.606	40.120	18.034	6.673	25.863
UltraFlux	65.943	4.481	42.254	17.588	18.681	9.492	28.147
SANA-4K	70.238	5.301	52.888	22.234	19.339	8.863	25.537
UltraPixel	67.632	7.428	48.158	23.260	19.546	8.915	27.721

Among these separately trained references, Diffusion-4K obtains the lowest 2K FID, KID, and pFID, whereas UltraPixel gives the highest 2K IS and CLIP Score and SANA-4K gives the highest pIS. At 4K, UltraFlux leads on FID, pFID, pKID, pIS, and CLIP Score, while Diffusion-4K retains the lowest KID and UltraPixel the highest IS. The variation across metrics reflects the distinct training data, architectures, and high-resolution objectives of these independently trained systems.

F.2 CLIP Encoder Comparison

Table 12 reports CLIP scores computed with both encoders under the protocol in Section E.6.

Table 12: CLIP Scores computed with ViT-B/16 and ViT-L/14 at 2K and 4K. The main comparison reports the ViT-L/14-based score only; the ViT-B/16-based score is retained here for completeness. Scores from different encoders are not directly comparable. Blue cell shading indicates 1st, 2nd, and 3rd place among training-free inference and super-resolution methods at each resolution. Pretrained high-resolution generators are reference only.

Method	CLIP Score (B/16) ↑	CLIP Score (L/14) ↑
2K		
Training-free inference and super-resolution		
FLUX-Direct	31.758	26.864
I-Max	32.726	27.864
HiFlow	32.656	27.779
ScaleDiff	32.709	27.843
DyPE	32.295	27.470
UltraImage	32.615	27.673
SEGA	32.272	27.320
HRDiT	32.693	27.734
BSRGAN	32.591	27.601
OSEDiff	32.264	27.375
WaveAlign(ours)	32.507	27.573
Pretrained high-resolution generators (reference only)		
Diffusion-4K	32.105	27.178
UltraFlux	32.425	27.669
SANA-4K	31.082	25.534
UltraPixel	33.146	28.064
4K		
Training-free inference and super-resolution		
FLUX-Direct	18.624	12.467
I-Max	32.497	27.669
HiFlow	32.523	27.667
ScaleDiff	32.678	27.871
UltraImage	32.436	27.605
SEGA	29.086	23.860
HRDiT	32.633	27.715
BSRGAN	32.569	27.587
OSEDiff	32.413	27.473
WaveAlign(ours)	32.352	27.473
Pretrained high-resolution generators (reference only)		
Diffusion-4K	30.997	25.863
UltraFlux	32.911	28.147
SANA-4K	31.086	25.537
UltraPixel	32.834	27.721

Within the FLUX.1-dev-based comparison, I-Max obtains the highest score with both encoders at 2K, followed by ScaleDiff. At 4K, ScaleDiff ranks first with both encoders (32.678/27.871), followed by HRDiT. WaveAlign changes from 32.507/27.573 at 2K to 32.352/27.473 at 4K for ViT-B/16 and ViT-L/14, respectively, showing substantially greater cross-resolution stability than direct inference. Among the separately pretrained references, UltraPixel gives the highest two scores at 2K, whereas UltraFlux does so at 4K.

F.3 Aesthetics and Artifact Diagnostic Results

Table 13 reports Aesthetics, Repeat, and Spectral for the methods included at each resolution, supplementing the main quantitative comparison. The top three results for each metric are highlighted according to their reported direction.

Table 13: Aesthetics and artifact diagnostics at 2K and 4K. Aesth. denotes Aesthetics, based on ViT-L/14 features. Blue cell shading indicates 1st, 2nd, and 3rd place for each metric among training-free inference and super-resolution methods at each resolution. Pretrained high-resolution generators are reference only and are not ranked.

Method	Aesth. ↑	Repeat ↓	Spectral ↓
2K			
Training-free inference and super-resolution			
FLUX-Direct	5.409	0.1557	0.0841
I-Max	5.447	0.1137	0.0575
HiFlow	5.507	0.1149	0.0592
ScaleDiff	5.550	0.1188	0.0587
DyPE	5.410	0.0907	0.0469
UltraImage	5.516	0.0787	0.0399
SEGA	5.372	0.0945	0.0495
HRDiT	5.473	0.1119	0.0568
BSRGAN	5.497	0.1451	0.0716
OSEDiff	5.551	0.1396	0.0678
WaveAlign(ours)	5.492	0.1235	0.0591
Pretrained high-resolution generators (reference only)			
Diffusion-4K	5.540	0.1379	0.0661
UltraFlux	5.687	0.1308	0.0604
SANA-4K	5.953	0.1044	0.0470
UltraPixel	5.865	0.1258	0.0610
4K			
Training-free inference and super-resolution			
FLUX-Direct	4.315	0.9130	0.8912
I-Max	5.492	0.1279	0.0674
HiFlow	5.537	0.1102	0.0575
ScaleDiff	5.575	0.1115	0.0551
UltraImage	5.515	0.1202	0.0635
SEGA	5.245	0.1097	0.0586
HRDiT	5.488	0.0997	0.0516
BSRGAN	5.497	0.1481	0.0718
OSEDiff	5.489	0.1406	0.0678
WaveAlign(ours)	5.555	0.0973	0.0464
Pretrained high-resolution generators (reference only)			
Diffusion-4K	5.368	0.1420	0.0724
UltraFlux	5.499	0.0771	0.0403
SANA-4K	5.954	0.1039	0.0467
UltraPixel	5.811	0.1187	0.0578

At 2K, OSEDiff obtains the highest Aesthetics score among the FLUX.1-dev-based methods, followed closely by ScaleDiff, while UltraImage records the lowest Repeat and Spectral values. At 4K, ScaleDiff achieves the highest Aesthetics score, followed by WaveAlign, whereas WaveAlign retains the lowest values on both artifact diagnostics. For the separately pretrained references, SANA-4K has the highest Aesthetics score at both resolutions, and UltraFlux records the lowest 4K Repeat and Spectral values.

G Related Work

G.1 Training-Based High-Resolution Image Generation

Training-based systems learn high-resolution representations or adapt parameters. UltraPixel (Ren et al., 2024) transfers semantics through learned implicit upsampling and scale-aware normalization. Diffusion-4K (Zhang et al., 2025) combines a partitioned VAE with wavelet-based latent fine-tuning. UltraFlux (Ye et al., 2025) couples 4K data with positional adaptation, VAE post-training, and a wavelet objective. SANA (Xie et al., 2024) improves efficiency through deep compression and linear attention. Latent Wavelet Diffusion (Sigillo et al., 2026) adds wavelet-guided detail training and scale-consistent VAE training.

G.2 Training-Free Progressive Generation and Flow Guidance

Training-free methods coordinate pretrained denoisers across spatial windows or resolution stages. MultiDiffusion (Bar-Tal et al., 2023) fuses overlapping paths, and SyncDiffusion (Lee et al., 2023) adds perceptual-gradient synchronization. For editing, W-EDIT (Sun et al., 2026) separates structure and detail with wavelets. DemoFusion (Du et al., 2024a) combines progressive upscaling, skip residuals, and dilated sampling. MrFlow (Zheng et al., 2026) accelerates generation using pretrained super-resolution and low-noise refinement.

For Rectified Flow, I-Max (Du et al., 2024b) guides projected flow toward a low-resolution endpoint, alongside decaying guidance and resolution-adaptation tools. HiFlow (Bu et al., 2025) constructs a virtual reference flow from intermediate clean predictions and coordinates initialization, Fourier low-frequency direction correction, and acceleration alignment. WaveAlign specifies the correspondence underlying trajectory guidance: shared stochastic realizations, a scale-adjusted parent phase, and recursively shared structural coordinates. Its distinction is this joint specification, with orthogonal corrections preserving complementary coordinates at each explicit update.

G.3 Positional and Attention Adaptation

DyPE (Issachar et al., 2025) progressively relaxes positional extrapolation during denoising. UltraImage (Zhao et al., 2025) combines rotary-frequency correction with entropy-guided attention concentration. SEGA (Rajabi et al., 2026) modulates rotary attention components using latent spectral energy and flatness. HRDiT (Xue et al., 2026) combines spatial position alignment with head-adaptive attention pruning. These methods modify positional encoding or attention within model evaluations; WaveAlign couples information across resolution stages. We combine this cross-stage mechanism with stage-specific positional and attention settings.