

UltraFlux: Data-Model Co-Design for High-quality Native 4K Text-to-Image Generation across Diverse Aspect Ratios

Tian Ye*
HKUST(GZ)

tye610@connect.hkust-gz.edu.cn

Song Fei*
HKUST(GZ)

secondauthor@i2.org

Lei Zhu†
HKUST, HKUST(GZ)

leizhu@ust.hk

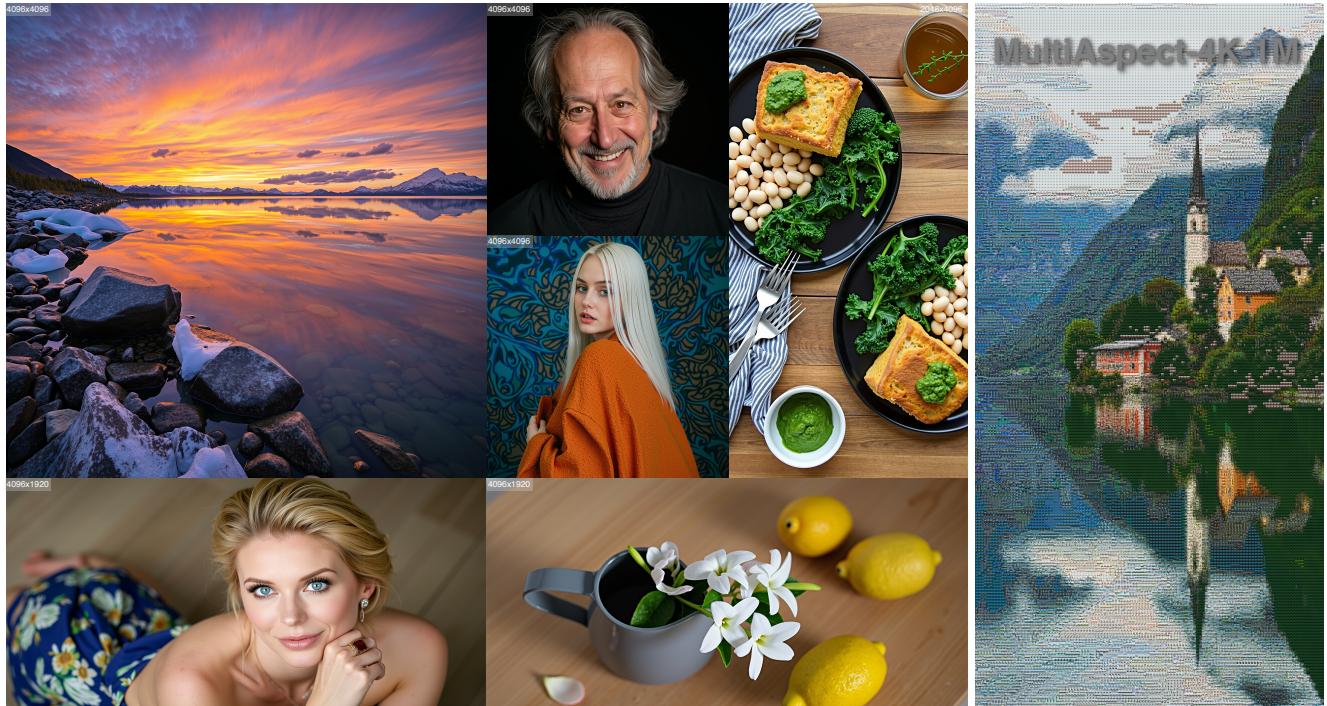


Figure 1. **Left:** UltraFlux generates photorealistic 4K images across diverse aspect ratios and topics while maintaining high aesthetic quality and faithful content depiction with a single unified text-to-image model. **Right:** Our MultiAspect-4K-1M is a large-scale high-quality dataset for 4K image synthesis.

Abstract

Diffusion transformers have recently delivered strong text-to-image generation around 1K resolution, but we show that extending them to native 4K across diverse aspect ratios exposes a tightly coupled failure mode spanning positional encoding, VAE compression, and optimization. Tackling any of these factors in isolation leaves substantial quality on the table. We therefore take a data-model co-design view and introduce UltraFlux, a Flux-based DiT trained natively at 4K on MultiAspect-4K-1M, a 1M-image 4K corpus

with controlled multi-AR coverage, bilingual captions, and rich VLM/IQA metadata for resolution- and AR-aware sampling. On the model side, UltraFlux couples (i) Resonance 2D RoPE with YaRN for training-window-, frequency-, and AR-aware positional encoding at 4K; (ii) a simple, non-adversarial VAE post-training scheme that improves 4K reconstruction fidelity; (iii) an SNR-Aware Huber Wavelet objective that rebalances gradients across timesteps and frequency bands; and (iv) a Stage-wise Aesthetic Curriculum Learning strategy that concentrates high-aesthetic supervision on high-noise steps governed by the model prior.

*Equal contribution, UltraFlux project leader: Tian Ye

†Corresponding author.

Together, these components yield a stable, detail-preserving 4K DiT that generalizes across wide, square, and tall ARs. On the Aesthetic-Eval@4096 benchmark and multi-AR 4K settings, UltraFlux consistently outperforms strong open-source baselines across fidelity, aesthetic, and alignment metrics, and—with a LLM prompt refiner—matches or surpasses the proprietary Seedream 4.0.

1. Introduction

Diffusion transformers (DiTs) [2, 5, 6, 22, 36] have recently pushed text-to-image generation to impressive quality around 1K resolution, enabled by efficient backbones, token compression, and carefully tuned training pipelines [5, 36]. However, extending these systems to *native 4K* while supporting a broad spectrum of aspect ratios (ARs) is not a simple matter of scaling resolution. At 4096×4096 and beyond, we empirically observe three coupled challenges: (i) *positional representation and AR extrapolation*, where 2D rotary embeddings calibrated on a single training window can drift or alias under large changes in resolution and AR [23, 42]; (ii) *high-frequency fidelity under VAE compression*, where higher downsampling factors improve throughput but tend to erase fine structures that dominate 4K perception [36, 40]; and (iii) *4K-aware optimization*, where gradient contributions become heavily skewed across timesteps and frequency bands, making standard objectives poorly matched to the statistics of 4K latents [9, 40]. These factors interact: the choice of positional scheme, VAE compression ratio, and training objective jointly determines whether a model can remain stable and detailed across native 4K resolutions and diverse ARs.

On the model side, several scaling strategies partially address these issues but leave the overall design space fragmented. Training-free high-resolution methods mitigate tiling artifacts and duplication at inference time, yet largely preserve the underlying positional encoding and were not designed for systematic multi-AR extrapolation [12, 43]. Decoder-side approaches based on global-local fusion or tiled diffusion improve size flexibility but introduce new failure modes, such as coherence gaps across tiles or heavy reliance on a global prior for consistency [1, 8]. Native-4K systems demonstrate that carefully engineered backbones can make 4K training tractable [5, 36], but most emphasize token/architecture efficiency and treat *positional robustness, VAE compression, and loss design* as largely orthogonal choices rather than a jointly optimized 4K regime.

Progress at 4K is further constrained by the data itself. Public 4K corpora are typically modest in scale (on the order of $10^4\text{--}10^5$ images), heavily biased toward near-square ARs and landscape-centric content, and curated with early CLIP-based aesthetic predictors. For example, Aesthetic-4K takes an important step by assembling high-

quality 4K image–text pairs with GPT-4O captions [40], yet its scale and AR coverage remain limited for studying *resolution–AR coupling*, and its subject distribution under-represents human-centric scenes. More critically, existing 4K datasets rarely provide the *structured metadata* needed for modern 4K training. As a result, practitioners have limited control over sampling data slices tailored to specific training regimes (e.g., high-detail or high-aesthetic subsets), and it becomes difficult to perform fine-grained aesthetic or AR-conditioned analyses.

On the optimization and adaptation side, recent work explores complementary—but still incomplete—directions. Wavelet-aware objectives at native resolution improve fidelity on strong backbones by better emphasizing high-frequency content [40], yet they typically combine simple quadratic or perceptual penalties and thus remain vulnerable to cross-scale dominance of low-frequency energy. Latent-space super-resolution and self-cascade schemes sharpen details beyond the original training resolution and reduce the cost of high-resolution transfer [7, 14], but they operate as post-hoc adapters on fixed backbones and do not resolve the underlying trade-off between VAE compression and 4K reconstruction fidelity. In parallel, timestep curricula adjust noise sampling while holding the data distribution fixed, and aesthetic post-training applies high-aesthetic data uniformly across timesteps, leaving unexplored the regime where *high-noise steps—those most governed by the model prior—are selectively sculpted by high-aesthetic supervision*. Finally, existing RoPE interpolation and NTK-style scaling strategies are primarily developed for 1D sequence length extrapolation, and provide little guidance for *2D token grids at native 4K under strongly varying ARs*, where misaligned phase behavior manifests as ghosting, drift, and striping artifacts. Altogether, native 4K multi-AR generation still lacks a unified framework that couples: (i) a large-scale, multi-AR, content-diverse, VLM-curated 4K corpus with rich metadata; (ii) an efficient, non-adversarial VAE post-training strategy that improves 4K reconstruction without sacrificing throughput; (iii) an SNR-Aware Huber Wavelet Training Objective and a stage-wise aesthetic curriculum matched to 4K statistics; and (iv) a training-window aware, band-aware, and AR-aware positional encoding scheme. In this work, we explicitly target this data–model co-design space.

We will open-source the full dataset, training code, pre-trained weights, and inference code. Concretely, we make the following contributions:

- **MultiAspect-4K-1M: a large-scale, multi-AR, aesthetically curated 4K corpus.** We construct a 1M-scale 4K dataset with native 4K and near-4K resolution, controlled aspect-ratio coverage, and a dual-channel pipeline that debiases landscape-heavy sources toward human-centric content. Each image is accompanied by decoupled VLM-

- based quality and aesthetic scores, classical IQA signals, bilingual captions, and subject tags, providing the structured metadata needed for data–model co-design.
- **UltraFlux: a data–model co-designed DiT for native 4K multi-AR generation.** We train a Flux-based backbone on MultiAspect-4K-1M with a co-designed recipe that couples (i) *Resonance 2D RoPE with YaRN* for training–window aware, band-aware, and AR-aware positional encoding, (ii) an *SNR-Aware Huber Wavelet Training Objective* tailored to 4K latents, (iii) a *Stage-wise Aesthetic Curriculum Learning (SACL)* scheme that concentrates high-aesthetic supervision on high-noise steps, and (iv) a simple, non-adversarial, data-efficient VAE post-training procedure that improves Flux VAE reconstructions at 4K. Together, these components yield a stable, detail-preserving DiT for native 4K synthesis across diverse ARs.
 - **State-of-the-art native 4K performance.** On standard 4K benchmarks and popular metrics covering fidelity, aesthetic quality, and text alignment, UltraFlux consistently outperforms strong 4K baselines, including recent native-4K and training-free scaling methods.

2. Related Work

This section reviews approaches to scaling text-to-image diffusion models to high-resolution T2I, native-4K and diverse aspect ratios. We group prior work into three lines: training-free inference-time scaling, lightweight adaptations (e.g., latent super-resolution and self-cascade), and native-4K training with 4K-capable backbones.

Training-Free High-Resolution Scaling. Training-free strategies extend pre-trained 512–1K models to 2K/4K and diverse aspect ratios by modifying inference-time computation, without re-training. *HiDiffusion* diagnoses duplication and quadratic self-attention costs at high resolutions and introduces a resolution-aware U-Net and windowed attention to improve quality and speed [43]. *FouriScale* approaches ultra-high resolution from the frequency view via Fourier-domain low-pass guidance and dilated convolutions, improving global structure while injecting high frequencies [12]. These approaches are effective for quick scaling, yet commonly *keep the original positional scheme unchanged*, which leaves *positional extrapolation stability under extreme ARs* only partially addressed [1, 8, 12, 43].

Lightweight Adaptation: Latent SR and Self-Cascade Models. Lightweight adaptations improve high-resolution quality with minimal cost by augmenting the sampling pipeline or attaching small modules. *LSRNA* maps low-res latents to a high-res manifold via latent-space super-resolution and injects region-wise noise to restore high-frequency detail without retraining the base model [14]. *Self-Cascade Diffusion* integrates low-resolution generation into the high-resolution denoising process and optionally

fine-tunes small multi-scale upsamplers, achieving rapid 4K adaptation at a fraction of full fine-tuning cost [7]. While these methods markedly sharpen details and reduce adaptation overhead, they typically *inherit the original positional scheme*, leaving *AR-generalized extrapolation* under-explored [7, 14].

Native 4K Training and 4K-Capable Foundation Models. A complementary direction trains or fine-tunes models directly at native-4K and curates high-quality 4K corpora. *Diffusion-4K* introduces Aesthetic-4K and a wavelet-based fine-tuning scheme that improves fidelity and prompt alignment on large modern backbones [40]. Meanwhile, efficient backbones such as *PixArt- Σ* (token-compression attention) and *Sana* ($32\times$ VAE with linear-attention DiT) make 4096×4096 synthesis computationally feasible at small model scales [5, 36]. Despite these advances, public corpora remain limited in *scale and AR diversity*, constraining systematic study of *resolution–AR coupling*. Moreover, end-to-end methodologies for stable native-4K training are under-documented and fragmented across implementations, slowing progress and curbing real-world adoption while masking the gains of true 4K training. A practical distinction concerns *native* versus *upscale-based 4K*. Unlike platform services that reach 4K primarily via $2\times/4\times$ upscalers (e.g., Midjourney [21]; Google Imagen [25] exposes a dedicated upscaler; Ideogram offers a $2\times$ Upscale endpoint [13]), recent closed-source leaders such as *See-dream 4* [27] explicitly support *multi-AR, native-4K* generation within a unified T2I/editing architecture. This distinction matters: cascade upscaling pipelines couple low-resolution synthesis with a separate restoration prior, conflating high-frequency fidelity and positional extrapolation, whereas native-4K training compels the backbone to learn long-range dependencies and cross-AR spatial alignment directly. *Therefore, we treat native-4K as a distinct training/evaluation regime and design both our data (MultiAspect-4K-1M) and recipe (UltraFlux) to isolate the gains of true 4K training from post-hoc upscaling.*

3. Method: Data–Model Co-Design for Native 4K Multi-AR Generation

3.1. MultiAspect-4K-1M Dataset

Design goals and scope. Public 4K corpora for text-to-image training remain modest in scale (typically below 10^5 images) and are usually curated with early CLIP-based aesthetic predictors such as LAION-Aesthetic. While these datasets already achieve reasonably good visual quality, their *aspect-ratio (AR) coverage* is coarse and imbalanced—only a few popular ARs are well-populated at native 4K—and the textual side (captions and aesthetic/quality supervision) is constrained by legacy CLIP-only scoring. Our data design therefore targets three comple-

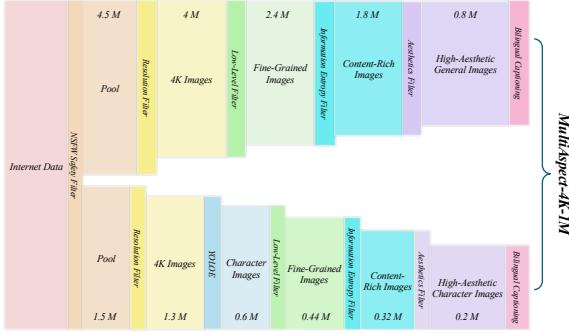
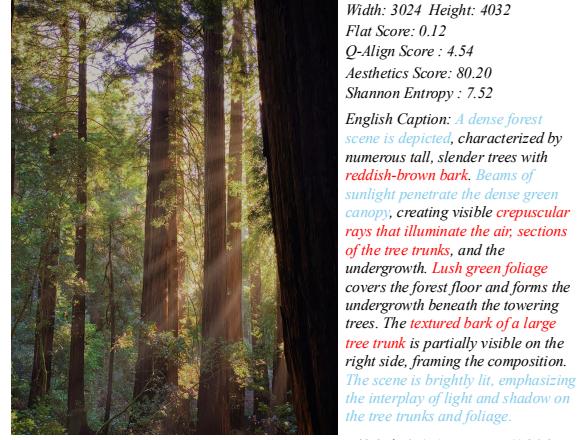


Figure 2. Data Pipeline overview.

mentary gaps: (i) *broad multi-AR coverage* at native 4K to avoid overfitting to a small set of AR buckets; (ii) *refreshed supervision quality*, coupling modern VLM-based quality/aesthetics estimators instead of relying solely on legacy CLIP-based predictors; and (iii) *distribution debiasing* that compensates for the over-representation of landscapes and the under-representation of human-centric content in existing 4K sources. We adopt a *VLM-driven* filtering strategy—semantic quality via *Q-Align* [35] and aesthetics via *ArtiMuse* [3]—complemented by interpretable classical signals (flatness and information entropy), and a dedicated *character augmentation* branch to improve recall for human subjects. Figure 2 sketches the two-channel pipeline and the final merge.

Sources and overall structure. After an NSFW safety check, we curate from a pool of approximately 6M high-resolution images whose subject distribution is skewed toward landscapes. To operationalize the goals in Sec. 3.1, we adopt a *dual-channel* pipeline: (i) a **general, AR-aware curation** path that enforces native/near-4K resolution and broad aspect-ratio (AR) coverage while filtering for quality and aesthetics; and (ii) a **human-centric augmentation** path that restores the underrepresented *character* category via open-vocabulary detection. The two channels are merged after de-duplication and with consistent metadata (resolution/AR, VLM scores, classical signals, caption, subject tags), yielding 1M images. Figure 2 provides a high-level overview and stage-wise retention.

VLM-based Filtering for HQ 4K Data. We begin with a safety screen, then enforce a *pixel-count threshold* as resolution filtering stage—images must have at least 3840×2160 total pixels; we *preserve each image’s native aspect ratio without any resizing*. This keeps the corpus artifact-free while naturally retaining a wide spectrum of ARs (e.g., 1:1, 16:9, 3:2, 4:3, 9:16), enabling transparent auditing of AR coverage. The resulting resolution and aspect-ratio distribution across 4K corpora is visualized in Figure 4. On this scaffold, we *decouple quality from aesthetics*: for quality we adopt *Q-Align*, a large multimodal model (LMM)-based visual scorer shown to deliver robust IQA judgments via discrete text-level supervision, and for aesthetics we use *ArtiMuse*, a recent MLLM-based image aesthetics evaluator that provides numeric scores together with *reasoned, expert-style explanations* (rather than score-only outputs). Classical, interpretable signals—*flatness* and



Width: 3024 Height: 4032
Flat Score: 0.12
Q-Align Score : 4.54
Aesthetics Score : 80.20
Shannon Entropy : 7.52
English Caption: A dense forest scene is depicted, characterized by numerous tall, slender trees with reddish-brown bark. Beams of sunlight penetrate the dense green canopy, creating visible crepuscular rays that illuminate the air, sections of the tree trunks, and the undergrowth. Lush green foliage covers the forest floor and forms the undergrowth beneath the towering trees. The textured bark of a large tree trunk is partially visible on the right side, framing the composition. The scene is brightly lit, emphasizing the interplay of light and shadow on the tree trunks and foliage.
Chinese Caption: 这个场景描绘了一片茂密的森林，其中有许多高大、纤细的树木，树皮呈红棕色。阳光穿过茂密的绿色树冠，形成明显的光线，照亮了空气、树干的部分区域以及林下植被。郁郁葱葱的绿叶覆盖了森林地面，构成了高耸树木下的植被层。右侧部分可以看到一棵大树树干上粗糙的树皮纹理，为整个画面提供了视觉上的框架。整个场景光线充足，突显了光线与树木及树叶之间微妙的相互作用。

Figure 3. Dataset example.

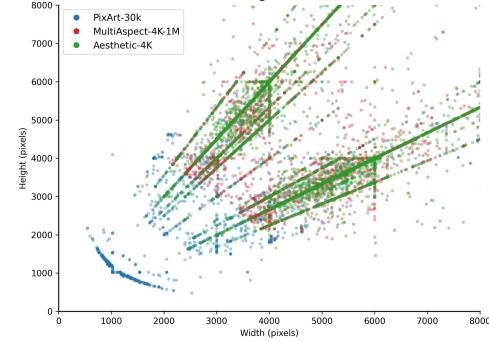


Figure 4. Dataset aspect and resolution analysis. All datasets use 10k samples. MultiAspect-4K-1M has a broader aspect ratio distribution.

information entropy—act as guardrails to suppress low-texture or overly smooth images that VLMs may tolerate, yielding a cleaner, high-frequency-preserving pool without sacrificing semantic clarity.

Human-centric augmentation via open-vocabulary detector. To correct the chronic under-representation of people in 4K sources, we run a targeted augmentation path. Candidate images are collected via person-related retrieval under the same safety and resolution/AR checks. We then apply the same *Q-Align* and *ArtiMuse* filters, strengthened by *information entropy* to suppress low-texture portraits. Crucially, we require *structured evidence of human presence* using **YOLOE** [32], a promptable open-vocabulary detector, which improves both recall and precision beyond fixed-class detectors. The accepted subset is merged into the main pool with a *character* flag.

Key statistics and comparisons to existing 4K and popular T2I datasets are summarized in Table 1.

Table 1. Dataset statistical comparisons.

Dataset	Number	Avg. Resolution	Avg. ArtiMuse	Avg. Caption Length	Bilingual Caption
PixArt-30k [5]	30,000	2,531 \times 2,656	63.67	87.9 tokens	✗
Aesthetic-4K [40]	12,009	4,576 \times 4,837	63.49	31.0 tokens	✗
MultiAspect-4K-1M	1,007,230	4,521 \times 4,703	64.59	125.1 tokens	✓

Bilingual captioning. Captioning is performed *last*. For the retained set, we generate detailed captions with *Gemini-2.5-Flash*, a production multimodal model suitable for fast, high-quality captioning; we then translate each caption into accurate Chinese with *Hunyuan-MT-7B* [44] to serve bilin-

gual users. An example image with its metadata and bilingual captions is shown in Figure 3.

The final *MultiAspect-4K-1M* corpus comprises *1M* 4K images with balanced coverage over standard AR buckets and a diversified subject mix (landscapes, people, objects). Each image includes resolution, *Q-Align*, *ArtiMuse*, flatness/entropy, English/Chinese caption, and the character tag. These fields are designed for transparent auditing and flexible data-model co-design: *they can act as analysis tags and stratified sampling keys for text-to-image training*.

3.2. UltraFlux: Scaling Flux to Native 4K Image Generation

With the data foundation in place, we now turn to the model side and build *UltraFlux*. Rather than redesigning the DiT architecture, we keep the core Flux transformer intact and focus on three components that bottleneck 4K performance: the VAE, the positional representation, and the training objective and strategy. We first post-train an *F16* VAE to recover fine details without giving up the efficiency gains of stronger compression, then introduce a Resonance 2D RoPE with a YaRN-style extrapolation scheme to stabilize attention across resolutions and aspect ratios. Finally, we couple an SNR-aware Huber wavelet loss with a stage-wise aesthetic curriculum that concentrates learning on high-frequency structure and high-aesthetic examples. Together, these lightweight but targeted changes upgrade Flux into an efficient, high-fidelity 4K generator that can fully exploit *MultiAspect-4K-1M*.

3.2.1. VAE Post-training for High-resolution Reconstruction Fidelity

A strong but efficient VAE is essential for practical native 4K image generation. The Flux backbone uses an *F8* VAE (height/width downsampling by 8), which at 4K yields a very large latent grid and makes sampling prohibitively slow: in our profiling, a single 4K image with 50 diffusion steps takes on the order of 30 minutes. Following Diffusion-4K [40], we instead adopt an *F16* VAE, halving the latent resolution while keeping comparable channel capacity, and focus on *post-training* the decoder to improve high-resolution reconstruction fidelity.

We fine-tune the Flux *F16* decoder on our curated *MultiAspect-4K-1M* corpus to enhance fine-scale 4K detail. Ablations on loss design and training recipes lead to three key findings: (*i*) explicitly targeting high-frequency content is essential—combining a wavelet reconstruction loss applied to high-frequency sub-bands with a feature-space perceptual loss consistently outperforms purely pixel-wise and perceptual objectives in sharpness and structural fidelity; (*ii*) once reconstruction reaches a reasonable regime, an adversarial discriminator offers negligible benefit—the GAN loss saturates quickly, induces optimization instability, and

fails to improve perceptual quality, so our final recipe omits the adversarial term and retains only wavelet, perceptual and L_2 losses; and (*iii*) stringent data curation substantially reduces post-training cost—by applying a flatness filter to select a high-detail subset of *MultiAspect-4K-1M*, we obtain the bulk of reconstruction gains within around 4k update steps, and observe that a few hundred thousand carefully screened, detail-rich images suffice to markedly upgrade the Flux *F16* VAE without multi-day GAN training or tens of millions of samples. This lightweight post-training stage produces a decoder that preserves fine 4K structures while maintaining the throughput advantages of *F16* compression, thereby enabling native 4K synthesis with both high fidelity and practical efficiency.

3.2.2. Resonance 2D RoPE for Multi-AR 4K Extrapolation

The official Flux backbone employs a fixed per-axis rotary spectrum with an optional *global* NTK factor [2], following the standard RoPE formulation [30] but without band-specific treatment or training-window awareness. Consequently, the frequencies do not adapt to the inference size $H \times W$, and phase grows purely with position, which empirically destabilizes multi-aspect-ratio generation at native 2K/4K.

Flux baseline: 2D RoPE. Following Flux, we assign rotary embeddings independently along height and width. For each axis $a \in \{H, W\}$ with channel size d_a and $m_a = d_a/2$ complex pairs, we use rotary base $b > 1$ and optional NTK factor $\eta \geq 1$ to define per-axis frequencies

$$\omega_k^{(a)} = (b \cdot \eta)^{-\alpha_k^{(a)}}, \quad \alpha_k^{(a)} = \frac{2k}{d_a}, \quad k = 0, \dots, m_a - 1, \quad (1)$$

and phases for a position $p = (p_H, p_W)$ (in patches)

$$\phi_k^{(a)}(p_a) = p_a \omega_k^{(a)}. \quad (2)$$

Writing each pair as $z_k^{(a)} = x_{2k-1}^{(a)} + i x_{2k}^{(a)}$, we apply the usual complex rotation $\tilde{z}_k^{(a)} = z_k^{(a)} e^{i \phi_k^{(a)}(p_a)}$, with wavelength (in patches) $\lambda_k^{(a)} = 2\pi/\omega_k^{(a)}$.

Resonance 2D RoPE. Motivated by the Resonance RoPE idea for train-short-test-long generalization in LLMs [33], we reinterpret the 2D rotary spectrum on a *finite* training window. Let L_H, L_W be the training-window lengths (in patches) along height and width, and let $\omega_k^{(a)}$ be the per-axis frequencies from Eq. (1). We define the number of cycles completed by component (a, k) inside the training window as

$$r_k^{(a)} = \frac{L_a}{\lambda_k^{(a)}} = \frac{L_a \omega_k^{(a)}}{2\pi}. \quad (3)$$

We then *snap* $r_k^{(a)}$ to the nearest nonzero integer:

$$\hat{r}_k^{(a)} = \max\left(1, \lfloor r_k^{(a)} + \frac{1}{2} \rfloor\right), \quad (4)$$

and replace $\omega_k^{(a)}$ by its integer-cycle projection

$$\hat{\omega}_k^{(a)} = \frac{2\pi \hat{r}_k^{(a)}}{L_a}. \quad (5)$$

The phase becomes

$$\hat{\phi}_k^{(a)}(p_a) = p_a \hat{\omega}_k^{(a)}, \quad (6)$$

and we reuse the same complex rotation as in the Flux baseline, now driven by $\hat{\phi}_k^{(a)}$. The snapped wavelength is $\lambda_k^{(a)} = 2\pi/\hat{\omega}_k^{(a)} = L_a/\hat{r}_k^{(a)}$.

Discussion. Snapping $r^{(a)} * k$ to the nearest nonzero integer turns each rotary band into a finite-window “standing wave” on $[0, L_a]$ that completes an exact integer number of cycles and has matching phase at $p_a = 0$ and $p_a = L_a$. In the original Flux spectrum, many bands traverse the training window with a fractional number of cycles, so reusing the same frequencies at larger resolutions or different ARs accumulates half-cycle phase error, which appears as spatial drift and faint striping, especially in high-frequency channels. By replacing $\omega^{(a)} * k$ in Eq. (1) with their resonant counterparts $\hat{\omega}_k^{(a)}$, Resonance 2D RoPE makes the spectrum explicitly training-window aware and prevents this fractional-cycle build-up, empirically reducing ghosting and striping artifacts when extrapolating to native-4K multi-AR grids.

Resonance 2D RoPE with YaRN. Inspired by the YaRN scheme for length extrapolation of 1D RoPE [23], we further make the extrapolation *band-aware*. Let $a \in \{H, W\}$ index spatial axes. Given the training-window length L_a (in patches) and the *resonant* frequency $\hat{\omega}_k^{(a)}$ with integer cycles $\hat{r}_k^{(a)}$ from the previous section, define the inference length L'_a and the extrapolation scale $s_a = L'_a/L_a \geq 1$. We use a linear ramp to map each band:

$$\gamma(r; \alpha, \beta) = \begin{cases} 0, & r < \alpha, \\ \frac{r - \alpha}{\beta - \alpha}, & \alpha \leq r \leq \beta, \\ 1, & r > \beta, \end{cases} \quad 0 \leq \alpha < \beta, \quad (7)$$

and interpolate between position-interpolation scaling and no scaling using the resonant cycle count:

$$\omega_{k,\text{yarn}}^{(a)} = (1 - \gamma(\hat{r}_k^{(a)}; \alpha, \beta)) \frac{\hat{\omega}_k^{(a)}}{s_a} + \gamma(\hat{r}_k^{(a)}; \alpha, \beta) \hat{\omega}_k^{(a)}. \quad (8)$$

The phase is $\phi_{k,\text{yarn}}^{(a)}(p_a) = p_a \omega_{k,\text{yarn}}^{(a)}$, and the complex rotation is identical to the Flux baseline. Compared to Flux’s fixed spectrum with a single global NTK factor, Resonance 2D RoPE with YaRN first snaps frequencies to finite-window resonant modes and then uses the axis-wise cycle counts $\hat{r}_k^{(a)}$ to decide how much to scale each band

for a given extrapolation factor s_a . This makes the positional encoding explicitly training-window aware, band-aware, and AR-aware, and empirically enables more stable 2K/4K multi-AR inference with negligible overhead.

3.2.3. SNR-Aware Huber Wavelet Training Objective

Motivation. Wavelet-space objectives such as *Diffusion4K* demonstrate that measuring errors in a multi-scale transform can materially improve 4K fidelity [40], but at native 4K we empirically observe that standard L_2 -based training on VAE latents still suffers from three coupled pathologies. (i) *Frequency imbalance*—natural-image wavelet coefficients are heavy-tailed [31], so large high-frequency residuals (textures, edges, micro-geometry) are aggressively shrunk by quadratic losses, leading to over-smoothing of detail. (ii) *Timestep imbalance*—gradients concentrate at extremely small or large noise levels, echoing Min-SNR analyses that show inefficient use of intermediate timesteps [9]. (iii) *Cross-scale energy coupling*—low-frequency bands dominate pixel/latent norms, so the high-frequency errors that largely govern 4K perceptual quality receive disproportionately small gradient signal [40]. To address these issues, we design a single objective that is simultaneously (a) *robust yet smooth*—using a Pseudo-Huber penalty that behaves like L_2 near zero and L_1 in the tails [29]; (b) *SNR-aware*—with an adaptive threshold $c(t)$ that is small under high noise and grows as signal dominates; (c) *frequency-aware*—by measuring residuals in an orthonormal wavelet space that decouples low and high bands; and (d) *time-rebalanced*—via Min-SNR weighting that emphasizes mid-SNR timesteps for stable, faster optimization [9]. These choices yield the *SNR-Aware Huber Wavelet* objective, a drop-in replacement for standard flow-matching losses tailored to the demands of native 4K generation.

Classical FM setup. Prior work on flow matching for DiTs adopts a straight-line interpolation between a clean latent z and Gaussian noise ε [2, 19]:

$$z_t = (1 - t)z + t\varepsilon, \quad t \in (0, 1), \quad \varepsilon \sim \mathcal{N}(0, I). \quad (9)$$

Under this parameterization, the DiT model predicts a velocity field $v_\theta(z_t, t)$ and the associated data-prediction is

$$\hat{z}_\theta(z_t, t) = z_t - t v_\theta(z_t, t). \quad (10)$$

To balance gradients across timesteps, we collapse the straight-path factor and Min-SNR into a single weight

$$\omega(t) = \frac{t}{1-t} \min\{\text{SNR}(t), \gamma\}^\beta. \quad (11)$$

Here $\text{SNR}(t) = (1-t)^2/t^2$ under the straight FM path, with $\gamma > 0$ and $\beta \geq 0$. We measure residuals in a wavelet space: letting $\mathcal{W}(\cdot)$ denote a one-level orthonormal DWT (sub-bands concatenated along channels), we

Table 2. Quantitative comparison under 4K res. with open-source methods.

Method	FID ↓	HPSv3 ↑	PickScore ↑	ArtiMuse ↑	CLIP Score ↑	Q-Align ↑	MUSIQ ↑
ScaleCrafter [10]	164.02	6.83	21.68	67.88	33.36	4.30	38.21
Fouriscale [12]	164.71	11.19	21.86	65.87	33.11	4.50	38.96
Sana [36]	144.17	10.83	23.18	63.72	35.49	4.89	45.08
Diffusion-4K [40]	152.43	8.92	21.88	63.76	33.00	4.69	27.51
UltraFlux	143.11	11.47	22.69	68.36	34.62	4.85	46.13

Table 3. Quantitative comparison with Sana at 4096×2048 (2:1) and 2048×4096 (1:2) resolutions.

Method	FID ↓	HPSv3 ↑	ArtiMuse ↑	Q-Align ↑
Sana (2:1)	150.35	9.01	63.61	4.80
UltraFlux (2:1)	147.53	9.91	64.81	4.86
Sana (1:2)	149.41	11.40	66.95	4.85
UltraFlux (1:2)	143.71	12.51	66.41	4.89

Table 5. Quantitative comparison under 4K res. with close-source method.

Method	FID ↓	HPSv3 ↑	PickScore ↑	ArtiMuse ↑	CLIP Score ↑	Q-Align ↑	MUSIQ ↑
Seedream 4.0 [27]	132.87	11.98	23.52	69.83	35.26	4.71	30.21
UltraFlux w. Prompt Refiner	147.06	12.03	23.25	68.75	34.50	4.93	45.93

compute the residual $R_\theta(x, \varepsilon, t) = \mathcal{W}(\hat{z}_\theta(z_t, t)) - \mathcal{W}(z)$. For robustness we use the Pseudo-Huber penalty $\rho_c(r) = c^2(\sqrt{1 + (r/c)^2} - 1)$ [29] and schedule its threshold as

$$c(t) = c_{\min} + (c_{\max} - c_{\min}) \left(\frac{\min\{\text{SNR}(t), \gamma\}}{\gamma} \right)^\alpha. \quad (12)$$

We take $\alpha \in [0, 1]$, so that $c(t)$ is small and robust at low SNR and grows smoothly toward high SNR. Let N denote the number of pixels after wavelet stacking. The per-pixel robust wavelet loss is

$$\ell_{\text{Huber}}(R_\theta; c(t)) = \frac{1}{N} \sum_{p=1}^N \rho_{c(t)}(R_{\theta,p}), \quad (13)$$

where $R_{\theta,p}$ is the p -th element of $R_\theta(x, \varepsilon, t)$. Our final objective for the DiT of our UltraFlux becomes

$$\mathcal{L}(\theta) = \mathbb{E}_{z, \varepsilon, t} [\omega(t) \ell_{\text{Huber}}(R_\theta; c(t))]. \quad (14)$$

Setting $c(t) \rightarrow \infty$ and $\beta = 0$ recovers the standard flow-matching objective on this path.

3.2.4. Stage-wise Aesthetic Curriculum Learning

Recent analyses highlight that diffusion timesteps correspond to qualitatively different generation tasks, with high-noise steps shaping global structure and low-noise steps refining local details [16, 38]. Building on this view, OmniSync [24] assigns different datasets to different timestep ranges for lip-synchronization, and several works propose timestep curricula or timestep-dependent adaptation to focus training on harder noise regimes [28, 37]. In parallel, aesthetic filtering and post-training on high-quality subsets (e.g., LAION-Aesthetics [26] and subsequent aesthetic post-training methods [18]) have proven effective for improving visual appeal, but they typically apply a static high-aesthetic prior uniformly across all timesteps.

We instead couple the noise and data axes in a simple two-stage scheme, which we term *Stage-wise Aesthetic Curriculum Learning (SACL)*. In Stage 1, we fine-tune the model on the full *MultiAspect-4K-1M* corpus with standard timestep sampling over the entire diffusion horizon, giving the DiT backbone broad coverage of aspect ratios, content types, and noise levels. Stage 2 then restricts training to

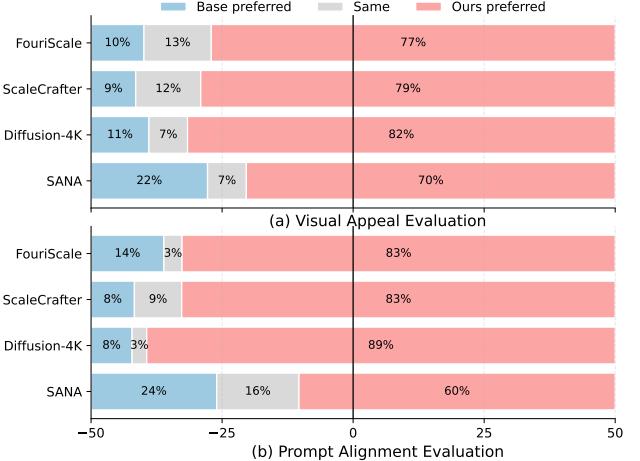


Figure 5. Gemini-2.5-Flash preference comparison.

a high-noise band—timesteps above a threshold where the model relies most on its generative prior—and to the top-5% images ranked by the ArtiMuse aesthetic score. Intuitively, Stage 1 learns a general 4K prior across all denoising tasks, while Stage 2 concentrates the remaining compute on the hardest regime with ultra high-aesthetic supervision. Unlike prior timestep curricula, which modulate timestep sampling under a fixed data distribution [16, 37, 38], or aesthetic post-training, which ignores the different roles of noise levels [18], SACL uses stage-wise aesthetic filtering to define a curriculum over high-noise timesteps, steering the global 4K generative prior toward high-aesthetic modes precisely where the sampling process is most underdetermined and yielding substantial 4K aesthetic and alignment gains at modest additional training cost (Sec. 4).

4. Experiment

4.1. Comparison with Open-source Methods.

Quantitative Comparison. As shown in Table 2, we compare our method against several strong baselines: ScaleCrafter [10] and Fouriscale [12] as representative training-free high-resolution scaling methods, Sana [36] as a recent native 4K text-to-image foundation model, and Diffusion-4K [40] as a Flux-based model trained natively at 4K resolution. All methods are evaluated on the Aesthetic-Eval@4096 benchmark [40] using FID [11], HPSv3 [20], PickScore [17], ArtiMuse [4], CLIP Score [39], Q-Align [34] and MUSIQ [15]. Beyond the square setting, Table 3 further shows that UltraFlux consistently matches or surpasses SANA across a range of non-square 4K aspect ratios, delivering better semantic alignment, and perceptual quality even in challenging panoramic and wide ARs. We also provide visual comparison as show in Figure 6.

Gemini-based Preference Evaluation. To complement automatic metrics, we conduct a large-scale pairwise preference study using *Gemini-2.5-Flash* in reasoning mode as an LMM judge. For each prompt and baseline, Gemini is shown the prompt and two anonymized images (baseline vs. UltraFlux) and asked which it prefers in terms of *visual appeal* and *prompt alignment*, with ties allowed. As sum-

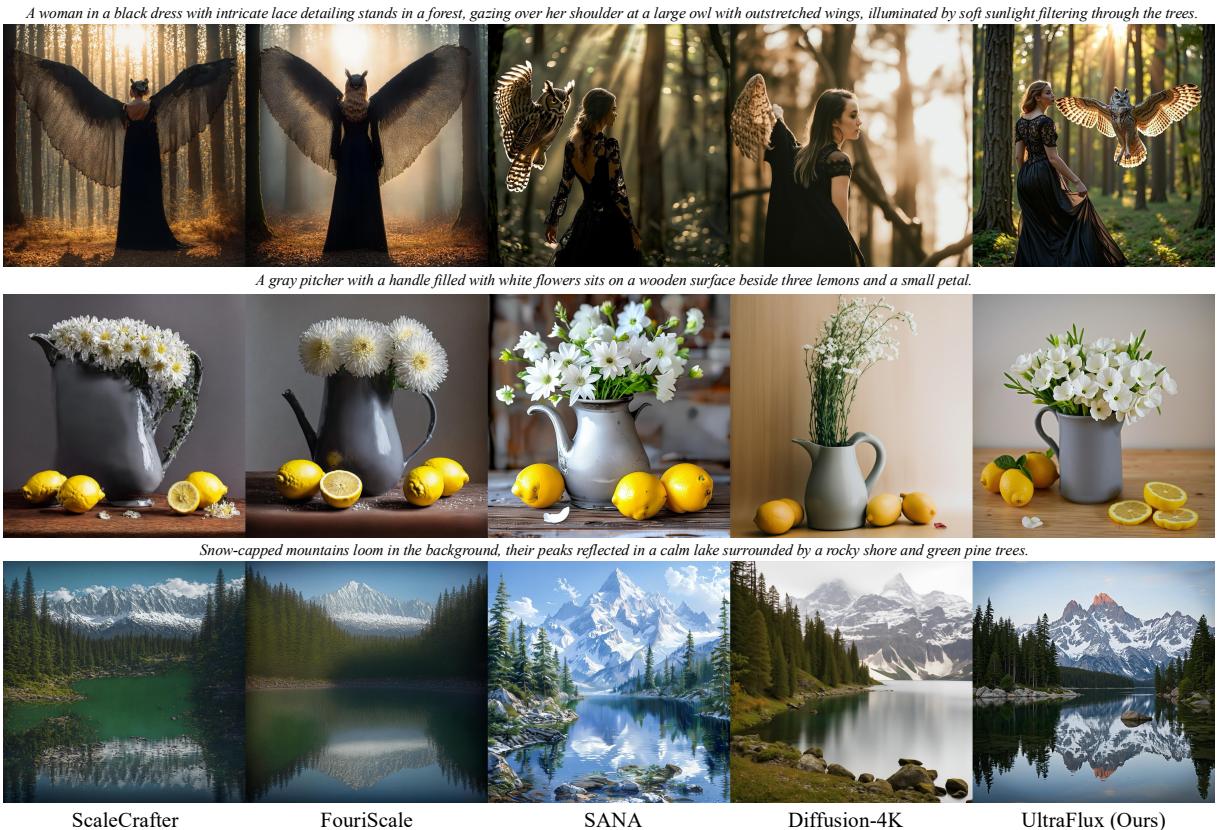


Figure 6. Visual comparison of open-source methods on the Aesthetic-Eval@4096 benchmark at 4096×4096 resolution.

marized in Figure 5, UltraFlux is preferred in 70–82% of cases on visual appeal and 60–89% on prompt alignment across all comparisons.

4.2. Comparison with Close-source 4K Native Generation Models

As shown in Table. 5 To make the comparison with *Seedream 4.0* [27] as fair as possible, we mirror its use of a large-scale LLM-based prompt refiner by evaluating *UltraFlux w. Prompt Refiner (Ours)* with a GPT-4O front-end. Under the same 4096×4096 evaluation protocol, Table 5 shows that UltraFlux with GPT-4O achieves a slightly higher HPSv3 score than Seedream 4.0 (12.03 vs. 11.98), while remaining competitive on PickScore, ArtiMuse, and CLIP Score, and clearly surpassing it on Q-Align and MUSIQ, which better capture semantic alignment and perceptual image quality. This indicates that, once both systems are equipped with strong prompt refiners, our open-source UltraFlux model trained on 1M images can closely track—and in some aspects exceed—the performance of a leading proprietary 4K generator. Notably, *Seedream 4.0* further benefits from large-scale RL post-training, whereas our full pipeline relies solely on stage-wise SFT.

4.3. Ablation Study

Starting from a Flux and trained F16 VAE baseline, replacing the standard latent regression loss with our SNR-Aware Huber Wavelet Training (SNR-HW) objective already yields consistent gains across metrics under the

Table 6. Ablation study on UltraFlux at 4K resolution training.

Variant	FID ↓	HPSv3 ↑	ArtiMuse ↑
Flux+F16 VAE (base)	151.40	9.22	66.39
+ SNR-HW	148.81	9.70	67.23
+ SNR-HW + SACL	147.32	10.30	67.31
+ SNR-HW + SACL + Resonance 2D RoPE w. YARN	146.93	10.91	68.13

same 500K data&10K steps fine-tuning schedule, indicating that SNR-aware wavelet supervision better balances high-frequency detail preservation with stable optimization. Introducing the SACL term on top of SNR-HW further improves both human preference and aesthetic scores, suggesting that stronger text–image alignment is especially beneficial at native 4K. Finally, equipping UltraFlux with Resonance 2D RoPE and YARN produces the best overall configuration, delivering monotonically improved perceptual and aesthetic metrics while also reducing FID. Taken together, these ablations show that the proposed objective, alignment loss, and positional encoding contribute complementary gains rather than merely redistributing performance among metrics.

5. Conclusion

UltraFlux couples a carefully curated MultiAspect-4K-1M corpus with AR-aware positional encoding, reconstruction, and optimization components into a unified framework for native 4K multi-AR generation. This data–model co-design

yields state-of-the-art fidelity, aesthetic quality, and text alignment on standard 4K benchmarks while remaining computationally practical.

References

- [1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. [2](#), [3](#)
- [2] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. [2](#), [5](#), [6](#)
- [3] Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhui Wang, Yu Qiao, Dajun Yao, and Yihao Liu. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. 2025. [4](#), [2](#)
- [4] Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, et al. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*, 2025. [7](#)
- [5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- σ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. [2](#), [3](#), [4](#)
- [6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. [2](#)
- [7] Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, et al. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. In *European conference on computer vision*, pages 39–55. Springer, 2024. [2](#), [3](#)
- [8] Moayed Haji-Ali, Guha Balakrishnan, and Vicente Ordonez. Elasticdiffusion: Training-free arbitrary size image generation through global-local content separation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6603–6612, 2024. [2](#), [3](#)
- [9] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7441–7451, 2023. [2](#), [6](#)
- [10] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. [7](#)
- [11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. [7](#)
- [12] Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *European conference on computer vision*, pages 196–212. Springer, 2024. [2](#), [3](#), [7](#)
- [13] Ideogram. Upscale - ideogram documentation. <https://docs.ideogram.ai/using-ideogram/features-and-tools/upscale>, 2025. Accessed: 2025-11-10. [3](#)
- [14] Jinho Jeong, Sangmin Han, Jinwoo Kim, and Seon Joo Kim. Latent space super-resolution for higher-resolution image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2355–2365, 2025. [2](#), [3](#)
- [15] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. [7](#)
- [16] Jin-Young Kim, Hyojun Go, Soonwoo Kwon, and Hyun-Gyo Kim. Denoising task difficulty-based curriculum for training diffusion models. *arXiv preprint arXiv:2403.10348*, 2024. [7](#)
- [17] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023. [7](#)
- [18] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13199–13208, 2025. [7](#)
- [19] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. [6](#)
- [20] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score, 2025. [7](#)
- [21] MidJourney. Midjourney: Ai image generation. <https://www.midjourney.com/>, 2025. Accessed: 2025-11-10. [3](#)
- [22] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. [2](#)
- [23] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023. [2](#), [6](#)
- [24] Ziqiao Peng, Jiwen Liu, Haoxian Zhang, Xiaoqiang Liu, Songlin Tang, Pengfei Wan, Di Zhang, Hongyan Liu, and

- Jun He. Omnisync: Towards universal lip synchronization via diffusion transformers. *arXiv preprint arXiv:2505.21448*, 2025. 7
- [25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3
- [26] Christoph Schuhmann. Laion-aesthetics. <https://laion.ai/blog/laion-aesthetics/>, 2022. Accessed: 2025-11-13. 7
- [27] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025. 3, 7, 8
- [28] Vera Soboleva, Aibek Alanov, Andrey Kuznetsov, and Konstantin Sobolev. T-lora: Single image diffusion model customization without overfitting. *arXiv preprint arXiv:2507.05964*, 2025. 7
- [29] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 6, 7
- [30] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Rofomer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5
- [31] Martin J Wainwright and Eero Simoncelli. Scale mixtures of gaussians and the statistics of natural images. *Advances in neural information processing systems*, 12, 1999. 6
- [32] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything, 2025. 4
- [33] Suyuchen Wang, Ivan Kobyzev, Peng Lu, Mehdi Rezagholizadeh, and Bang Liu. Resonance rope: Improving context length generalization of large language models. *arXiv preprint arXiv:2403.00071*, 2024. 5
- [34] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching Imms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi. 7
- [35] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching Imms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi. 4, 2
- [36] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 2, 3, 7
- [37] Tianshuo Xu, Peng Mi, Ruilin Wang, and Yingcong Chen. Towards faster training of diffusion models: An inspiration of a consistency phenomenon. *arXiv preprint arXiv:2404.07946*, 2024. 7
- [38] Xuanyu Yi, Zike Wu, Qingshan Xu, Pan Zhou, Joo-Hwee Lim, and Hanwang Zhang. Diffusion time-step curriculum for one image to 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9948–9958, 2024. 7
- [39] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024. 7
- [40] Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23464–23473, 2025. 2, 3, 4, 5, 6, 7
- [41] Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Ultra-high-resolution image synthesis: Data, method and evaluation. *arXiv preprint arXiv:2506.01331*, 2025. 1, 2
- [42] Kechi Zhang, Ge Li, Huangzhao Zhang, and Zhi Jin. Hirope: Length extrapolation for code models using hierarchical position. *arXiv preprint arXiv:2403.19115*, 2024. 2
- [43] Shen Zhang, Zhaowei Chen, Zhenyu Zhao, Yuhan Chen, Yao Tang, and Jiajun Liang. Hidiffusion: Unlocking higher-resolution creativity and efficiency in pretrained diffusion models. In *European Conference on Computer Vision*, pages 145–161. Springer, 2024. 2, 3
- [44] Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. Hunyuan-mt technical report. *arXiv preprint arXiv:2509.05209*, 2025. 4

UltraFlux: Data-Model Co-Design for High-quality Native 4K Text-to-Image Generation across Diverse Aspect Ratios

Supplementary Material

Supplementary Overview

This document complements the main paper by (i) clarifying the regime-level novelty of *UltraFlux* as a data–model co-designed system for native 4K, multi-AR text-to-image generation, and (ii) providing the analyses, metrics, and implementation details needed to faithfully reproduce and stress-test our setup. Rather than introducing isolated primitives, we make explicit how dataset curation, positional representation, VAE compression, and optimization objectives must be co-designed to operate robustly in the 4K, multi-AR regime.

Organization. Sec. S1 (*Clarifying Novelty and Data-Model Co-Design*) positions UltraFlux as a regime-level, system-oriented contribution and introduces a 2×2 data–model ablation that disentangles the roles of MULTIAspect-4K-1M and the UltraFlux architecture/loss. Sec. S2 (*Implementation Details*) expands on the dataset pipeline, DiT training schedule, VAE post-training recipe, and key hyperparameters, including reconstruction metrics for our F16 VAE. Sec. S3 and Sec. S4 provide focused analyses of Resonance 2D RoPE with YaRN and wavelet-space statistics of 4K VAE latents, together with qualitative 4K visualizations, motivating our positional design and SNR-Aware Huber Wavelet objective. Sec. S5–S7 report additional ablations, 4K runtime measurements, and more extensive quantitative comparisons at challenging wide aspect ratios, as well as extended visual comparisons against open-source baselines. Sec. S8 discusses the main limitations of UltraFlux in terms of sampling cost, memory footprint, and aesthetic ceiling, while Sec. S9 details our Gemini-based preference evaluation and GPT-4o prompt-refiner setup used for large-scale automatic assessment and prompt expansion.

Taken together, these sections are intended to show that UltraFlux is a *regime-level, system-oriented contribution* rather than a mere aggregation of existing tricks, and to document the concrete choices required to make native-4K, multi-AR generation work in practice.

6. Clarifying Novelty and Data-Model Co-Design

The main paper positions UltraFlux as a *data-model co-designed recipe* for native-4K, multi-AR text-to-image generation. Several of the building blocks—resonance-style rotary encodings, wavelet objectives, Min-SNR weighting, and aesthetic curricula—indeed draw inspiration from prior

Table 7. 2×2 data–model co-design ablation. A: baseline; B: data only; C: model/loss only; D: full co-design.

Variant	Dataset	Model / Loss	FID \downarrow	HPSv3 \uparrow
A	Diffusion-4K-v2 [41]	Flux, latent L2	152.09	8.57
B	MultiAspect-4K-1M	Flux, latent L2	151.41	9.17
C	Diffusion-4K-v2 [41]	UltraFlux	147.41	10.03
D	MultiAspect-4K-1M	UltraFlux	145.81	10.78

work. Our contribution is not to claim each primitive as a standalone invention, but to show that: (i) at 4K with diverse aspect ratios, positional encoding, VAE compression, and optimization objectives form a *coupled regime* that existing methods treat largely in isolation; and (ii) a carefully unified design across *dataset, representation, and loss* yields behaviors that cannot be reproduced by swapping in any single component in isolation.

To make this clearer, we provide in this supplement:

- A data–model ablation (Table 7) showing that neither a stronger 4K dataset nor architectural changes alone are sufficient: MultiAspect-4K-1M and UltraFlux each yield modest gains in isolation, while their combination delivers the full non-additive improvements in 4K, multi-AR fidelity.
- One-dimensional and two-dimensional diagnostics of Resonance 2D RoPE with YaRN (Sec. 8), analyzing cycle snapping, phase closure on the training window, and the stability of phase geometry under aspect-ratio extrapolation.
- Wavelet-space statistics of 4K VAE latents (Sec. 9) that empirically confirm the low-frequency-dominated yet heavy-tailed structure motivating our SNR-Aware Huber Wavelet objective, clarifying why a robust, SNR-aware wavelet loss is better aligned with the 4K regime than a pure latent L_2 objective.
- Expanded implementation details for the dataset pipeline, DiT training, and VAE post-training (Sec. S2), to facilitate faithful reproduction of our 4K native, multi-AR training setup.

We hope these analyses better convey that UltraFlux is a *regime-level, system-oriented contribution* rather than a mere aggregation of existing tricks.

7. Implementation Details

7.1. Dataset Pipeline

Flat-region detection. For each image, we first partition it into non-overlapping 240×240 patches and quantify the

edge richness of every patch with a Sobel-based score,

$$S_{\text{flat}} = \text{Var} \left(\sqrt{(\partial_x I)^2 + (\partial_y I)^2} \right).$$

Patches with $S_{\text{flat}} < 800$ are flagged as texture-poor, and any image in which more than 50% of the patches are flagged is removed from the dataset. The patch-level threshold of 800 and the 50% image-level ratio are selected empirically via manual inspection of edge-statistic histograms and visual audits. This conservative configuration effectively filters out images dominated by large uniform regions while still retaining plausible low-texture content such as sky and water, ensuring that the remaining images maintain sufficient edge and texture diversity for high-fidelity generation.

Information Entropy Filtering. Each image is analyzed for its Shannon entropy to quantify the amount of information it contains. The Shannon entropy H of an image is defined as:

$$H = - \sum_{i=1}^N p(x_i) \log_2 p(x_i),$$

where $p(x_i)$ denotes the probability of the pixel value x_i within the image. Images with an entropy value $H < 7.0$ are flagged as texture-poor, and any image in which $H < 7.0$ is removed from the dataset. The threshold of 7.0 is selected empirically based on the observed distribution of entropy values across the dataset. This threshold effectively filters out images with insufficient texture or information, ensuring that the remaining images exhibit adequate variability for high-quality processing while preserving content diversity.

Image Quality Filtering. To ensure semantic quality, we compute the quality score for each image using *Q-Align* [35]. Images with a quality score greater than 4.0 are retained, while those below this threshold are discarded. This threshold is determined empirically based on the distribution of quality scores across data sources, ensuring that only images with sufficient semantic clarity are kept for further analysis.

Aesthetic Quality Filtering. For aesthetic evaluation, we use the *ArtiMuse* [3] model to compute aesthetic scores for each image. Only the top 30% of images, based on their aesthetic rating, are preserved. This strategy ensures that images with higher aesthetic appeal are prioritized, while lower-rated images are excluded from the dataset. This filtering method helps maintain a diverse and aesthetically pleasing selection of images for further processing.

7.2. Training Details

DiT Training. We train *UltraFlux*, a large Flux-based DiT model for native 4K text-to-image generation. During DiT training, we freeze the VAE and text encoders and

Table 8. Reconstruction metrics of F16 VAEs on the Aesthetic-4K@4096 Eval set [40].

Model	rFID ↓	NMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Flux-VAE-F16 [40]	2.201	0.01522	26.90	0.784	0.168
Flux-VAE-F16-SC [41]	0.588	0.00736	30.19	0.846	0.097
UltraFlux-F16-VAE	0.547	0.00657	30.70	0.852	0.102

update all DiT blocks end-to-end. Training is conducted on 8×NVIDIA H800 GPUs using DeepSpeed ZeRO-2 (without CPU offload). We choose ZeRO-2 because it shards optimizer states and gradients without partitioning model parameters, which substantially reduces memory usage while yielding higher throughput than ZeRO-3 in our setting, enabling efficient 4K training. We use AdamW with a learning rate of 1×10^{-6} and an effective batch size of 64; the full training run takes roughly 12 days. We adopt a two-stage training schedule, with approximately 30K steps in the first stage and a further 2K steps in the second fine-tuning stage (Stage-wise Aesthetic Curriculum Learning). To support multi-AR native 4K generation, we adopt a bucketed resolution scheme: for each image, we snap its resolution to the nearest target from a fixed set of landscape buckets (e.g., 5120×2880 for 16:9, 4704×3136 for 3:2), portrait buckets (e.g., 2880×5120 for 9:16, 3136×4704 for 2:3), and a single square bucket at 3840×3840 , then center-crop and resize the image to the selected bucket resolution.

VAE Training. For VAE post-training, we fine-tune the decoder on the proposed *MultiAspect-4K-1M* dataset, retaining the top 50% of images according to the flatness score and training at 512×512 resolution with an effective batch size of 384. We use AdamW with a learning rate of 1×10^{-5} .

VAE reconstruction metrics and post-training gains. Table 8 quantitatively compares our **UltraFlux-F16-VAE** with the Flux-VAE-F16 baseline on the AESTHETIC-4K@4096 evaluation set [40]. Despite using the same F16 compression ratio, UltraFlux-F16-VAE achieves substantially better reconstruction quality across all metrics. These consistent gains indicate that our post-trained decoder not only preserves low-frequency structure, but also better reconstructs high-frequency details that are typically washed out under aggressive F16 compression. Combined with the wavelet-space analysis above, this suggests that the proposed post-training scheme effectively aligns the VAE with the heavy-tailed, cross-scale statistics of native 4K images, narrowing the reconstruction gap.

8. Analyses of Resonance 2D RoPE with YaRN

Figure 7 gives a 1D band-wise diagnostic of Resonance 2D RoPE on a single spatial axis, which is then used by YaRN. In panel (a), we plot the cycles completed in the training

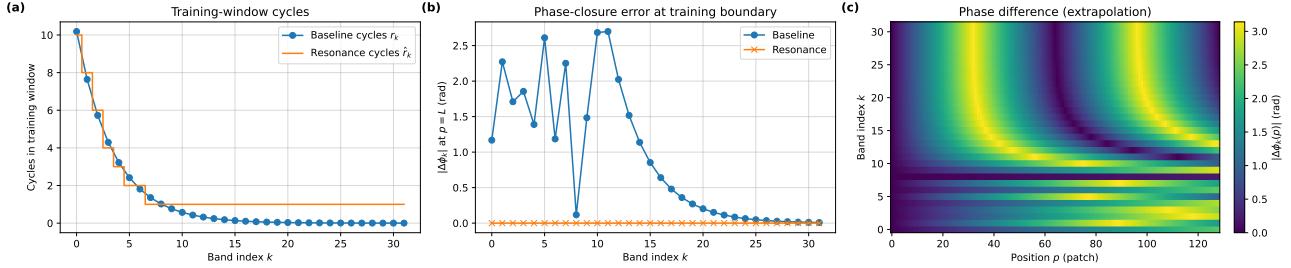


Figure 7. 1D band-wise analysis of Resonance 2D RoPE with YaRN. (a) Number of cycles r_k in the training window and their integer-snapped counterparts \hat{r}_k . (b) Phase-closure error $|\Delta\phi_k|$ at $p = L$, showing exact closure for Resonance RoPE. (c) Phase difference $|\Delta\phi_k(p)|$ between baseline and Resonance under $2\times$ length extrapolation, illustrating how fractional cycles in the baseline accumulate into large out-of-distribution phase errors.

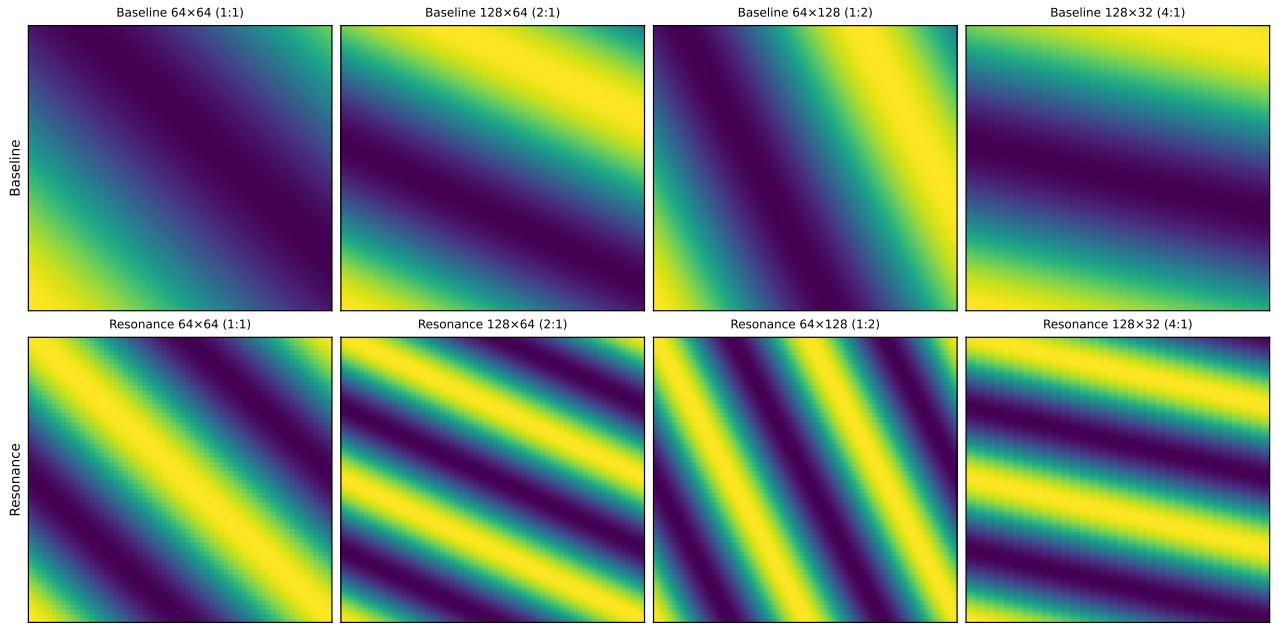


Figure 8. Height–width cosine phase patterns for a representative rotary band under different aspect ratios. Each panel displays $f(h, w) = \cos(h\omega_H + w\omega_W)$ evaluated on an $H \times W$ patch grid, with the top row using the Flux baseline frequencies and the bottom row using our resonant frequencies. The first column shows the training resolution (64×64 patches, AR 1:1); the remaining columns visualize extrapolation to 128×64 (2:1), 64×128 (1:2), and 128×32 (4:1). At train scale, baseline and Resonance RoPE induce similar diagonal stripe patterns, indicating that resonance acts as a mild reparameterization of the spectrum. Under aspect-ratio extrapolation, the baseline stripes rotate and change spacing more aggressively, while our resonant variant maintains more regular, coherent patterns along both height and width, qualitatively echoing the improved phase stability observed in the 1D analyses.

window of length L_a by each rotary band,

$$r_k^{(a)} = \frac{L_a \omega_k^{(a)}}{2\pi},$$

together with their snapped counterparts

$$\hat{r}_k^{(a)} = \max(1, \text{round}(r_k^{(a)})).$$

The Flux baseline (blue) yields a dense sequence of non-integer $r_k^{(a)}$, whereas Resonance RoPE (orange) projects ev-

ery band onto the nearest nonzero integer $\hat{r}_k^{(a)}$, producing a piecewise-constant spectrum that leaves low-frequency modes almost unchanged and regularizes higher ones.

Panel (b) measures phase closure at the boundary of the training window. For each band we evaluate the phase at $p_a = L_a$ using both the original frequency $\omega_k^{(a)}$ and the resonant frequency

$$\hat{\omega}_k^{(a)} = \frac{2\pi \hat{r}_k^{(a)}}{L_a},$$

Table 9. Summary of training-related hyperparameters for UltraFlux and associated components. Values are left blank to be filled with the final configuration.

Component	Hyperparameter	Value
Stage-wise Aesthetic Curriculum	Stage 1 timestep range	0–999
	Stage 2 timestep range	0–459
	Stage 2 aesthetic filter (AriMuse percentile)	top-5%
DiT objective	Wavelet type / number of levels Pseudo-Huber thresholds (c_{\min}, c_{\max})	Haar, $J=1$ $c_{\min} \approx 0.2, c_{\max} \approx 1.0$
Resonance 2D RoPE with YaRN	RoPE base b	10,000
	NTK scaling factor η	1.0
	YaRN ramp parameters (α, β)	(1.25, 0.75)
F16 VAE post-training	Maximum extrapolation scale $s_a = L'_a/L_a$	2.0
	Training resolution	512 × 512
	Global batch size (images/step)	384
	Optimizer / learning rate / weight decay	AdamW, $1 \times 10^{-4}, 1 \times 10^{-2}$
Multi-AR 4K DiT training	Loss weights $(\lambda_{\text{wave}}, \lambda_{\text{perc}}, \lambda_{L_2})$	0.2, 0.1, 1
	Landscape target sizes (W×H)	5440×3072, 5184×3264, 4992×3328 4736×3520, 5824×2880, 6272×2688
	Portrait target sizes (W×H)	5568×3008, 6336×2624, 5632×3008 4608×3648 3072×5440, 3648×4608, 3520×4736
	Square target sizes (W×H)	3328×4992 4096×4096

and plot the absolute phase mismatch $|\Delta\phi_k|$ between $p_a = 0$ and $p_a = L_a$. The baseline shows up to several radians of mismatch, while Resonance RoPE drives $|\Delta\phi_k|$ to zero for all bands, confirming that every component becomes an exact standing wave on $[0, L_a]$.

Panel (c) visualizes the phase difference between the baseline and Resonance RoPE under a $2\times$ resolution extrapolation. For positions $p_a \in [0, 2L_a]$ we compute

$$\Delta\phi_k(p_a) = \text{wrap}(p_a \omega_k^{(a)} - p_a \hat{\omega}_k^{(a)}),$$

where $\text{wrap}(\cdot)$ maps angles to $[-\pi, \pi]$, and plot $|\Delta\phi_k(p_a)|$ as a heatmap over (k, p_a) . The discrepancy is small near the training window but grows systematically with both position and frequency, illustrating how fractional cycles in the original spectrum accumulate into large out-of-distribution phase errors. Since YaRN subsequently applies band-wise scaling to these already integer-cycle-aligned modes, the combined Resonance 2D RoPE with YaRN inherits training-window awareness while achieving stable, AR-robust extrapolation in 2D.

2D spatial visualization. Figure 7 analyzes Resonance 2D RoPE with YaRN along a single spatial axis. To understand how these band-wise changes translate into actual image-plane geometry, we further visualize 2D cosine patterns in Figure 8. For a representative rotary band, we construct

$$f(h, w) = \cos(h \omega_H + w \omega_W),$$

on different height-width grids, where (ω_H, ω_W) are taken either from the Flux baseline or from the resonant frequencies. The leftmost column corresponds to the training resolution (64×64 patches, AR 1:1), while the remaining columns show extrapolation to 128×64 (2:1), 64×128 (1:2), and 128×32 (4:1). At the training scale, baseline and Resonance RoPE produce very similar diagonal stripe

patterns, consistent with the fact that snapping r_k to \hat{r}_k only slightly perturbs low-frequency modes. Across more extreme aspect ratios, however, the baseline stripes exhibit more pronounced changes in orientation and spacing, whereas the Resonance patterns remain more regular and coherent. This 2D view complements the 1D diagnostics: once each band forms an integer-cycle standing wave on the training window, spatial phase geometry varies more smoothly when scaling to multi-AR 2K/4K grids.

9. Wavelet-Space Statistics of 4K VAE Latents

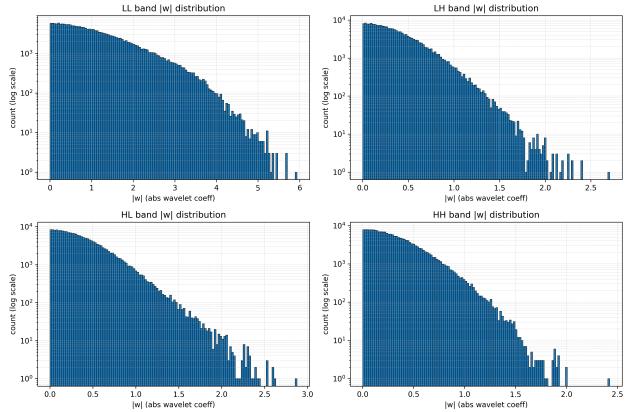


Figure 9. Wavelet-space statistics of 4K VAE latents. We show log-count histograms of absolute coefficients for the LL, LH, HL, and HH subbands over 400 samples. Most energy resides in the LL band, while high-frequency bands carry sparse but large-magnitude coefficients, indicating heavy-tailed behavior. This cross-scale structure motivates our SNR-Aware Huber Wavelet objective.

In the main paper (Section 3.2.3, *SNR-Aware Huber Wavelet Training Objective*) we argue that native 4K generation suffers from **(i) frequency imbalance and (ii) cross-scale energy coupling: low-frequency bands dominate latent norms, while high-frequency, perceptually critical structures appear as sparse, large-magnitude coefficients that are poorly handled by purely quadratic losses**. Here we provide an empirical characterization of this effect in the VAE latent space used by UltraFlux. We sample 400 images from MULTIAspect-4K-1M, encode them with our F16 VAE, and apply a one-level orthonormal DWT to the resulting latents. Figure 9 shows log-count histograms of the absolute wavelet coefficients in the LL, LH, HL, and HH subbands. The energy distribution is strongly skewed across scales: the LL band accounts for **87.4%** of the total latent energy (mean per-band energy 3.55), while each high-frequency band contributes only 3.5–4.7%. At the same time, all bands exhibit pronounced heavy tails. For example, in the LH band 20.8% of coefficients satisfy $|w| > 0.5$,

Table 10. Inference time per 4K sample at 4096×4096 resolution.

	ScaleCrafter	FouriScale	Sana	UltraFlux
Time (s)	195.67	216.27	48.42	49.50

3.2% exceed $|w| > 1.0$, and values up to $|w| \approx 7.2$ occur; HL and HH show similar tail behavior.

These statistics quantitatively support the motivation in the main paper: at 4K resolution, VAE latents are dominated by low-frequency energy, yet contain sparse, large-magnitude high-frequency coefficients that encode textures, edges, and micro-geometry. Under a standard L_2 latent loss, these heavy-tailed residuals are aggressively shrunk, and gradients are largely governed by the LL band, leading to over-smoothing of detail and weak supervision for high-frequency errors. This empirical evidence motivates our design of the SNR-Aware Huber Wavelet objective, which replaces pure L_2 with a wavelet-space, Pseudo-Huber penalty with SNR-dependent thresholds to better balance low- and high-frequency reconstruction errors in the 4K regime.

10. Additional Ablations

Figure 10 provides a visual counterpart to the analyses in Sec. 8. The Flux.1 2D RoPE baseline without scaling (a) reuses its training-time spectrum at 4K and produces noticeable geometric drift: objects appear slightly stretched or misaligned and backgrounds show faint striping. Introducing YaRN scaling alone (b) reduces these artifacts by making the spectrum resolution-aware, but residual phase misalignment still leads to mild warping along long contours. Our Resonance 2D RoPE with YaRN (c) first snaps each band to an integer-cycle standing wave on the training window and then applies band-wise YaRN scaling, yielding visibly more stable composition and cleaner high-frequency details, especially in the delicate structures of fur, foliage, and planetary rings.

11. Efficiency Comparison

Table 10 reports the wall-clock time required for each method to generate a single 4096×4096 sample under the same hardware and sampler configuration. UltraFlux and Sana operate in a similar runtime regime, while both are several times faster than ScaleCrafter and FouriScale, whose 4K pipelines incur substantially higher latency. In other words, UltraFlux achieves our best 4K fidelity and aesthetic metrics without introducing extra inference cost relative to the strongest open baseline, and remains markedly more efficient than earlier 4K upsampling-based approaches.

Table 11. Quantitative comparison with SOTA methods at different aspect ratios, including 4096×2048 (2:1), 2048×4096 (1:2), 5120×2880 (16:9), and 5952×2496 (2.39:1) resolutions.

Aspect Ratio	Method	FID \downarrow	HPSv3 \uparrow	Artimuse \uparrow	Q-Align \uparrow
2:1	ScaleCrafter	168.29	6.26	65.62	4.29
	FouriScale	169.30	5.89	64.29	4.38
	Sana	150.36	9.01	63.61	4.81
	UltraFlux	147.54	9.91	64.81	4.86
1:2	ScaleCrafter	157.21	8.92	68.74	4.41
	FouriScale	159.87	8.09	66.64	4.38
	Sana	149.42	11.40	66.95	4.86
	UltraFlux	143.71	12.51	66.41	4.89
16:9	ScaleCrafter	175.97	5.30	65.05	4.14
	FouriScale	173.84	5.46	64.14	4.36
	Sana	153.31	9.04	63.02	4.82
	UltraFlux	142.43	9.92	67.22	4.85
2.39:1	ScaleCrafter	196.60	3.69	64.02	4.00
	FouriScale	196.30	3.61	63.09	4.14
	Sana	153.10	8.57	62.48	4.77
	UltraFlux	151.99	11.76	66.36	4.82

12. More Quantitative Comparison with SOTA Methods at Wide Aspect Ratios

To provide a more comprehensive evaluation of performance at challenging wide aspect ratios, including 2:1 (4096×2048), 1:2 (2048×4096), 16:9 (5120×2880), and the cinematic 2.39:1, we compare with SOTA methods across four distinct aspect ratios and resolutions, as detailed in Table 11. The results show that UltraFlux consistently surpasses the performance all competing methods across all tested aspect ratios and metrics, demonstrating its effectiveness in generating high-quality images for diverse wide-format scenarios.

13. Additional Visual Comparison With Open-Source methods

In Figures 11–13, additional visual comparisons are provided. From the results, we observe that ScaleCrafter sometimes produces images with noticeable distortions, while FouriScale occasionally struggles to fully capture textual content. The images generated by SANA, on the other hand, can appear somewhat overly smoothed or "oily." In contrast, compared to Diffusion-4K, our method consistently delivers higher-quality images with more visually appealing results, offering a more pleasant overall experience.

14. Limitations

Although UltraFlux substantially improves native-4K, multi-AR generation over prior open-source baselines, the system still has several practical limitations.

Sampling cost and memory footprint. First, UltraFlux is not yet a *efficient* 4K generator. Even with the F16 VAE and our optimized DiT backbone, sampling at na-



Figure 10. **Qualitative effect of Resonance 2D RoPE with YaRN.** We compare three positional encodings at native 4K resolution for the same prompts. (a) Flux.1 2D RoPE baseline *without* any scaling at inference time, which tends to exhibit geometric drift and mild striping or warping artifacts in both foreground objects and backgrounds. (b) 2D RoPE with YaRN scaling, which stabilizes the overall layout but still shows subtle distortions along long contours and in extreme regions of the image. (c) Our proposed *Resonance 2D RoPE with YaRN*, which yields the most coherent global geometry and sharper, more regular fine structures (e.g., ring edges and tree trunks).

tive 4K with 50–60 flow-matching steps remains noticeably slower than 1K-class models and requires a high-end 50GB GPU to avoid aggressive offloading. This compute and memory footprint limits deployment to research- or data-center-grade hardware, and makes large-scale 4K sampling expensive compared to lower-resolution pipelines or distilled student models.

Aesthetic ceiling and robustness. Second, while our data–model co-design delivers consistent gains in automatic metrics and Gemini-based preference studies, the aesthetic quality is not uniformly top-tier across all prompts and domains. In challenging cases, UltraFlux can still produce occasional over-smoothed textures, minor geometric artifacts, or compositions that are less polished than those from heavily engineered proprietary systems. Our co-design focuses on the 4K + multi-AR regime rather than absolute peak aesthetics, and there remains headroom for further preference alignment, prompt understanding, and content diversity.

Scope of co-design. Finally, the present work primarily co-designs dataset, positional encoding, VAE, and loss under a single large DiT backbone. We do not address complemen-

tary axes such as sparse or low-rank attention, lightweight decoders, or distillation to smaller 4K models, which could significantly reduce memory usage and latency. Extending UltraFlux-style co-design to more parameter-efficient architectures and to broader data domains (e.g., specialized scientific or medical imagery) is an important direction for future work.

15. Details About Gemini-based Preference Evaluation.

In this section, we provide additional details on the Gemini-based preference evaluation used to assess the visual quality and prompt alignment of different models. As part of this evaluation, Gemini-2.5-Flash, in reasoning mode, is employed to judge image pairs based on their aesthetic appeal and alignment with the given prompt. The following is an example of the exact prompt used for evaluating *aesthetic preferences* in our study. For each image pair, Gemini is asked to assess various aspects such as composition, sharpness, lighting, and overall visual appeal, ensuring that the evaluation process is both consistent and reproducible.

Prompt 15.1 (Pairwise Preference for Aesthetics)

You are an impartial image aesthetics judge. Compare Image A and Image B, and decide which one better fits human aesthetic preferences overall.

Evaluate:

- Composition
- Sharpness / clarity
- Lighting / contrast
- Color harmony
- Noise / compression artifacts
- Overall visual appeal

Be decisive; only return "tie" if the two images are nearly identical in quality.

Return strictly in the following JSON format (no explanations, no extra text):

```
{  
    "preferred": "A | B | tie",  
    "a_score": 0-100,  
    "b_score": 0-100,  
    "reasons": "short explanation"  
}
```

16. Details About Prompt Refiner using GPT-4O.

Prompt 16.1 (GPT-4O Prompt Refining Process)

System prompt: You are a senior prompt refiner for AI image generation. Expand each short prompt into a single rich, high-aesthetic prompt.

Requirements:

- Length: 55–100 words; one line per item; no new-lines, numbering, or quotes.
- Preserve the original subject and intent; do not invent brands, copyrighted IP, or named people.
- Composition: camera angle, shot size/framing, focal length or lens type, foreground/midground/background, environment context.
- Subject attributes: age range, gender expression where implied, appearance details (hair/eyes/skin or material), clothing/fabric, pose, expression/action.
- Lighting and color: key light quality/direction, color temperature, time of day/season/weather, palette or dominant hues.
- Style/medium: photographic or cinematic unless the input implies another medium; mention film look or post-processing if appropriate.
- Quality: tasteful, coherent, non-repetitive language; avoid keyword stuffing.

User prompt: Short prompts: {list of input prompts}

Language: Write outputs in *[language]* (Chinese/English); one line per item. For each input, produce exactly one refined prompt; avoid lists, bullets, or line breaks inside items.

Expected output format:

```
[  
    "Refined prompt 1",  
    "Refined prompt 2",  
    ...  
]
```

To further refine the quality of input prompts, we employ GPT-4o as a front-end for our *UltraFlux w. Prompt Refiner (Ours)* configuration. The process of prompt refinement involves transforming short and concise prompts into more detailed, high-aesthetic descriptions suitable for image generation tasks. The GPT-4o model expands each input prompt into a rich description, incorporating essential elements such as composition, lighting, subject attributes, and stylistic choices. The refined prompts follow a strict set of guidelines to maintain coherence, clarity, and aesthetic quality, ensuring that they meet the requirements for high-fidelity image generation. In the following example, we provide the exact system prompt and user instructions used to guide GPT-4o in refining a list of short prompts. The prompts are designed to ensure that the model generates vi-

sually appealing and contextually appropriate descriptions for each input. This prompt refining process ensures that the generated prompts are detailed, high-quality, and aligned with the intended visual aesthetics. The use of GPT-4o to refine short prompt significantly enhances the input quality, making it suitable for use in high-fidelity image generation tasks.

A woman with a chic hairstyle wears elegant eyeglasses and multiple pieces of sparkling jewelry, including earrings and layered necklaces, against a textured gray background.



Blonde model with tousled hair wearing a form-fitting olive green dress, looking intently at the camera with a soft expression.



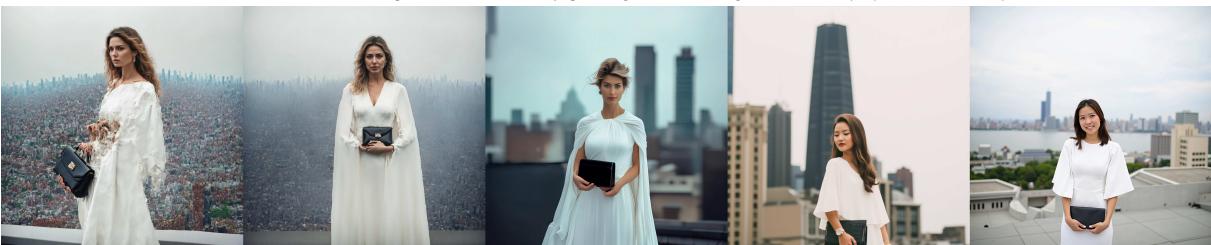
An older man in a dark suit and fedora sits casually on a balcony, wearing sunglasses and holding a cigarette, with a serious expression against a muted background.



A ballerina in a flowing yellow dress performs an elegant pose on one leg, showcasing her strength and grace against a backdrop of geometric wall art and soft natural light.



A woman in a white dress with cape sleeves stands on a rooftop holding a black clutch, against a blurred city skyline and overcast sky.



A woman in a light blue dress stands at a balcony, holding onto the door frame, with a view of a serene coastal landscape featuring mountains and a peaceful shoreline.



ScaleCrafter

Fouriscale

SANA

Diffusion-4K

UltraFlux (Ours)

Figure 11. More visual comparison of open-source methods on the Aesthetic-Eval@4096 benchmark at 4096×4096 resolution.

A still life composition featuring two blue glass bottles next to a black bowl filled with several pears, set against a textured backdrop with warm tones.



A surfer rides a wave beneath the surface, showcasing skill and agility on a yellow board, with coral formations visible below.



A grand interior with soaring arches, intricate stonework, and vibrant stained glass windows, illuminated by ornate chandeliers and soft light, creating an atmosphere of reverence and awe.



A majestic waterfall cascades into a lush, green valley, surrounded by dense forest and towering mountains, with a mist hovering over the lower areas and a volcano rising in the background under a soft sky.



Snow-capped mountains loom in the background, their peaks reflected in a calm lake surrounded by a rocky shore and green pine trees.



A woman in a vintage, elegantly tailored gown with intricate embroidery gazes thoughtfully over a balcony, with a scenic river and historic buildings in the background.



ScaleCrafter

FourIScale

SANA

Diffusion-4K

UltraFlux (Ours)

Figure 12. More visual comparison of open-source methods on the Aesthetic-Eval@4096 benchmark at 4096×4096 resolution.

A dramatic landscape featuring majestic, snow-capped mountains under a colorful rainbow, with dark, sandy terrain dotted with patches of grass.



A vibrant night sky filled with a milky way galaxy and swirling colors, reflected in a still, steaming geothermal pool surrounded by dark silhouettes of trees and misty atmosphere.



A picturesque riverside scene featuring a medieval castle atop a hill, surrounded by vibrant autumn foliage, with colorful homes lining the waterfront and several boats docked along the river.



Two individuals set up a tent on a grassy hillside, surrounded by mountains, with the sun rising in the background.



A middle-aged man with long, gray hair and a short beard smiles gently, his warm, expressive eyes capturing attention against a dark background.



A stone bridge with arched spans reflects golden light on the water, flanked by buildings and a leafy tree under a clear blue sky.



ScaleCrafter

FouriScale

SANA

Diffusion-4K

UltraFlux (Ours)

Figure 13. More visual comparison of open-source methods on the Aesthetic-Eval@4096 benchmark at 4096×4096 resolution.

