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. The supplement is organized as follows:

- **Sec. S1 (Novelty & Co-Design):** Positions UltraFlux as a regime-level contribution via a 2×2 data–model ablation.
- **Sec. S2 (Implementation Details):** Details the dataset pipeline, DiT training, and our F16 VAE post-training recipe.
- **Sec. S3–S4 (Positioning & Wavelets):** Analyzes Resonance 2D RoPE and wavelet-space statistics to motivate our architectural choices.
- **Sec. S5–S7 (Ablations & Benchmarks):** Reports 4K runtime, wide-AR performance, and comparisons with open-source baselines.
- **Sec. S8–S9 (Limitations & Evaluation):** Discusses system constraints and our LLM-based (Gemini/GPT-4O) evaluation pipeline.

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.

1. 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 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.

Table 1. 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 ↓	HPSv3 ↑
A	Diffusion-4K-v2 [4]	Flux, latent L2	152.09	8.57
B	MultiAspect-4K-1M	Flux, latent L2	151.41	9.17
C	Diffusion-4K-v2 [4]	UltraFlux	147.41	10.03
D	MultiAspect-4K-1M	UltraFlux	145.81	10.78

To make this clearer, we provide in this supplement:

- A data-model ablation (Table 1) 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. 3), 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. 4) 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.

2. Implementation Details

2.1. 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* [2]. 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* [1] 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.

Table 2. Reconstruction metrics of F16 VAEs on the Aesthetic-4K@4096 Eval set [3].

Model	rFID ↓	NMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Flux-VAE-F16 [3]	2.201	0.01522	26.90	0.784	0.168
Flux-VAE-F16-SC [4]	0.588	0.00736	30.19	0.846	0.097
UltraFlux-F16-VAE	**0.547**	**0.00657**	**30.70**	**0.852**	**0.102**

2.2. 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 update all DiT blocks end-to-end. Training is conducted on $8\times$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 2 quantitatively compares our **UltraFlux-F16-VAE** with the Flux-VAE-F16 baseline on the AESTHETIC-4K@4096 evaluation set [3]. 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.

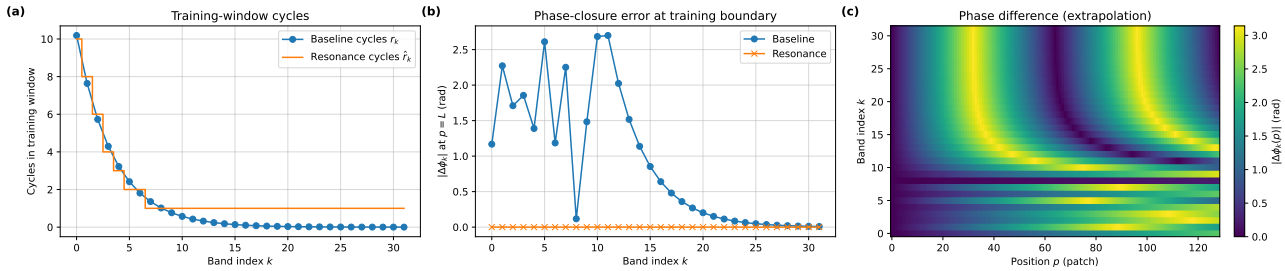


Figure 1. 1D band-wise analysis of Resonance 2D RoPE with YaRN. (a) Number of cycles r_k in the training window and their integer-snapped counterparts $\hat{r}_k$. (b) Phase-closure error $|\Delta\phi_k|$ at $p = L$, showing exact closure for Resonance RoPE. (c) Phase difference $|\Delta\phi_k(p)|$ between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

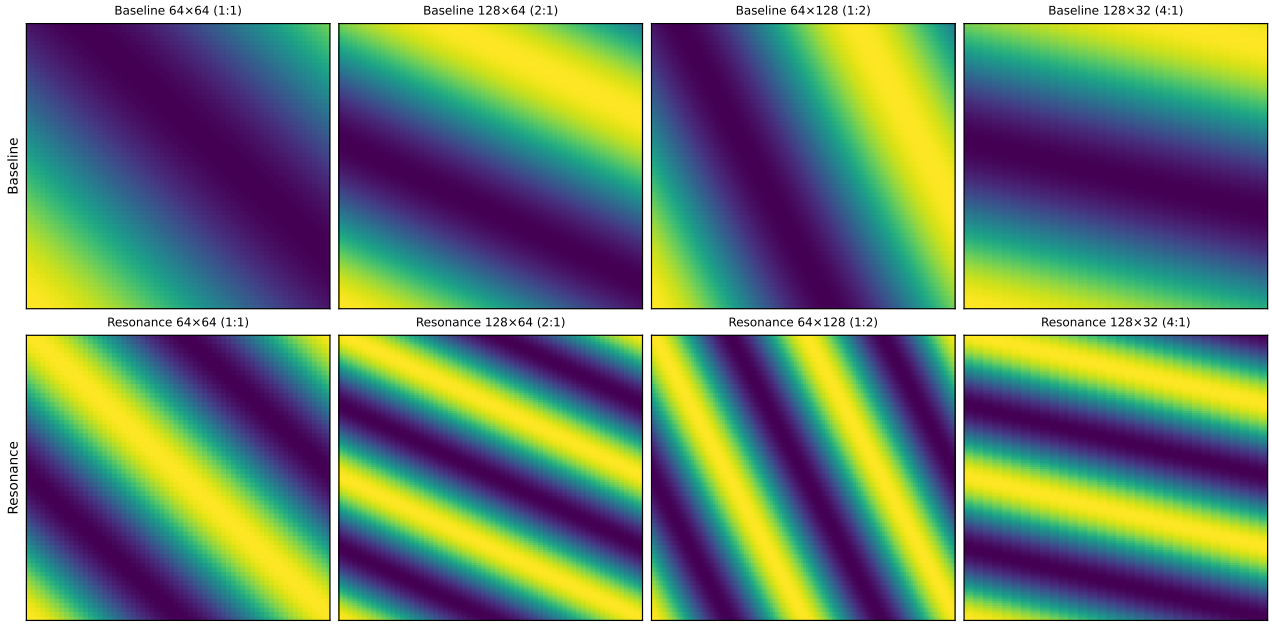


Figure 2. Height–width cosine phase patterns for a representative rotary band under different aspect ratios. Each panel displays $f(h, w) = \cos(h\,\omega_H + w\,\omega_W)$ evaluated on an $H \times W$ patch grid, with the top row using the Flux baseline frequencies and the bottom row using our resonant 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.

3. Analyses of Resonance 2D RoPE with YaRN

Figure 1 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 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

$$\hat{r}_k^{(a)} \;=\; \max\!\big(1,\ \mathrm{round}(r_k^{(a)})\big).$$

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

Table 3. 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 (ArtiMuse percentile)	top-5%
DiT objective	Wavelet type / number of levels	Haar, J=1
	Pseudo-Huber thresholds ($c_{\min}, c_{\max}$)	$c_{\min}\approx 0.2,\ c_{\max}\approx 1.0$
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)
	Maximum extrapolation scale $s_a = L'_a/L_a$	2.0
F16 VAE post-training	Training resolution	512×512
	Global batch size (images/step)	384
	Optimizer / learning rate / weight decay	AdamW, 1×10^{-4}, 1×10^{-2}
	Loss weights ($\lambda_{\text{wav}}, \lambda_{\text{perc}}, \lambda_{L_2}$)	0.2, 0.1, 1
Multi-AR 4K DiT training	Landscape target sizes (W×H)	5440×3072, 5184×3264, 4992×3328, 4736×3520, 5824×2880, 6272×2688, 5568×3008, 6336×2624, 5632×3008, 4608×3648
	Portrait target sizes (W×H)	3072×5440, 3648×4608, 3520×4736, 3328×4992
	Square target sizes (W×H)	4096×4096

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

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

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

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

where $\mathrm{wrap}(\cdot)$ 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 1 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 2. For a representative rotary band, we construct

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

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.

4. Wavelet-Space Statistics of 4K VAE Latents

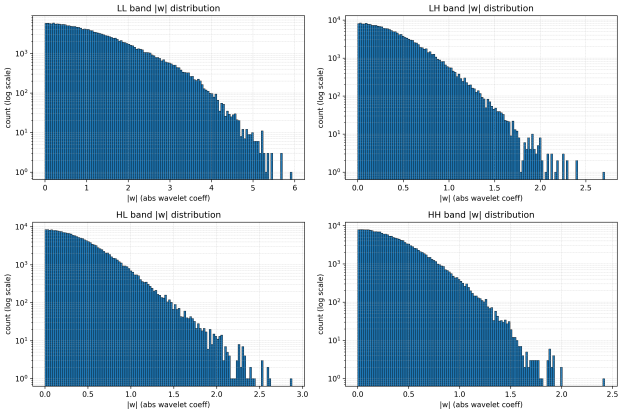


Figure 3. Wavelet-space statistics of 4K VAE latents. We show log-count histograms of absolute coefficients for the LL, LH, HL, and HH subbands over 400 samples. Most energy resides in the LL band, while high-frequency bands carry sparse but large-magnitude coefficients, indicating heavy-tailed behavior. 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 3 shows log-count histograms of the absolute wavelet coefficients in the LL, LH, HL, and HH

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

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

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 $|w| > 0.5$, 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.

5. Additional Ablations

Figure 4 provides a visual counterpart to the analyses in Sec. 3. 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.

6. Efficiency Comparison

Table 4 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 5. Quantitative comparison with SOTA methods at different aspect ratios, including 4096×2048 (2:1), 2048×4096 (1:2), 5120×2880 (16:9) and 2496×5292 (1:2.39) resolutions.

Aspect Ratio	Method	FID ↓	HPSv3 ↑	Artimuse ↑	Q-Align ↑
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**
1:2.39	ScaleCrafter	162.39	7.88	**69.01**	4.17
	FouriScale	167.04	6.73	65.89	4.17
	Sana	**150.45**	**11.91**	65.49	4.82
	UltraFlux	151.98	11.76	66.36	**4.84**

7. Additional Visual Comparison With Open-Source methods

In Figures 5–7, 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.

8. 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 acpect ratios and resolutions, as detailed in Table 5. The results show that UltraFlux consistently match or 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.

Figure 4. **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).

9. Qualitative Comparison with SOTA Methods at Wide Aspect Ratios

This section provides visual comparisons with SOTA methods at wide aspect ratios, complementing our quantitative analysis in Table 5. The results are presented in Fig. 8 (1:2), Fig. 9 (2:1), Fig. 10 (16:9) and Fig. 11, respectively.

At the 1:2 aspect ratio, all methods produce visually plausible results without severe artifacts. However, our results are more visually appealing with better composition and aesthetic quality. In the 2:1 case, methods such as Scalecrafter and Fouriscale exhibit noticeable structural distortions and artifacts, while Sana also shows visible flaws. In contrast, our method generates remarkably natural and coherent images. At the 2.39:1 ultra-wide ratio, both Scalecrafter and Fouriscale suffer from mild misalignment with text prompts as well as detail degradation. Our results not only avoid these issues but also outperform Sana in overall visual quality.

These observations demonstrate that our approach consistently maintains state-of-the-art performance and high visual fidelity across a spectrum of challenging aspect ratios.

10. 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 native 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 heav-

ily 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 complementary 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.

11. 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 11.1 (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"
}
```

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

Prompt 12.1 (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 newlines, 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.

References

[1] Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajuin Yao, and Yihao Liu. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding, 2025. 2

[2] 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 lmms 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. 2

[3] 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

[4] 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

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.

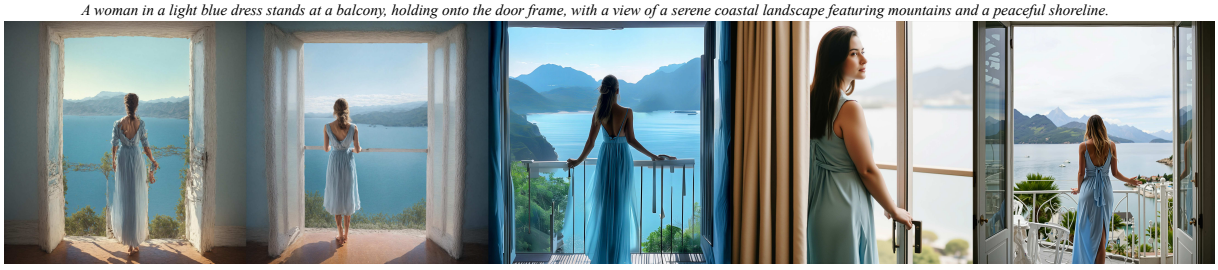
Blonde model with tousled hair wearing a form-fitting olive green dress, looking intently at the camera with a soft expression.

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

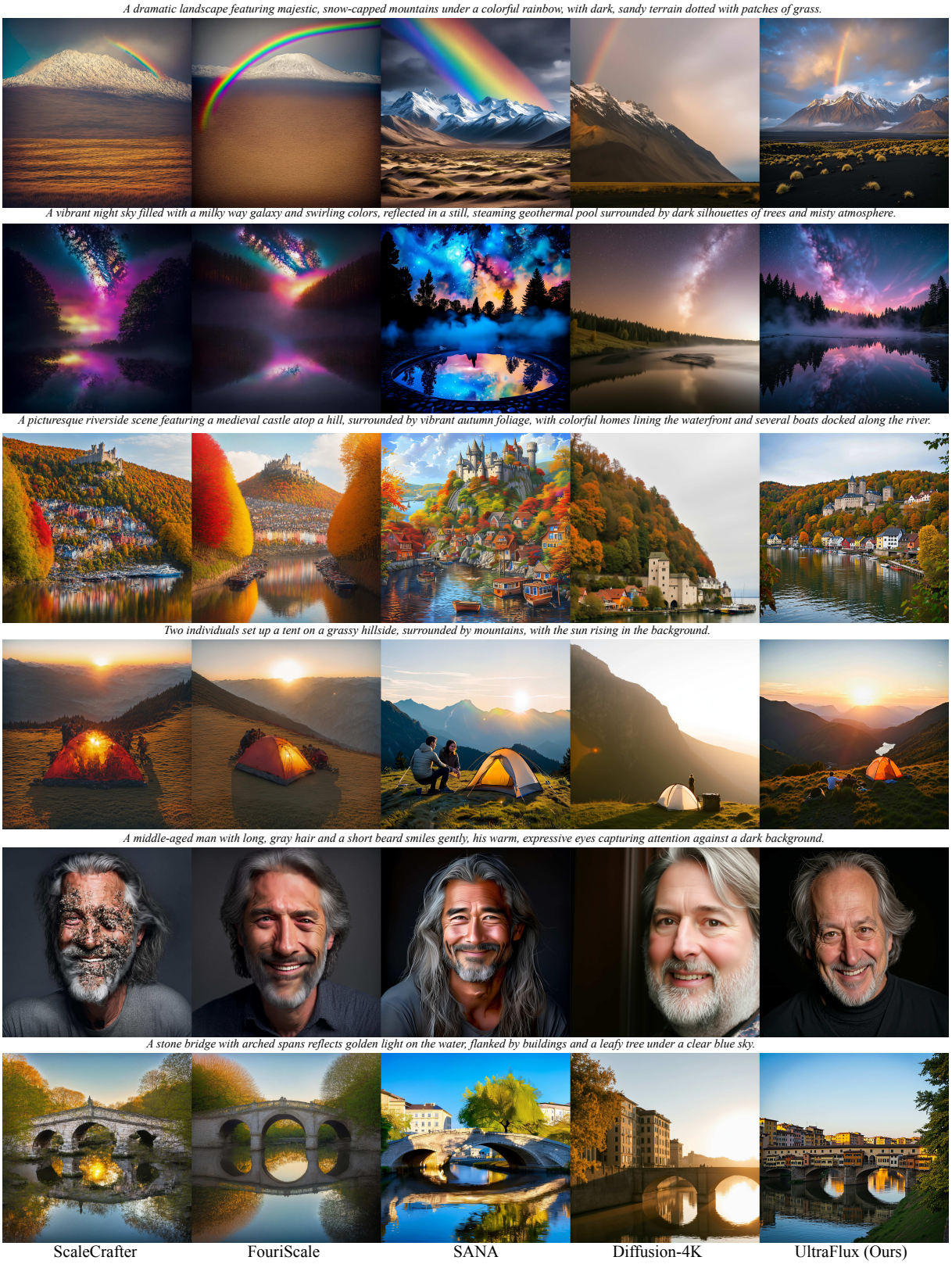


Figure 7. More visual comparison of open-source methods on the Aesthetic-Eval@4096 benchmark at 4096×4096 resolution.

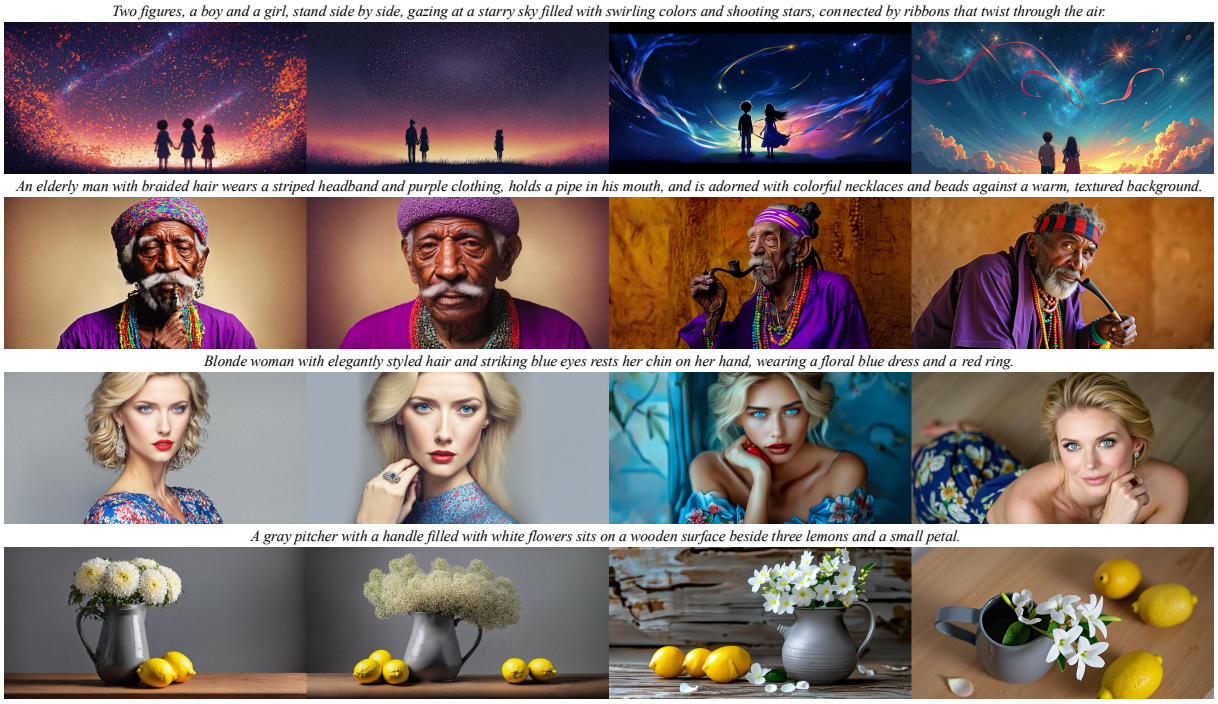


Figure 8. Visual comparison of open-source methods at 1:2 aspect ratio (2048x4096).

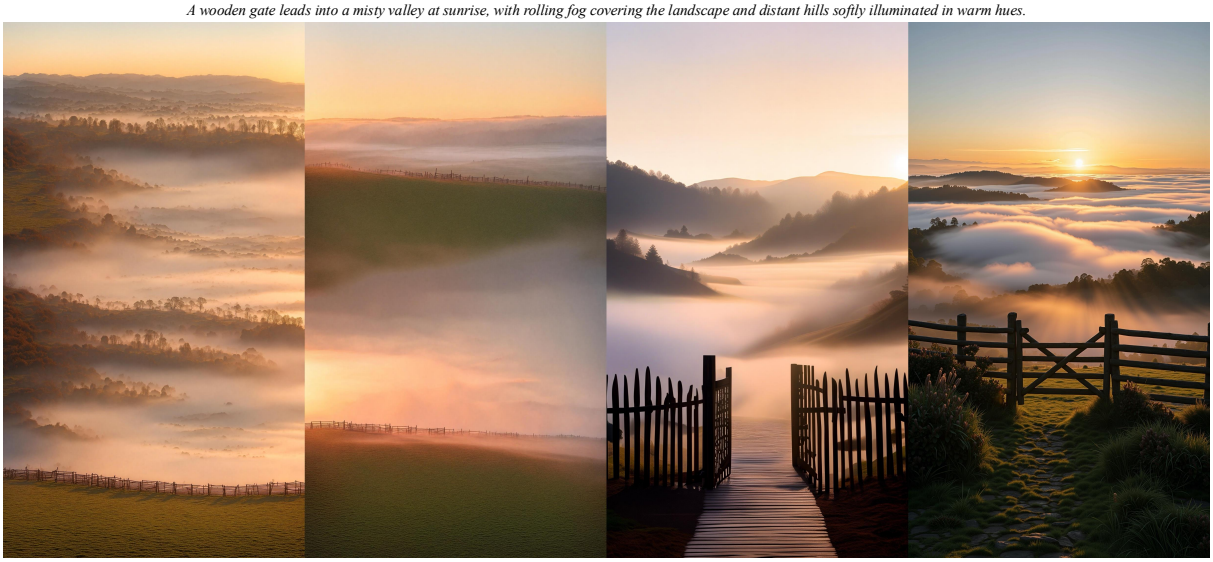
A woman in a white tank top and light blue leggings sits casually on a modern white chair, one leg draped over the side, resting her head on her hand, with an acoustic guitar on the floor beside her and a black chair nearby. A silver pitcher and a dark wooden table are visible in the background.

A young woman with curly hair and rosy cheeks leans out of a stone window, holding a red fan adorned with floral patterns, dressed in a traditional European blouse with intricate detailing and a necklace.

ScaleCrafter FouriScale Sana UltraFlux (Ours)

Figure 9. Visual comparison of open-source methods at 2:1 aspect ratio (4096x2048).

A wooden gate leads into a misty valley at sunrise, with rolling fog covering the landscape and distant hills softly illuminated in warm hues.



A woman kneels on the forest floor, smiling as she offers grapes to a large brown bear beside her, surrounded by tall birch trees.

ScaleCrafter FouriScale Sana UltraFlux (Ours)

Figure 10. Visual comparison of open-source methods at 16:9 aspect ratio (5120x2880).

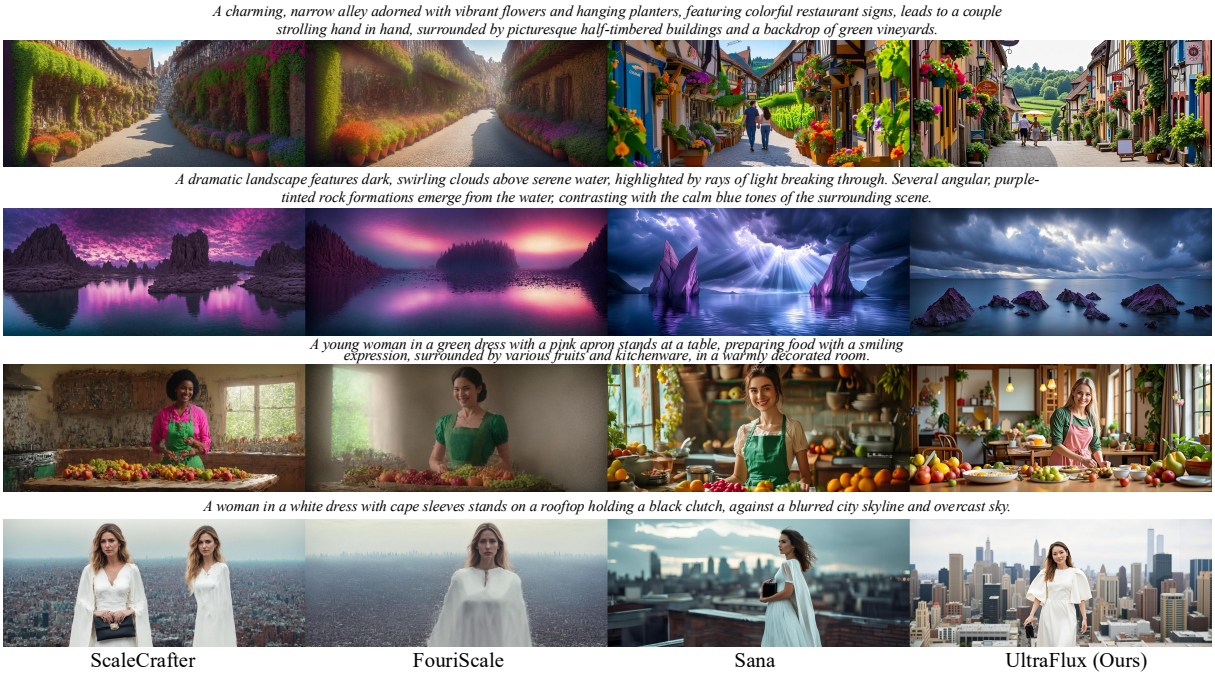


Figure 11. Visual comparison of open-source methods at 1:2.39 aspect ratio (2496x5952).