Phase-Aligned RoPE for Mixed-Resolution Diffusion Transformer

Haoyu Wu[1], Jingyi Xu[1], Qiaomu Miao[1], Dimitris Samaras[1], and Hieu Le[2]

[1] Stony Brook University, NY, USA
[2] UNC-Charlotte, NC, USA

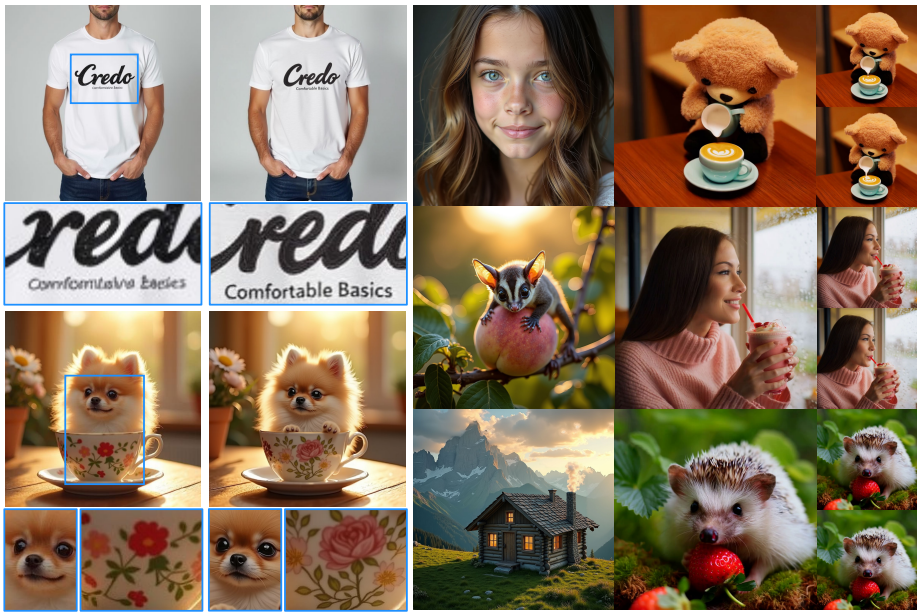
Abstract. Rotary positional embeddings (RoPE) are widely used in diffusion transformers (DiTs) to encode spatial relationships, yet their behavior with mixed-resolution tokens remains underexplored. A natural approach is to rescale token positions from different resolutions into a unified coordinate system before attention, but we show this fails. Our analysis shows that with RoPE, the attention similarity score is a highly structured and periodic function of token distance, so rescaling distances across resolutions moves token pairs to different regions of this periodic function, leading to incorrect attention scores. Motivated by this, we introduce Phase-Aligned Mixed-Resolution Attention (PMA), a training-free mechanism that stabilizes mixed-resolution attention. PMA modifies the RoPE position mapping to enforce a consistent positional scale for every query–key pair, ensuring that relative distances are evaluated under a single reference scale. To further improve local coherence near resolution transitions, we incorporate a lightweight boundary refinement module that softly exchanges features across adjacent scales. Experiments on image and video diffusion models validate our analysis and demonstrate consistent improvements in visual fidelity and computational efficiency. Project page: `https://hao-yu-wu.github.io/mixed_res/`.

Keywords: Mixed-Resolution Attention · Diffusion Models

1 Introduction

Improving the efficiency of diffusion models has become a central challenge in modern generative modeling. As the resolution of images and videos increases, the quadratic cost of attention quickly dominates computation, making high-resolution generation increasingly expensive [5, 99, 109]. A promising strategy to mitigate this cost is mixed-resolution processing [15, 38, 77, 84]: coarse resolution for background or non-essential regions, and higher resolution for important regions to capture fine-grained details. By allocating computation adaptively, mixed-resolution denoising can, in principle, deliver sharper details without the full cost of uniformly high-resolution attention.

In practice, however, we observe diffusion transformers (DiTs) [38,43,95] often struggle under mixed-resolution processing. When tokens from different resolutions are processed together in a single attention operation, the results often



(a) LR (b) Ours (c) Our Generated Samples

Fig. 1: Our method enables **stable mixed-resolution denoising** (b), performing high-resolution (HR) denoising on salient regions (blue boxes) while simultaneously denoising the remaining areas at low resolution (LR). (a) shows a low-resolution baseline. (c) presents image and video samples generated by our method.

exhibit blur and unstable artifacts, even when positions are carefully rescaled into a unified coordinate system, as shown in Sec. 3. We find that the root cause lies in the interaction between attention and Rotary Positional Embeddings (RoPE) [85], which is the standard positional encoding in modern large language models [2,3,23,31,58,93] and diffusion transformers [32,43,95,104,118].

In essence, RoPE works by rotating token embeddings according to their positions. We find that such transformation introduces a strong position-dependent scale bias into every attention score. As we will show empirically (see Fig. 3) and theoretically (see Eq. (6)), the expected cosine similarity between a query–key pair follows a sinusoidal-like curve $\kappa(\Delta)$, as a function of their relative token distance Δ. This means that token pairs at certain distances tend to have artificially higher or lower similarity scores, on top of what token content alone would dictate. When all tokens share the same resolution, this bias is consistent and the model learns to work with it naturally. The problem arises in mixed-resolution settings: any attempt to unify LR and HR tokens into a shared positional space will distort the native distance scale of at least one group, compressing or stretching their pairwise distances and shifting them to different regions (phases) of the $\kappa(\Delta)$ curve. Some pairs get artificially high attention scores, others get suppressed, rendering the outputs blurry or with random image artifacts.

Motivated by this diagnosis, we introduce **P**hase-Aligned **M**ixed-Resolution **A**ttention (**PMA**), a training-free mechanism that stabilizes mixed-resolution attention. PMA modifies the RoPE position mapping so that, for every token-level attention, the token positions always share the same reference scale that the model was trained on. To further improve local continuity near resolution transitions, we incorporate a lightweight boundary refinement module that softly exchanges features across adjacent scales. Combined with a practical coarse-to-fine denoising schedule, our approach enables stable, high-fidelity, and efficient generation for images and videos. As illustrated in Fig. 1, by supporting targeted regions to be processed at higher resolution, our method unlocks fine-grained detail and new possibilities for controllable, high-quality content generation.

Our contributions are:

- We uncover and formally analyze a structural limitation of RoPE that makes mixed-resolution attention inherently unstable.
- We propose Phase-Aligned Mixed-Resolution Attention (PMA), a training-free mechanism that restores stability by enforcing a consistent and native positional scale, along with a lightweight boundary refinement module.
- We demonstrate stable and efficient mixed-resolution image and video generation with substantially enhanced fine-grained details.

2 Related Work

Diffusion Transformers (DiTs). Replacing U-Nets with transformers has proven highly scalable for diffusion models, with DiTs operating on latent patches and achieving state-of-the-art generation [66]. Multimodal variants further improve text-to-image [8,24,43,45,50,87,100] and text-to-video [27,32,61,72,95,117] alignment and visual fidelity [47, 48, 102, 103].

Positional Encoding. A key choice in DiTs is the positional encoding. RoPE [85] rotates queries and keys using frequency-based phases and is now standard across vision [17,29,33,58,92] and language transformers [14,16,23,88,89]. Work on extending RoPE beyond its trained context—linear interpolation, NTK-aware scaling, YaRN—targets long-range extrapolation [12, 19, 67, 68]. Trainable or Lie-group generalizations [65,107] aim to improve robustness across resolutions, while alternatives like ALiBi [73], XPos [86], and other relative schemes pursue greater stability. However, prior efforts concentrate on language or long-context vision tokens [114], instead of our mixed-resolution setting.

Diffusion Acceleration. Many acceleration methods reduce either step count or per-step compute and are orthogonal to our focus [22]. Token-reduction techniques merge, prune, or route tokens to lower cost with minimal degradation [5,7,11,106,108,109]. Model-compression approaches—quantization [9,18,46,81], distillation [28,49,112], block pruning [25,62,80], and structured or unstructured pruning [6, 26, 30, 94, 113]—trade capacity for efficiency. Feature-caching methods [10, 60, 62, 97] reuse hidden states across steps [1, 40, 53, 63, 119], sometimes with selective token routing [54, 55, 106]. Sampler advancements cut step counts

without retraining [56, 57, 82, 115, 116], and progressive distillation or consistency models push toward few- or one-step generation [59, 79, 83]. Coarse-to-fine and multi-stage diffusion pipelines reduce compute while maintaining fidelity, often operating in VAE latents [13, 76]. Two-stage [69, 71] and cascaded systems are common [34, 39, 78, 90]. Other approaches explore high-low-high strategies [91], region-adaptive sampling [54], latent-space super-resolution [37], and patch-based schemes [20].

To the best of our knowledge, our work is the first to focus on mixed-resolution generation. The closest prior work is RALU [38], which uses mixed-resolution as the intermediate stage of a coarse-to-fine pipeline; however, it relies on linear interpolation of RoPE, overlooks RoPE phase mismatch across resolutions, and compensates with additional noise and extra steps.

3 Understanding RoPE Interpolation Failures in Mixed-Resolution Denoising

In this section, we investigate why mixed-resolution denoising destabilizes DiTs. We show that RoPE imposes a strong positional scale bias on attention scores as a function of token distance, following a structured sinusoidal pattern and systematically inflates or deflates similarity scores on top of what token content alone would dictate. In mixed-resolution settings, forcing LR and HR tokens into a shared positional coordinate space corrupts this distance structure, producing systematically wrong attention scores that manifest as blur and visual artifacts.

3.1 Preliminary - RoPE

We first recall a simple case of RoPE in 1D. Let an attention head have even dimensionality d, and let ω_i denote the i-th angular frequency, typically a geometric sequence: $\omega_i = 10000^{-2i/d},\ i \in \{0, 1, \ldots, d/2-1\}$. For a scalar position $p \in \mathbb{R}$, RoPE rotates each $(2i, 2i{+}1)$ pair by angle $\theta_i(p) = \omega_i\, p$:

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \tag{1}$$

$$\mathcal{R}(p) = \mathrm{diag}\big(R(\theta_0(p)), \ldots, R(\theta_{d/2-1}(p))\big). \tag{2}$$

Given tokens $q, k \in \mathbb{R}^d$, RoPE applies $q \to \mathcal{R}(p_q)q,\ k \to \mathcal{R}(p_k)k$. The attention score obeys the relative property

$$\big(\mathcal{R}(p_q)q\big)^\top \big(\mathcal{R}(p_k)k\big) \;=\; q^\top \mathcal{R}(\Delta)\, k, \tag{3}$$

i.e., RoPE converts absolute positions to a frequency-coded phase that depends only on the relative offset $\Delta = p_k - p_q$.

(a) PI-LR (b) PI-HR (c) Ours

Fig. 2: Results for RoPE with linear position interpolation (PI) [12] to the low- or high-resolution grid, versus our method.

3.2 Position Interpolation Fails for Mixed-Resolution Tokens

We study the problem of mixed-resolution denoising: allocating high resolution to salient regions and lower resolution elsewhere. How attention behaves in this setting remains an open problem. The natural baseline is linear position interpolation (PI) [12], widely used in language models, vision transformers, and DiTs, which maps all tokens onto a unified positional space.

Consider a 1D example. Starting from indices [0, 1, 2, **3, 4**, 5, 6, 7, 8], suppose the middle segment [3, 4] is upsampled into a HR block. Two natural unification strategies are:
(i) Fractional unification (LR grid with fractional HR indices).

$$\underbrace{0\ \ 1\ \ 2}_{\text{LR}}\ \ \underbrace{3.0\ \ 3.5\ \ 4.0\ \ 4.5}_{\text{HR}}\ \ \underbrace{5\ \ 6\ \ 7\ \ 8}_{\text{LR}}.$$

(ii) Integerized unification (warp LR zones to keep integer indices). Here, LR zones are stretched by 2 while the HR block remains dense:

$$\underbrace{0\ \ 2\ \ 4}_{\text{LR}}\ \ \underbrace{6\ \ 7\ \ 8\ \ 9}_{\text{HR}}\ \ \underbrace{10\ \ 12\ \ 14\ \ 16}_{\text{LR}}.$$

Both can be expressed as a piecewise-affine map on RoPE positions p:

$$\phi(p) = a_r + s_r\,(p - b_r), \qquad p \in r, \tag{4}$$

where r represents the region (LR or HR), s_r is the per-region scale factor, b_r is the region's original start index, and a_r ensures continuity across boundaries. However, as shown in Fig. 2, both strategies fail systematically: one produces plausible LR regions but collapses HR, and the other does the opposite. We explain why next.

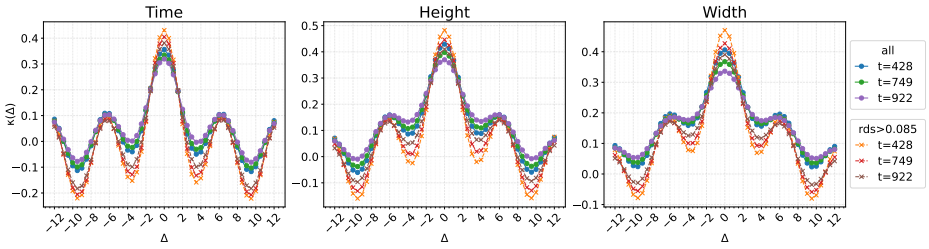


Fig. 3: RoPE imposes a sinusoidal scale bias on the attention function. We plot the mean normalized attention score $\kappa(\Delta)$ as a function of relative distance Δ on the Wan model [95], measured along three axes (time, height, width) across diffusion steps $t \in \{428, 749, 922\}$. $\kappa(\Delta)$ peaks sharply near $\Delta \approx 0$ and oscillates with a clear sinusoidal structure at larger offsets. This is independent of token content since we measure this with random token pairs. This bias is amplified in RoPE-dominant heads, *i.e.*, heads with a RoPE-dominance score greater than 0.085 (*rds*), and remains stable across timesteps.

3.3 RoPE Imposes a Sinusoidal Scale Bias That Breaks Under Mixed Resolution

To understand the failure above, we directly measure the relationship between attention scores and relative distance across attention heads in pre-trained DiTs. Specifically, we compute the expected cosine similarity (scale-free proxy for the pre-softmax attention score), between a query and a RoPE-rotated key, as a function of their relative token distance Δ:

$$\kappa(\Delta) := \mathbb{E}_{(q,k)}\big[\,\langle \widehat{q},\, \mathcal{R}(\Delta)\,\widehat{k}\rangle\,\big], \tag{5}$$

where hats denote ℓ_2-normalization. Importantly, we average over ***random*** token pairs. Thus, $\kappa(\Delta)$ isolates the pure positional scale bias imposed by RoPE—independent of what the tokens actually represent. In Fig. 3, we show the aggregated $\kappa(\Delta)$ across time, height, and width axes, and additionally analyze *RoPE-dominant* heads selected by the RoPE-dominance score (*rds*) following [14].

Empirical findings. As shown in Fig. 3, $\kappa(\Delta)$ exhibits three consistent properties across all layers and timesteps:

1. **Sharp peak near $\Delta \approx 0$.** The similarity is highest for nearby tokens and drops steeply within the first 2–3 offsets.
2. **Sinusoidal oscillations.** Beyond the initial peak, $\kappa(\Delta)$ does not decay smoothly—it oscillates, with alternating regions of high and low similarity at larger offsets.
3. **Stable across timesteps.** This structure persists from early to late diffusion steps and is amplified in RoPE-dominant heads, confirming it reflects a *pretrained phase prior* rather than a denoising artifact.

This sinusoidal scale bias is what makes mixed-resolution attention fundamentally problematic. In a single-resolution setting, all tokens share the same

distance scale, *i.e.*, the unit of positional spacing, so $\kappa(\Delta)$ acts consistently and the model learns to work within it. But when the positions of LR and HR tokens are forced into a shared coordinate space, any remapping inevitably assigns inconsistent pairwise distances to at least one group of tokens: either HR token positions get compressed, or LR token positions get stretched. Because $\kappa(\Delta)$ oscillates rather than decays monotonically, these distortions are unpredictable: compressing a distance from 4 to 2 may land on a peak, while compressing from 2 to 1 may land on a trough, or vice versa. Some token pairs become artificially over-attended, others spuriously suppressed, rendering the output blurry or with random visual artifacts. No single rescaling can fix this, because there is no position remapping that can preserve the oscillatory structure of $\kappa(\Delta)$ across two different positional scales simultaneously.

Theoretical interpretation. Using the relative form of RoPE (Eq. (3)), the attention score can be written as a mixture of sinusoids:

$$\text{score}(q, k, \Delta) = \sum_i C_i(q, k) \cos(\omega_i \Delta + \phi_i), \tag{6}$$

where ω_i are fixed RoPE frequencies and (C_i, ϕ_i) are content- and head-dependent coefficients (full proof in supplementary material). This confirms that each attention head implements a *learned sinusoidal phase filter*: high-frequency components produce the steep near-zero peak and strong oscillations observed empirically. The filter is calibrated to a single positional scale during training. When distances are rescaled, token pairs get shifted to different phases (regions) of the same periodic pattern, so pairs that should align can become misaligned. Mixed-resolution processing therefore samples the filter at two incompatible scales simultaneously, which is the root cause of failures observed in Sec. 3.2.

Overall, the issue is not the choice of interpolation scheme but the lack of a uniform positional scale for Δ. Stable mixed-resolution attention therefore requires all token positions to be expressed on a consistent scale that aligns with the attention heads in pre-trained DiTs. We introduce such a mechanism next.

4 Method

To stabilize attention with mixed-resolution tokens, we propose Phase-Aligned Mixed-Resolution Attention (PMA), a training-free mechanism that enforces a consistent positional scale across all tokens and is native to pre-trained DiT attention heads. To further smooth content transitions between LR and HR regions, we also introduce a Boundary Expand-and-Replace module. Together, these components form a stable mixed-resolution denoising pipeline.

4.1 Phase-Aligned Mixed-Resolution Attention (PMA)

The core insight from Sec. 3.3 is that attention scores are only meaningful when relative token distances are expressed at the same positional scale as seen during pre-training. Any remapping that forces LR and HR tokens into a shared

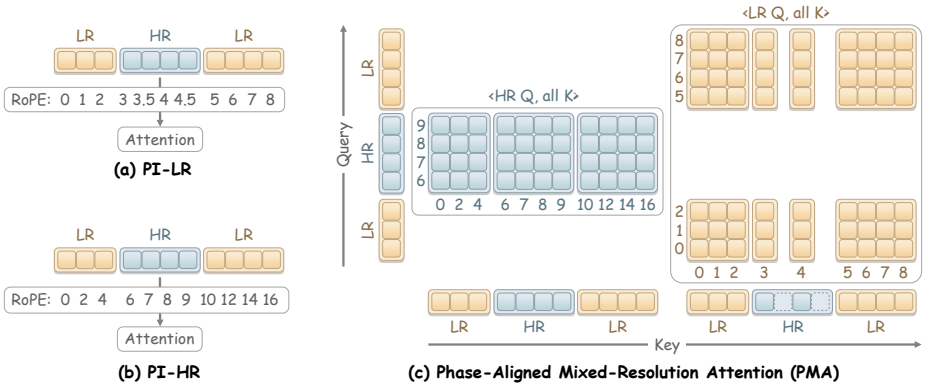


Fig. 4: Phase-Aligned Mixed-Resolution Attention (PMA). PMA measures RoPE offsets in the query's native units by rescaling key positions to the query grid, aligning RoPE phases across resolutions for stable mixed-resolution denoising. (a) and (b) illustrate baselines that interpolate positions to LR and HR grids.

coordinate space violates this: it inevitably distorts the pairwise distances of at least one group of tokens, causing their attention scores to be evaluated at the wrong region of the sinusoidal-like curve $\kappa(\Delta)$, *i.e.*, at a different phase of this periodic function.

Thus, the fix is straightforward: rather than remapping all tokens into a single global coordinate space, we perform remapping *locally* for each query–key pair: always expressing the key's position in the query's native positional scale. This ensures that for every dot product attention, the relative offset Δ is measured in units consistent with pre-training.

Let S denote the native resolution scale of tokens, *i.e.*, HR has a larger S than LR, and define the query–key scale ratio: $\alpha_{k\to q} = S_q/S_k$. For attention between each key and query, we re-index the position of the key to the query's reference grid:

$$p_k^{(q)} = \alpha_{k\to q}\, p_k, \tag{7}$$

and compute the RoPE-based attention score

$$\langle \widetilde{q},\, \widetilde{k} \rangle = \big\langle \mathcal{R}(p_q)\, q,\ \mathcal{R}\big(p_k^{(q)}\big)\, k \big\rangle = \big\langle q,\, \mathcal{R}\big(\alpha_{k\to q}\, p_k - p_q\big)\, k \big\rangle. \tag{8}$$

This leads to two cases in practice (Fig. 4):

- **HR query vs. all keys.** The HR query's native scale is the reference. We stretch LR key positions to match the HR grid. In the toy example, HR query positions remain at $[6, 7, 8, 9]$, and LR key positions are stretched: $[0, 1, 2][5, 6, 7, 8] \mapsto [0, 2, 4][10, 12, 14, 16]$.
- **LR query vs. all keys.** The LR query's native scale is the reference. We compress HR key positions to match the LR grid. Since LR queries only require coarse context, we also downsample HR keys to the LR grid via strided sampling, with their positions compressed: $[6, 7, 8, 9] \mapsto [3, 4]$ (stride=2).

PMA is training-free and requires no architectural changes: it only modifies token positions before attention. It ensures that, for every attention between a query and a key, their token positions always share the same scale that is consistent with the pre-trained attention heads.

4.2 Extension: Boundary Expand-and-Replace (BER)

PMA resolves the attention scale mismatch between resolutions, but small visual discontinuities may still appear near LR–HR boundaries (*e.g.*, slight density or texture shifts). We mitigate these with a lightweight, localized harmonization step that operates only in a narrow band around the boundaries (Fig. 5).

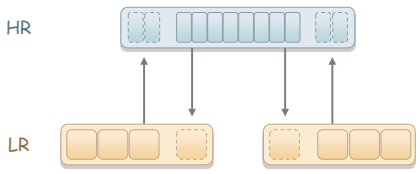


Fig. 5: Boundary Expand-and-Replace. Around LR–HR boundaries, we dilate the masks and bidirectionally exchange upsampled and downsampled latent content within the narrow band.

Given the LR and HR token masks, we first dilate each mask by n_{pad} (*e.g.*, 2) tokens to form an overlapping boundary band. At each diffusion step t, with noisy latent x_t and current clean estimate x_0, we perform a bidirectional content exchange within this band:

- **Low→high.** Upsample the LR portion of x_0 using a learned latent upsampler, re-noise it to timestep $t-1$, and replace HR tokens in the band for the next denoising step.
- **High→low.** Downsample the HR portion of x_0, re-noise it similarly, and replace the LR tokens in the band.

We train the latent up/downsamplers using paired targets obtained by pixel-space resizing and re-encoding, optimized with ℓ_1 in latent space and $\ell_1 + \text{LPIPS}$ in pixel space. Because the exchanged tokens serve only as localized attention context rather than final output, a small resizer model (25M) is sufficient.

4.3 Use Case: Saliency-Guided Mixed-Resolution Denoising

PMA is an attention mechanism that, given LR tokens and HR tokens, stabilizes attention for mixed-resolution generation. To demonstrate a concrete use case, we propose pairing PMA with an off-the-shelf lightweight saliency model to produce a spatial importance signal and enable mixed-resolution generation. Our contribution is orthogonal to the specific choice of importance localizer: it only requires a reasonable spatial weighting over regions, such as the one produced by off-the-shelf saliency predictors or user-provided masks. Moreover, these saliency predictors are lightweight relative to the diffusion model itself, adding negligible overhead to the overall pipeline. We adopt a coarse-to-fine pipeline:

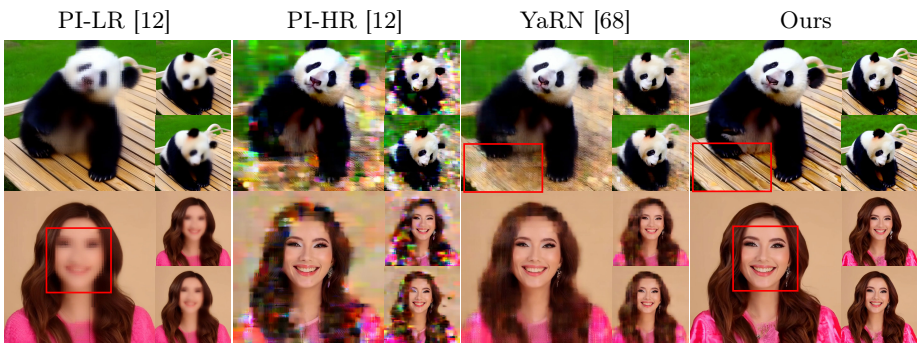


Fig. 6: Mixed-resolution video generation on Wan 2.1 [95]. We compare our method with linear interpolation [12] to low- or high-resolution grids (PI-LR/PI-HR), YaRN [68]. As highlighted in red boxes, our method produces the most stable results.

Fig. 7: Mixed-resolution image generation on FLUX [43] comparing RoPE with linear interpolation to low- and high-resolution grids against our method.

1. **Coarse.** Low-resolution generation for a few denoising steps to quickly establish the global structure and motion of the scene.
2. **Mixed-resolution.** An off-the-shelf saliency model identifies regions of higher importance from the coarse result; these regions are switched to high resolution while the remainder stays at low resolution. Denoising then proceeds using PMA and BER.
3. **Fine (optional).** A small number of final steps at full resolution to enhance visual quality.

As we will show in Sec. 5, this simple pipeline already achieves state-of-the-art results, and can also be seamlessly integrated with other diffusion acceleration techniques, highlighting its broad applicability.

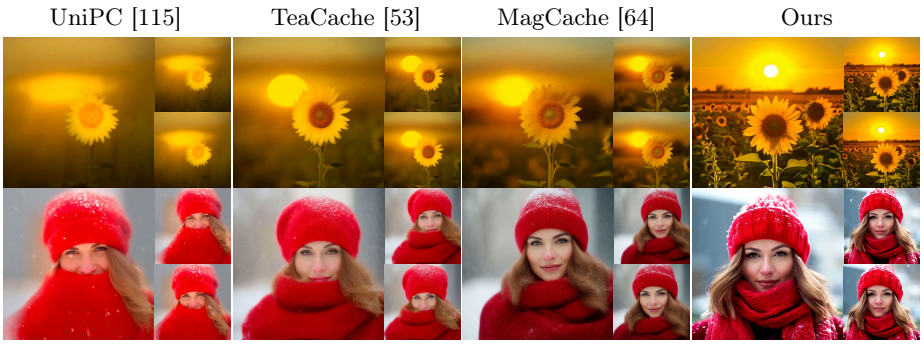


Fig. 8: Comparison with diffusion acceleration methods for video generation on Wan 2.1 [95]. Our method achieves higher quality and greater fidelity.

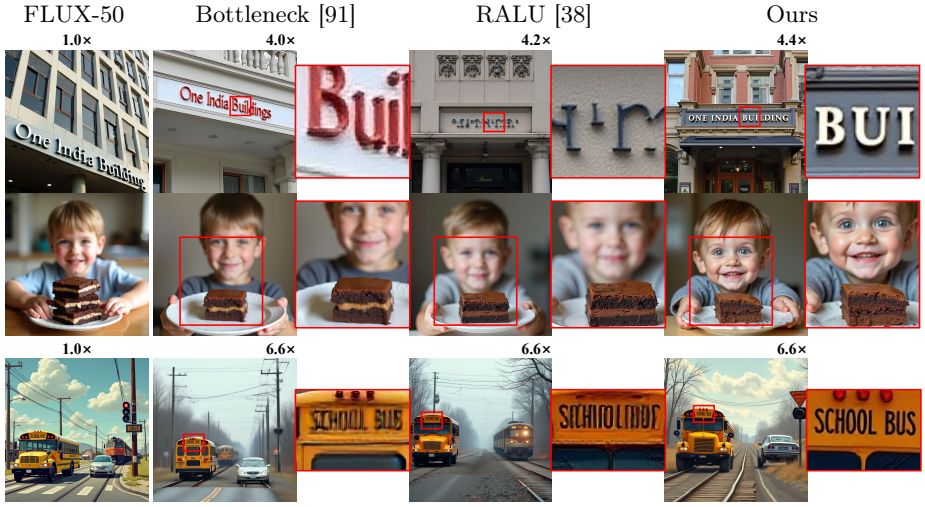


Fig. 9: Comparison with diffusion acceleration methods for image generation on FLUX [43]. We compare at 4× and 6× speedups, with 50-step FLUX generations as reference. Red boxes highlight the improved level of detail in our generation.

5 Experiments

5.1 Experimental Settings

Model Configurations and Metrics. We use Wan2.1-1.3B [95] as the text-to-video model and FLUX.1-dev [43] as the text-to-image model. For video evaluation, we use VBench [35] and DOVER [98], and report generation latency, with the full set of VBench prompts. For images, we report ImageReward [101] for photorealism; MUSIQ [41] and CLIP-IQA [96] for image quality assessment; CLIP score [75] for prompt alignment, and generation latency. We use MSCOCO 2014 validation dataset [51], with 5K randomly sampled image and caption pairs.

Table 1: Quantitative comparison with RoPE interpolation methods for mixed-resolution denoising on Wan2.1-1.3B [95] using DOVER [98] and VBench [35]. We include an HR baseline that denoises at high resolution as a reference.

	DOVER ↑			VBench ↑			Time
Method	Aesthetic	Technical	Overall	Quality	Semantics	Total	(s) ↓
HR	*99.83*	*10.43*	*79.12*	*80.12*	*62.30*	*76.56*	*172.1*
PI-LR [12]	98.10	8.01	63.39	75.93	54.92	71.73	
PI-HR [12]	86.52	4.94	35.04	70.38	49.41	66.18	
NTK [67]	92.76	5.89	44.52	71.80	52.93	68.02	43.2
PI+NTK	98.07	7.71	62.67	75.60	56.09	71.70	
YARN [68]	98.56	8.96	66.38	76.39	56.72	72.46	
Ours	**99.63**	**10.01**	**75.34**	**80.76**	**62.17**	**77.04**	43.2

Table 2: Quantitative comparison with RoPE interpolation methods for mixed-resolution denosing on FLUX.1-dev [43].

Method	**ImgReward ↑**	**CLIP-IQA ↑**	**MUSIQ ↑**	**CLIP ↑**	**Time ↓**
HR	*1.062*	*0.621*	*70.47*	*31.12*	*3.4 s*
PI-LR [12]	0.659	0.411	53.96	31.41	
PI-HR [12]	0.935	0.523	70.94	31.41	
NTK [67]	0.953	0.542	70.62	31.37	2.4 s
PI+NTK	0.810	0.479	60.23	**31.45**	
YARN [68]	0.926	0.548	69.99	31.29	
Ours	**0.978**	**0.623**	**71.81**	31.31	2.4 s

Implementation Details. For video generation, we adopt a two-stage inference scheme with 50 denoising steps: 15 steps at 480p and 35 mixed-resolution steps, with a high-resolution token ratio of 15% during the mixed-resolution stage. For two-stage image generation, we use 15 steps: 5 steps at 512 and 10 mixed-resolution steps, with a 60% high-resolution token ratio. For the three-stage image generation involving coarse, mixed, and fine stages, we adopt a RALU-like schedule [38] with a small number of steps at the final high-resolution stage and apply noise rescheduling accordingly. To select salient regions, we first use tiny VAE decoders [4] to reconstruct coarse images or videos (only 0.03s for video) following the low-resolution denoising stage, and then apply the off-the-shelf pre-trained DeepGaze model [52] to detect these regions. For training latent resizers, we use a batch size of 1, with Pexels [44, 70] and Aesthetic-Train-V2 datasets [110, 111].

5.2 Results

Comparison with RoPE interpolation methods. As shown in Tab. 1 and Fig. 6, our approach yields more stable video generation than RoPE interpolation baselines, including linear position interpolation [12] (to low- or high-resolution grids), NTK-aware interpolation [67], their hybrids, and YaRN [68]. Table 2 and Fig. 7 shows a similar trend for image generation.

Table 3: Quantitative comparison with diffusion acceleration methods on Wan2.1-1.3B [95]. We compare our method with caching (TeaCache [53], MagCache [64]), advanced samplers (UniPC [115], DPM++ [57]), and token merging (ToMe [5]).

	DOVER ↑			VBench ↑			Acceleration	
Method	Aes.	Tech.	Overall	Qual.	Sem.	Total	Time(s)↓	Speed
HR	*99.83*	*10.43*	*79.12*	*80.12*	*62.30*	*76.56*	*172.1*	1.0×
UniPC [115]	99.42	8.54	70.43	78.50	53.44	73.49	45.6	3.8×
DPM++ [57]	99.10	7.89	66.78	77.91	51.12	72.55	44.7	3.9×
ToMe [5]	89.34	6.50	48.47	68.80	28.80	60.80	48.0	3.6×
TeaCache [53]	99.30	8.54	69.53	76.78	53.19	72.06	43.5	4.0×
MagCache [64]	99.49	9.84	74.33	77.87	57.91	73.88	45.5	3.8×
Ours	**99.63**	**10.01**	**75.34**	**80.76**	**62.17**	**77.04**	**43.2**	**4.0×**

Table 4: Quantitative comparison with diffusion acceleration methods on FLUX [43]. We compare 4× and 6× speedups against temporal acceleration methods, including TeaCache [53], MagCache [64], and ToCa [119], spatial acceleration methods such as RALU [38] and Bottleneck Sampling [91], as well as reducing steps of FLUX.

Method	**ImgReward ↑**	**CLIP-IQA ↑**	**MUSIQ ↑**	**CLIP ↑**	**Time(s) ↓**	**Speed**
FLUX-50	*1.085*	*0.647*	*71.90*	*30.82*	*11.5*	1.0×
FLUX-12	0.985	0.588	68.94	31.19	2.8	4.1×
TeaCache [53]	0.803	0.519	63.36	30.91	2.8	4.1×
MagCache [64]	0.993	0.512	67.37	30.98	3.0	3.9×
RALU [38]	0.940	0.592	70.06	31.04	2.7	4.2×
ToCa [119]	0.956	0.498	66.49	**31.25**	2.9	4.0×
Bottleneck [91]	0.903	0.485	65.29	31.22	2.9	4.0×
Ours	**1.027**	**0.616**	**72.29**	31.16	**2.6**	**4.4×**
FLUX-7	0.905	0.484	62.85	31.24	1.7	6.6×
MagCache [64]	0.487	0.425	52.82	31.21	2.0	5.9×
RALU [38]	0.900	0.533	66.87	31.07	1.7	6.6×
ToCa [119]	0.345	0.435	50.23	30.99	1.8	6.5×
Bottleneck [91]	0.753	0.424	58.34	31.18	1.7	6.6×
Ours	**0.929**	**0.565**	**69.01**	**31.25**	**1.7**	**6.6×**

Comparison with diffusion acceleration methods. In Tab. 3, we report video generation results showing that our method outperforms other acceleration approaches, including temporal feature caching methods (TeaCache [53] and MagCache [64]), advanced samplers (UniPC [115] and DPM++ [57]), and token merging (ToMe [5]), with comparable or faster speed. We also evaluate our three-stage generation setting (coarse, mixed, fine) for image generation in Tab. 4. We compare 4× and 6× speedups against temporal acceleration methods, including TeaCache [53], MagCache [64], and ToCa [119], as well as spatial acceleration methods such as RALU [38] and Bottleneck Sampling [91]. As shown in Figs. 8 and 9, while other methods often produce reasonable outputs, our method better preserves key visual details at comparable or higher speed.

Integration with orthogonal diffusion acceleration methods. Our method is composable with other acceleration techniques. In Tab. 5, we show it combines

Table 5: Integration with orthogonal diffusion acceleration methods. We combine our method with feature caching methods and with the step-distillation model, DMD [105]. Speedups are reported relative to Wan2.1-1.3B [95] for video generation.

	DOVER ↑			VBench ↑			Acceleration	
Method	Aes.	Tech.	Overall	Qual.	Sem.	Total	Time(s)↓	Speed
Ours	99.63	10.01	75.34	80.76	62.17	77.04	43.2	4.0×
+ TeaCache [53]	99.59	10.05	75.34	80.33	61.28	76.52	23.9	7.2×
+ MagCache [64]	99.63	10.10	75.57	80.42	61.98	76.73	22.0	7.8×
DMD (8-step) [105]	99.96	12.70	87.13	82.51	71.31	80.27	22.6	7.6×
+ Ours	99.95	12.62	86.32	82.63	69.72	80.05	14.1	12.2×
DMD (4-step)	99.97	13.83	88.95	82.04	72.55	80.15	10.9	15.8×
+ Ours	99.96	14.03	87.97	82.32	71.31	80.11	5.6	30.7×

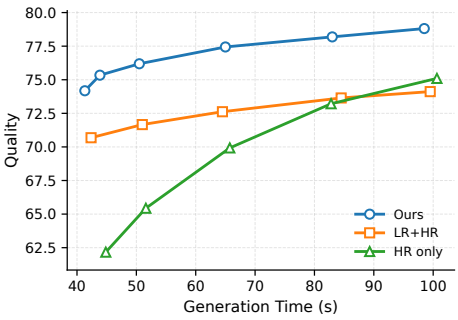


Fig. 10: Quality–cost trade-off on Wan 2.1 [95]. We report generation quality using DOVER [98] versus generation time.

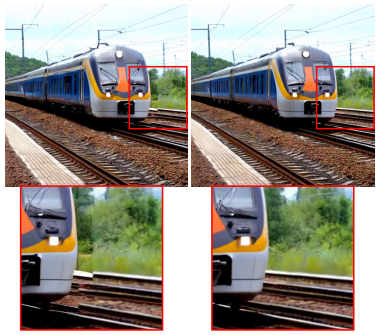


Fig. 11: Generation without (left) vs. with (right) Boundary Expand-and-Replace module (BER).

effectively with caching and with step-distillation models like DMD [105], improving efficiency while maintaining comparable generation quality.

Quality–cost trade-off. In Fig. 10, we compare against LR+HR (LR steps followed by HR steps) and HR-only (all steps in HR). We match generation time for fairness: for our method, we adjust the time budget by changing the ratio of HR tokens, LR+HR uses the same total steps as ours while varying the HR-step ratio, and HR-only varies total steps. Across all time budgets, our method achieves a consistently better quality–cost trade-off.

Ablation on BER. As shown in Fig. 11 and Tab. 7, removing our BER module degrades generation quality. Outputs look plausible but show subtle content discrepancies near the boundary when the underlying token resolution changes. We found padding 2 tokens for both LR and HR to be a reasonable setting.

Robustness to saliency models. Table 6 shows that our method performs consistently well across various off-the-shelf saliency detectors, and even when using a fixed center square as HR regions achieves competitive results. We also show that these detectors are lightweight and add negligible overhead.

Table 6: Robustness of our method across saliency models, with each model's runtime cost and model size reported.

	DOVER ↑			VBench ↑			Size	Time
Model	Aes.	Tech.	All	Qual.	Sem.	Tot.	(M)	(s)
DeepGazeI [42]	99.64	10.26	76.13	80.79	62.01	77.04	2.5	0.01
UNISAL [21]	99.66	10.24	76.26	80.64	61.83	76.88	3.7	0.01
DeepGazeIIE [52]	99.63	10.01	75.34	80.76	62.17	77.04	104	0.27
Center Square	99.59	9.69	74.22	80.50	61.74	76.75	-	-

Table 7: Ablation on overlapping boundary band size n_{pad} for low (LR) and high resolution (HR).

n_{pad}		DOVER ↑			VBench ↑		
LR	HR	Aes.	Tech.	All	Qual.	Sem.	Tot.
0	0	98.90	8.94	68.43	78.08	61.38	74.74
2	2	**99.63**	**10.01**	**75.34**	**80.76**	**62.17**	**77.04**
2	4	99.62	9.88	75.18	80.69	61.76	76.90

Table 8: 3-mixed-resolution (480 + 960 + 1920p) video generation.

	DOVER ↑			VBench ↑			Time
Method	Aes.	Tech.	All	Qual.	Sem.	Tot.	(s) ↓
Wan-2k	*93.99*	*5.85*	*52.14*	*82.72*	*42.75*	*74.73*	*1995*
PI-LR	62.29	4.73	19.87	75.18	41.39	68.42	288
PI-HR	86.31	6.48	45.23	74.01	39.96	67.20	
NTK	88.69	6.37	46.76	73.52	38.91	66.60	
YaRN	88.92	6.07	45.07	73.51	40.17	66.84	
Ours	**99.70**	**10.62**	**76.78**	**81.77**	**61.55**	**77.73**	**288**

Table 9: 3-mixed-resolution (512 + 1024 + 2048) image generation. Higher is better for all metrics except for time.

Method	**ImgR.**	**C.IQA**	**MUSIQ**	**CLIP**	**Time(s)**
FLUX-2k	*0.919*	*0.458*	*55.05*	*30.93*	*21.2*
PI-LR	0.536	0.292	40.19	28.63	9.8
PI-HR	0.391	0.276	45.82	30.32	
NTK	0.328	0.295	44.12	30.23	
YaRN	0.400	0.230	52.97	29.99	
Ours	**0.983**	**0.468**	**57.20**	**31.28**	**9.8**

Multi-resolution study. Our method naturally extends to multi-resolutions. Table 8 reports 3-mixed-resolution results for video: we outperform the linear interpolation baseline and direct 2K generation while being 7× faster. Table 9 reports 3-mixed-resolution image results with a similar trend. All methods use resolution extrapolation (CineScale [74], DyPE [36]).

6 Conclusion

We show that standard RoPE interpolation fundamentally breaks mixed resolution attention in DiTs: the model is forced to compare phases sampled at incompatible spatial rates, producing cross-scale aliasing and chaotic attention. To eliminate this failure mode, we introduce Phase-Aligned Mixed-Resolution Attention, a training-free mechanism that enforces a consistent and native positional scale, and adds a lightweight Boundary Expand-and-Replace step to smooth resolution transitions. Together, these changes enable reliable, high-fidelity image and video generation at reduced cost, supporting more sustainable and scalable deployment of diffusion models.

Acknowledgements

We are grateful to Meher Gitika Karumuri, Brandon Smith, Amogh Gupta, and Vidya Narayanan for their insightful comments and valuable discussions. This work was supported in part by NSF grants IIS-2123920, IIS-2212046, and by the CCI startup fund at UNC Charlotte.

References

1. Agarwal, S., Mitra, S., Chakraborty, S., Karanam, S., Mukherjee, K., Saini, S.K.: Approximate caching for efficiently serving {Text-to-Image} diffusion models. In: 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24). pp. 1173–1189 (2024)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., et al.: Gpt-neox-20b: An open-source autoregressive language model. In: Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models. pp. 95–136 (2022)
4. Boer Bohan, O.: Taehv: Tiny autoencoder for hunyuan video. `https://github.com/madebyollin/taehv` (2025), accessed: 2026-02-28
5. Bolya, D., Hoffman, J.: Token merging for fast stable diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4599–4603 (2023)
6. Castells, T., Song, H.K., Kim, B.K., Choi, S.: Ld-pruner: Efficient pruning of latent diffusion models using task-agnostic insights. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 821–830 (2024)
7. Chang, S., Pichao, W., Tang, J., Wang, F., Yang, Y.: Sparsedit: Token sparsification for efficient diffusion transformer. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
8. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis (2023)
9. Chen, L., Meng, Y., Tang, C., Ma, X., Jiang, J., Wang, X., Wang, Z., Zhu, W.: Q-dit: Accurate post-training quantization for diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28306–28315 (2025)
10. Chen, P., Shen, M., Ye, P., Cao, J., Tu, C., Bouganis, C.S., Zhao, Y., Chen, T.: A training-free acceleration method tailored for diffusion transformers. arXiv preprint arXiv:2406.01125 (2024)
11. Chen, P., Zeng, X., Zhao, M., Ye, P., Shen, M., Cheng, W., Yu, G., Chen, T.: Sparse-vdit: Unleashing the power of sparse attention to accelerate video diffusion transformers. arXiv preprint arXiv:2506.03065 (2025)
12. Chen, S., Wong, S., Chen, L., Tian, Y.: Extending context window of large language models via positional interpolation. arXiv preprint arXiv:2306.15595 (2023)
13. Chen, X., Liu, N., Zhu, Y., Feng, F., Tang, J.: Edt: An efficient diffusion transformer framework inspired by human-like sketching. Advances in Neural Information Processing Systems **37**, 134075–134106 (2024)
14. Chen, Y., Yan, J.: What rotary position embedding can tell us: Identifying query and key weights corresponding to basic syntactic or high-level semantic information. Advances in Neural Information Processing Systems **37**, 54507–54528 (2024)
15. Choudhury, R., Kim, J., Park, J., Yang, E., Jeni, L.A., Kitani, K.M.: Accelerating vision transformers with adaptive patch sizes. arXiv preprint arXiv:2510.18091 (2025)

16. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023)
17. Crowson, K., Baumann, S.A., Birch, A., Abraham, T.M., Kaplan, D.Z., Shippole, E.: Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In: Forty-first International Conference on Machine Learning (2024)
18. Deng, J., Li, S., Wang, Z., Gu, H., Xu, K., Huang, K.: Vq4dit: Efficient post-training vector quantization for diffusion transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 16226–16234 (2025)
19. Ding, Y., Zhang, L.L., Zhang, C., Xu, Y., Shang, N., Xu, J., Yang, F., Yang, M.: Longrope: Extending llm context window beyond 2 million tokens. arXiv preprint arXiv:2402.13753 (2024)
20. Ding, Z., Zhang, M., Wu, J., Tu, Z.: Patched denoising diffusion models for high-resolution image synthesis. In: The Twelfth International Conference on Learning Representations (2024)
21. Droste, R., Jiao, J., Noble, J.A.: Unified image and video saliency modeling. In: European Conference on Computer Vision. pp. 419–435. Springer (2020)
22. Du, Z., Fu, Y., Wang, L., Qian, J., Luo, X., Lin, Y.C.: Fewer denoising steps or cheaper per-step inference: Towards compute-optimal diffusion model deployment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3001–3010 (2025)
23. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv e-prints pp. arXiv–2407 (2024)
24. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
25. Fang, G., Li, K., Ma, X., Wang, X.: Tinyfusion: Diffusion transformers learned shallow. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18144–18154 (2025)
26. Fang, G., Ma, X., Wang, X.: Structural pruning for diffusion models. In: Advances in Neural Information Processing Systems (2023)
27. Fei, Z., Qiu, D., Li, D., Yu, C., Fan, M.: Video diffusion transformers are in-context learners. arXiv preprint arXiv:2412.10783 (2024)
28. Feng, W., Yang, C., An, Z., Huang, L., Diao, B., Wang, F., Xu, Y.: Relational diffusion distillation for efficient image generation. In: Proceedings of the 32nd ACM international conference on multimedia. pp. 205–213 (2024)
29. Feng, Y., Yang, M., Yang, S., Zhang, S., Yu, J., Zhao, Z., Liu, Y., Jiang, J., Guo, C.: Romantex: Decoupling 3d-aware rotary positional embedded multi-attention network for texture synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17203–17213 (2025)
30. Ganjdanesh, A., Shirkavand, R., Gao, S., Huang, H.: Not all prompts are made equal: Prompt-based pruning of text-to-image diffusion models. arXiv preprint arXiv:2406.12042 (2024)
31. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

32. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
33. Heo, B., Park, S., Han, D., Yun, S.: Rotary position embedding for vision transformer. In: European Conference on Computer Vision. pp. 289–305. Springer (2024)
34. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. Journal of Machine Learning Research **23**(47), 1–33 (2022)
35. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
36. Issachar, N., Yariv, G., Benaim, S., Adi, Y., Lischinski, D., Fattal, R.: Dype: Dynamic position extrapolation for ultra high resolution diffusion. arXiv preprint arXiv:2510.20766 (2025)
37. Jeong, J., Han, S., Kim, J., Kim, S.J.: Latent space super-resolution for higher-resolution image generation with diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2355–2365 (2025)
38. Jeong, W., Lee, K., Seo, H., Chun, S.Y.: Upsample what matters: Region-adaptive latent sampling for accelerated diffusion transformers. arXiv preprint arXiv:2507.08422 (2025)
39. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. arXiv preprint arXiv:2410.05954 (2024)
40. Kahatapitiya, K., Liu, H., He, S., Liu, D., Jia, M., Zhang, C., Ryoo, M.S., Xie, T.: Adaptive caching for faster video generation with diffusion transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15240–15252 (2025)
41. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021)
42. Kümmerer, M., Theis, L., Bethge, M.: Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. arXiv preprint arXiv:1411.1045 (2014)
43. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742, accessed: 2026-02-28
44. LanguageBind: Open-sora-plan-v1.1.0 dataset. https://huggingface.co/datasets/LanguageBind/Open-Sora-Plan-v1.1.0 (2024), accessed: 2026-02-28. This dataset package includes content sourced from Pexels.
45. Li, H., Lal, S., Li, Z., Xie, Y., Wang, Y., Zou, Y., Majumder, O., Manmatha, R., Tu, Z., Ermon, S., et al.: Efficient scaling of diffusion transformers for text-to-image generation. arXiv preprint arXiv:2412.12391 (2024)
46. Li, M., Lin, Y., Zhang, Z., Cai, T., Li, X., Guo, J., Xie, E., Meng, C., Zhu, J.Y., Han, S.: Svdquant: Absorbing outliers by low-rank components for 4-bit diffusion models. arXiv preprint arXiv:2411.05007 (2024)
47. Li, S., Le, H., Xu, J., Salzmann, M.: Enhancing compositional text-to-image generation with reliable random seeds. ICLR (2025)

48. Li, S., Liu, C., Zhang, T., Le, H., Süsstrunk, S., Salzmann, M.: Controlling the fidelity and diversity of deep generative models via pseudo density. TMLR (2024)
49. Li, Y., Wang, H., Jin, Q., Hu, J., Chemerys, P., Fu, Y., Wang, Y., Tulyakov, S., Ren, J.: Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. Advances in Neural Information Processing Systems **36**, 20662–20678 (2023)
50. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., et al.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748 (2024)
51. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
52. Linardos, A., Kümmerer, M., Press, O., Bethge, M.: Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12919–12928 (2021)
53. Liu, F., Zhang, S., Wang, X., Wei, Y., Qiu, H., Zhao, Y., Zhang, Y., Ye, Q., Wan, F.: Timestep embedding tells: It's time to cache for video diffusion model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7353–7363 (2025)
54. Liu, Z., Yang, Y., Zhang, C., Zhang, Y., Qiu, L., You, Y., Yang, Y.: Region-adaptive sampling for diffusion transformers. arXiv preprint arXiv:2502.10389 (2025)
55. Lou, J., Luo, W., Liu, Y., Li, B., Ding, X., Hu, W., Cao, J., Li, Y., Ma, C.: Token caching for diffusion transformer acceleration. arXiv preprint arXiv:2409.18523 (2024)
56. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in neural information processing systems **35**, 5775–5787 (2022)
57. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. Machine Intelligence Research pp. 1–22 (2025)
58. Lu, Z., Wang, Z., Huang, D., Wu, C., Liu, X., Ouyang, W., Bai, L.: Fit: Flexible vision transformer for diffusion model. arXiv preprint arXiv:2402.12376 (2024)
59. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
60. Lv, Z., Si, C., Song, J., Yang, Z., Qiao, Y., Liu, Z., Wong, K.Y.K.: Fastercache: Training-free video diffusion model acceleration with high quality. arXiv preprint arXiv:2410.19355 (2024)
61. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
62. Ma, X., Fang, G., Bi Mi, M., Wang, X.: Learning-to-cache: Accelerating diffusion transformer via layer caching. Advances in Neural Information Processing Systems **37**, 133282–133304 (2024)
63. Ma, X., Fang, G., Wang, X.: Deepcache: Accelerating diffusion models for free. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15762–15772 (2024)

64. Ma, Z., Wei, L., Wang, F., Zhang, S., Tian, Q.: Magcache: Fast video generation with magnitude-aware cache. arXiv preprint arXiv:2506.09045 (2025)
65. Ostmeier, S., Axelrod, B., Moseley, M.E., Chaudhari, A., Langlotz, C.: Liere: Generalizing rotary position encodings. arXiv preprint arXiv:2406.10322 **2**(4) (2024)
66. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
67. Peng, B., Quesnelle, J.: Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation (2023), `https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/`, accessed: 2026-02-28
68. Peng, B., Quesnelle, J., Fan, H., Shippole, E.: Yarn: Efficient context window extension of large language models. arXiv preprint arXiv:2309.00071 (2023)
69. Pernias, P., Rampas, D., Richter, M.L., Pal, C.J., Aubreville, M.: Würstchen: An efficient architecture for large-scale text-to-image diffusion models. arXiv preprint arXiv:2306.00637 (2023)
70. Pexels: Pexels license. `https://www.pexels.com/license/` (2024), accessed: 2026-02-28
71. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
72. Pondaven, A., Siarohin, A., Tulyakov, S., Torr, P., Pizzati, F.: Video motion transfer with diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22911–22921 (2025)
73. Press, O., Smith, N.A., Lewis, M.: Train short, test long: Attention with linear biases enables input length extrapolation. arXiv preprint arXiv:2108.12409 (2021)
74. Qiu, H., Yu, N., Huang, Z., Debevec, P., Liu, Z.: Cinescale: Free lunch in high-resolution cinematic visual generation. arXiv preprint arXiv:2508.15774 (2025)
75. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
76. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
77. Ronen, T., Levy, O., Golbert, A.: Vision transformers with mixed-resolution tokenization. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 4613–4622 (2023), `https://api.semanticscholar.org/CorpusID:257913635`, accessed: 2026-02-28
78. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
79. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
80. Seo, H., Jeong, W., Seo, J.s., Chun, S.Y.: Skrr: Skip and re-use text encoder layers for memory efficient text-to-image generation. arXiv preprint arXiv:2502.08690 (2025)
81. Shang, Y., Yuan, Z., Xie, B., Wu, B., Yan, Y.: Post-training quantization on diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1972–1981 (2023)

82. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
83. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. arXiv preprint arXiv:2303.01469 (2023)
84. Su, C., Ma, X., Su, J., Wang, Y.: Sat-hmr: Real-time multi-person 3d mesh estimation via scale-adaptive tokens. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16796–16806 (2025)
85. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
86. Sun, Y., Dong, L., Patra, B., Ma, S., Huang, S., Benhaim, A., Chaudhary, V., Song, X., Wei, F.: A length-extrapolatable transformer. In: Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers). pp. 14590–14604 (2023)
87. Tang, B., Zheng, B., Paul, S., Xie, S.: Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28586–28595 (2025)
88. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., et al.: Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)
89. Team, I.: Internlm: A multilingual language model with progressively enhanced capabilities (2023)
90. Teng, J., Zheng, W., Ding, M., Hong, W., Wangni, J., Yang, Z., Tang, J.: Relay diffusion: Unifying diffusion process across resolutions for image synthesis. arXiv preprint arXiv:2309.03350 (2023)
91. Tian, Y., Xia, X., Ren, Y., Lin, S., Wang, X., Xiao, X., Tong, Y., Yang, L., Cui, B.: Training-free diffusion acceleration with bottleneck sampling. arXiv preprint arXiv:2503.18940 (2025)
92. Tian, Y., Tu, Z., Chen, H., Hu, J., Xu, C., Wang, Y.: U-dits: Downsample tokens in u-shaped diffusion transformers. Advances in Neural Information Processing Systems **37**, 51994–52013 (2024)
93. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
94. Wan, B., Zheng, T., Chen, Z., Wang, Y., Wang, J.: Pruning for sparse diffusion models based on gradient flow. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
95. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
96. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: AAAI (2023)
97. Wimbauer, F., Wu, B., Schoenfeld, E., Dai, X., Hou, J., He, Z., Sanakoyeu, A., Zhang, P., Tsai, S., Kohler, J., et al.: Cache me if you can: Accelerating diffusion models through block caching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6211–6220 (2024)
98. Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J.H., Wang, A., Sun, W.S., Yan, Q., Lin, W.: Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: International Conference on Computer Vision (ICCV) (2023)

99. Wu, H., Xu, J., Le, H., Samaras, D.: Importance-based token merging for efficient image and video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4983–4995 (October 2025)
100. Xie, E., Chen, J., Zhao, Y., Yu, J., Zhu, L., Wu, C., Lin, Y., Zhang, Z., Li, M., Chen, J., et al.: Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. arXiv preprint arXiv:2501.18427 (2025)
101. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
102. Xu, J., Le, H., Samaras, D.: Assessing sample quality via the latent space of generative models. In: ECCV (2024)
103. Xu, J., Le, H., Shu, Z., Wang, Y., Tsai, Y.H., Samaras, D.: Learning frame-wise emotion intensity for audio-driven talking-head generation. arXiv (2024)
104. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
105. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
106. You, H., Barnes, C., Zhou, Y., Kang, Y., Du, Z., Zhou, W., Zhang, L., Nitzan, Y., Liu, X., Lin, Z., et al.: Layer-and timestep-adaptive differentiable token compression ratios for efficient diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18072–18082 (2025)
107. Yu, H., Jiang, T., Jia, S., Yan, S., Liu, S., Qian, H., Li, G., Dong, S., Yuan, C.: Comrope: Scalable and robust rotary position embedding parameterized by trainable commuting angle matrices. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4508–4517 (2025)
108. Zhang, E., Tang, J., Ning, X., Zhang, L.: Training-free and hardware-friendly acceleration for diffusion models via similarity-based token pruning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9878–9886 (2025)
109. Zhang, E., Xiao, B., Tang, J., Ma, Q., Zou, C., Ning, X., Hu, X., Zhang, L.: Token pruning for caching better: 9 times acceleration on stable diffusion for free. arXiv preprint arXiv:2501.00375 (2024)
110. Zhang, J., Huang, Q., Liu, J., Guo, X., Huang, D.: Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
111. Zhang, J., Huang, Q., Liu, J., Guo, X., Huang, D.: Ultra-high-resolution image synthesis: Data, method and evaluation (2025), arXiv:2506.01331
112. Zhang, L., Ma, K.: Accelerating diffusion models with one-to-many knowledge distillation. arXiv preprint arXiv:2410.04191 (2024)
113. Zhang, Y., Jin, E., Dong, Y., Khakzar, A., Torr, P., Stegmaier, J., Kawaguchi, K.: Effortless efficiency: Low-cost pruning of diffusion models. arXiv preprint arXiv:2412.02852 (2024)
114. Zhao, M., He, G., Chen, Y., Zhu, H., Li, C., Zhu, J.: Riflex: A free lunch for length extrapolation in video diffusion transformers. arXiv preprint arXiv:2502.15894 (2025)
115. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. NeurIPS (2023)

116. Zheng, K., Lu, C., Chen, J., Zhu, J.: Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. Advances in Neural Information Processing Systems **36**, 55502–55542 (2023)
117. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
118. Zhuo, L., Du, R., Xiao, H., Li, Y., Liu, D., Huang, R., Liu, W., Zhu, X., Wang, F.Y., Ma, Z., et al.: Lumina-next: Making lumina-t2x stronger and faster with next-dit. Advances in Neural Information Processing Systems **37**, 131278–131315 (2024)
119. Zou, C., Liu, X., Liu, T., Huang, S., Zhang, L.: Accelerating diffusion transformers with token-wise feature caching. arXiv preprint arXiv:2410.05317 (2024)