SPATIALLY ADAPTIVE MULTI-RESOLUTION TOKEN MERGING FOR EFFICIENT DIFFUSION INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion transformers repeatedly process dense spatial token grids, making high-resolution generation computationally expensive. Existing training-free token-merging methods primarily use global token matching or ranking, overlooking that redundancy in generated images is spatially structured. We formulate diffusion token reduction as budgeted hierarchical spatial-resolution allocation and introduce QUADMERGE, which converts classifier-free-guidance residuals into a multi-resolution partition of 1×1, 2×2, and 4×4 regions. QUADMERGE conservatively coarsens only uniformly low-importance regions, aggregates them with importance-weighted pooling, couples the plan with timestep-adaptive compression, and caches allocation plans across compatible transformer blocks. At a matched token budget, QUADMERGE improves both generation quality and measured latency over global token-merging baselines, and the margin grows under aggressive compression. On PixArt-α, a 30% token reduction costs 0.09 FID ($36.60 \rightarrow 36.69$) versus 1.05 for importance matching and 12.48 for ToMe; unlike those matchers, it stays below uncompressed wall-clock cost. At a 70% reduction the FID gap over importance matching grows to 7.92, and the same allocation transfers to Stable Diffusion 2.1. The method requires no retraining, no additional learned modules, and no modification to the diffusion sampler.

1 INTRODUCTION

Modern text-to-image diffusion models Ho et al. (2020); Rombach et al. (2022) rely on attention-based backbones, and diffusion transformers in particular repeatedly process a dense latent token grid over many denoising steps, so inference cost scales with both spatial resolution and sampling length Peebles & Xie (2023); Chen et al. (2024); Ma et al. (2024a). As the field has moved to larger backbones and higher output resolutions Podell et al. (2024); Esser et al. (2024), efficiency has become a first-class design concern Xie et al. (2024). Reducing redundant token computation without retraining the generator is therefore central to efficient diffusion inference.

Token merging is an attractive training-free approach: instead of evaluating all spatial tokens independently, it aggregates redundant tokens before expensive operations and restores the original layout afterward. The idea originates in discriminative vision transformers Dosovitskiy (2020), where pruning Rao et al. (2021); Liang et al. (2022) and adaptive token budgets Yin et al. (2022) showed that many patch tokens are dispensable. ToMe introduced bipartite token merging Bolya et al. (2023), ToMeSD adapted it to Stable Diffusion Bolya & Hoffman (2023), and importance-based merging showed that the classifier-free-guidance residual is a useful prompt-sensitive importance signal for diffusion models Wu et al. (2025).

However, these methods still treat the latent grid largely as an unordered token set: global source–destination matching or ranking decides which individual tokens to merge, without modelling that image redundancy is local and multi-scale. Smooth backgrounds contain contiguous redundant regions that tolerate coarser representation, whereas foreground objects, boundaries, and prompt-sensitive details require fine spatial resolution. Under aggressive compression, ignoring this structure merges unrelated areas, blurs boundaries, and removes detail unnecessarily.

We argue that diffusion token reduction should be formulated not as token selection but as *spatial-resolution allocation*: under a fixed removal budget, each local region is assigned an appropriate

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

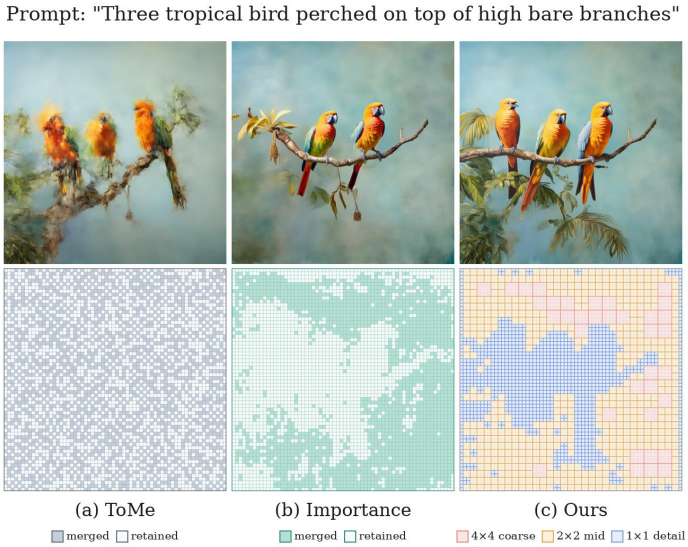


Figure 1: **Spatially Adaptive Quadtree Token Merging.** Our method adapts merge granularity across the latent grid, retaining fine groups in higher-risk areas and representing lower-risk regions coarsely. At a common target ratio and near-matched realized compute, it better preserves structure than ToMe and importance-based merging. The second row visualizes the resulting token layout.

resolution. We introduce QUADMERGE, a spatially adaptive multi-resolution token-merging framework. Given a guidance-derived importance map, QUADMERGE constructs a non-overlapping partition of dyadic regions: higher-risk positions are favored for 1×1 representation, while low-risk redundant regions are represented by 2×2 or 4×4 groups. This converts a continuous importance field into a coherent merge plan, replacing global token matching with budgeted local resolution planning.

The design follows the structure of diffusion inference. Conservative max-risk scoring ranks a coarse region favorably only when every token inside it has low importance, discouraging foreground and boundary tokens from being absorbed into coarse groups; selected regions are combined by importance-normalized aggregation and broadcast back to their original positions after compressed attention. Timestep-adaptive compression merges more aggressively in early noisy steps and less during late refinement, and plans are cached across compatible blocks within a step to amortize planning overhead. Timestep-aware reduction has precedents; our contribution is its coupling with CFG-guided multi-resolution planning. Acting within a step, token reduction is conceptually complementary to shortening the trajectory by distillation Meng et al. (2023), caching activations across adjacent steps Ma et al. (2024b); Li et al. (2024); So et al. (2024), and compressing the network by pruning **?** or quantization Li et al. (2023).

Experiments on PixArt-α show that QUADMERGE improves both quality and measured latency over global matching at near-matched realized compute, with the largest gains under aggressive compression, and that it transfers to Stable Diffusion 2.1. Throughout, we report distribution-level quality (FID, CLIP) separately from matched-output consistency, because smooth outputs can stay close to a baseline sample under full-reference metrics while losing distributional quality and prompt alignment.

Our contributions are threefold:

- We formulate diffusion token reduction as hierarchical spatial-resolution allocation under an explicit removal budget, replacing globally selected token pairs with a coherent multi-resolution partition.
- We introduce QUADMERGE, a training-free method combining conservative CFG-guided region allocation, importance-normalized local aggregation, timestep-adaptive compression, and cached merge plans.

- We provide controlled comparisons on both a diffusion-transformer and a UNet backbone, including realized token-count, full-transformer FLOP, and latency audits on the primary PixArt benchmark, and analyze the interaction of guidance, spatial planning, and timestep scheduling.

2 Related Work

Token reduction in vision transformers. Visual transformers process images as sequences of patch tokens (Dosovitskiy, 2020), making token reduction an effective way to lower attention cost. Token Merging (ToMe) progressively combines similar tokens through training-free bipartite matching (Bolya et al., 2023), while QuadTree Attention exploits spatial hierarchy through coarse-to-fine attention (Tang et al., 2022). These methods motivate reducing spatially redundant computation, but they do not construct prompt-dependent resolution assignments for the repeated denoising process.

Token reduction for diffusion inference. ToMeSD adapts bipartite feature matching to Stable Diffusion and restores the dense layout after compressed attention (Bolya & Hoffman, 2023). Importance-based token merging uses the classifier-free-guidance residual to preserve prompt-relevant tokens during global matching (Wu et al., 2025). ToMA improves merge–unmerge efficiency through representative selection, latent locality, and pattern reuse (Lu et al., 2025), whereas ToDo avoids matching by downsampling keys and values on a fixed spatial grid (Smith et al., 2024). These approaches either select token correspondences or impose a fixed reduction pattern rather than explicitly allocating spatial resolution according to prompt-conditioned importance. Denoising-stage-aware reduction has been explored by AT-EDM and SDTM (Wang et al., 2024; Fang et al., 2025). The distinctive contribution of QuadMerge is to formulate token reduction as prompt-conditioned spatial-resolution allocation: given a CFG-derived importance field and an explicit removal budget, its planner directly constructs a non-overlapping partition whose regions remain at 1×1 resolution or are represented at 2×2 or 4×4 resolution. This spatial planner is coupled with a mean-preserving capped timestep schedule, coordinating where and when compression is applied. Unlike fixed-grid downsampling, its resolution pattern adapts across images; unlike pair matching, it directly produces a coherent multi-resolution partition.

3 Method

3.1 Overview and Formulation

Figure 2 summarizes QuadMerge. It derives a prompt-sensitive importance field from the classifier-free-guidance (CFG) residual, allocates a removal budget among non-overlapping 1×1, 2×2, and 4×4 regions, and shifts more compression to early denoising steps. The resulting plan is reused within each step to compress the attention branches and broadcast their outputs to the original layout; the feed-forward branch remains dense.

At step t and block ℓ, let $X_{t,\ell} \in \mathbb{R}^{N_\ell \times C_\ell}$ denote tokens on an $H_\ell \times W_\ell$ lattice, where $N_\ell = H_\ell W_\ell$. A group assignment $g_{t,\ell}$ induces merge and broadcast operators M_g and U_g:

$$\widetilde{X}_{t,\ell} = M_{g_{t,\ell}}(X_{t,\ell}), \qquad \widehat{Y}_{t,\ell} = U_{g_{t,\ell}}\Big(F_\ell(\widetilde{X}_{t,\ell})\Big), \tag{1}$$

where $\widetilde{X}_{t,\ell} \in \mathbb{R}^{K_{t,\ell} \times C_\ell}$ is processed by self-attention or the spatial-query stream of cross-attention. Restoring the dense layout preserves pretrained parameters and downstream tensor shapes. For target ratio r_t, the requested and realized removal budgets are

$$\overline{R}_{t,\ell} = \lfloor r_t N_\ell \rfloor, \qquad R_{t,\ell} = N_\ell - K_{t,\ell} \le \overline{R}_{t,\ell}. \tag{2}$$

The inequality reflects discrete group sizes, boundaries, and non-overlap constraints. We therefore report compute using realized $K_{t,\ell}$ rather than treating the configured ratio as an exact sequence length.

3.2 Guidance-Derived Spatial Importance

Classifier-free guidance Ho & Salimans (2022) already computes conditional and unconditional predictions. Following importance-based merging Wu et al. (2025), we use their residual $\Delta_t =$

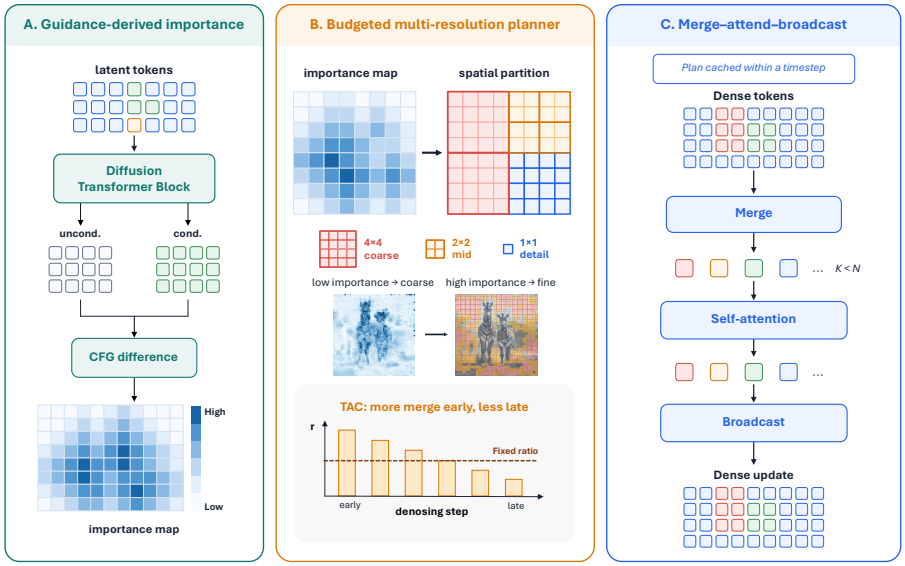


Figure 2: **Overview of QUADMERGE.** A CFG-derived importance field guides a budgeted multi-resolution partition. TAC varies compression over denoising time, and the cached plan merges tokens for attention before broadcasting a dense update.

$\epsilon_t^y - \epsilon_t^\emptyset$ to define

$$I_t(u,v) = \frac{1}{C_\epsilon} \sum_{m=1}^{C_\epsilon} \left|\Delta_t(m,u,v)\right|, \tag{3}$$

where C_ϵ covers the noise-prediction channels. Area interpolation maps I_t to each block resolution. The signal requires no additional denoiser evaluation; planning uses the latest available residual, introducing a one-step lag.

3.3 MULTI-RESOLUTION SPATIAL PLANNING

The planner turns the importance field into a merge plan: under budget $\overline{R}_{t,\ell}$, each location is assigned a resolution in $\mathcal{L} = \{1, 2, 4\}$ so that smooth, low-risk regions share a token while prompt-sensitive structure stays at 1×1. This replaces global source–destination matching. A candidate $s\times s$ block $B_s(p)$ aligned to a shifted grid is represented by one token and removes $\Delta(B_s) = s^2 - 1$ tokens; uncovered locations remain singletons. Candidates are scored by their maximum importance,

$$\rho(B_s(p)) = \max_{i \in B_s(p)} I_{t,\ell}(i), \tag{4}$$

so a block is cheap only when *every* token inside it is unimportant. Mean pooling would let one salient corner be absorbed by a quiet neighbourhood; the max rejects that merge. Mean-risk scoring is used only as an ablation.

Planning is coarse-to-fine. The 4×4 level runs first and marks accepted cells as assigned, so a later 2×2 block cannot overlap them. At each size s, unassigned candidates are sorted by ρ and accepted from lowest risk until the level quota is spent or no feasible tile remains. The 4×4 quota is

$$\overline{R}_{t,\ell}^{(4)} = \left\lfloor \beta_4 \, \overline{R}_{t,\ell} \right\rfloor, \tag{5}$$

and any unused portion, together with $\overline{R}_{t,\ell} - \overline{R}_{t,\ell}^{(4)}$, is given to 2×2 groups. Raising β_4 spends more budget on large quiet regions and frees singletons for the subject; $\beta_4 = 0$ forbids 4×4 tiles, and $\beta_4 = 1$ allows the coarse level to take the entire budget. The displayed configuration uses $\beta_4 = .25$. Because tiles have discrete sizes and cannot overlap, the realized removal satisfies $R_{t,\ell} \le \overline{R}_{t,\ell}$.

A shared offset (o_x, o_y) drawn once per step, with $o_x, o_y \in \{0, \ldots, s_{\max} - 1\}$, shifts every candidate grid and reduces fixed-boundary artefacts. The output is a contiguous resolution map and a group index for each token. No all-pairs feature-similarity matrix is formed: the planner scores $\mathcal{O}(N_\ell)$ aligned tiles and sorts them in $\mathcal{O}(N_\ell \log N_\ell)$.

3.4 MERGE–ATTEND–BROADCAST

For group $\mathcal{G}_q = \{i \mid g(i) = q\}$, QUADMERGE forms an importance-weighted representative and broadcasts its processed output:

$$\widetilde{x}_q = \sum_{i \in \mathcal{G}_q} \omega_i x_i, \qquad \omega_i = \frac{I_{t,\ell}(i) + \varepsilon}{\sum_{j \in \mathcal{G}_q} (I_{t,\ell}(j) + \varepsilon)}, \qquad \widehat{y}_i = \widetilde{y}_{g(i)}. \tag{6}$$

Here $\varepsilon > 0$ prevents division by zero and $\widetilde{Y} = F_\ell(\widetilde{X})$. The normalized weights define a convex representative; attention itself is unchanged. Broadcasting restores the dense residual update, making the operation training-free and shape preserving.

3.5 TIMESTEP-ADAPTIVE COMPRESSION AND PLAN REUSE

Early denoising steps set global layout and tolerate coarser spatial tokens; late steps refine edges and text and need a longer sequence. Timestep-adaptive compression (TAC) therefore varies the requested ratio over time while keeping a prescribed mean r. Merging starts at step t_0 (step 4 of 20 on PixArt; step 10 on Stable Diffusion). For $t \in \{t_0, \dots, T-1\}$,

$$r_t = r_{\min} + (r_{\max} - r_{\min})\phi_t^\alpha, \qquad \phi_t = \frac{T-1-t}{T-1-t_0}, \tag{7}$$

and $r_t = 0$ before t_0. The normalized time ϕ_t is 1 at the first merged step and 0 at the last, so r_t falls from $r_{\max}$ to $r_{\min}$. We use a linear ramp $\alpha = 1$. The endpoints are mean-preserving and capped,

$$r_{\max} = \min\big((1+\delta)r, r_{\text{cap}}\big), \qquad r_{\min} = 2r - r_{\max}, \tag{8}$$

with $\delta = .3$. Because a linear schedule averages $\frac{1}{2}$, Eq. 8 keeps the arithmetic mean equal to r even after the upper tail hits r_{cap}. The cap is operator-dependent: global importance matching must retain destination tokens and is limited to $r_{\text{cap}} = .75$, whereas the dyadic planner uses $r_{\text{cap}} = .95$. The two TAC instantiations therefore share the target mean but not the same feasible interval, which is why the ablation treats planner–schedule coupling as an interaction rather than an identical-schedule decomposition. Attention cost is quadratic in K, so a mean-preserving r_t still does not imply equal realized FLOPs; we compare recorded $K_{t,\ell}$ and transformer compute.

The spatial plan depends on (I_t, r_t) and the step offset, not on block activations. Blocks that share a lattice and budget therefore reuse one assignment and its merge weights within the step. Planning and sorting run once; each patched block only gathers, attends, and broadcasts. Features remain block-specific. On PixArt this cache covers the 14 merged transformer blocks at the 64×64 grid.

Computational Cost. Self-attention decreases from $\mathcal{O}(N_\ell C_\ell^2 + N_\ell^2 C_\ell)$ to $\mathcal{O}(K_{t,\ell} C_\ell^2 + K_{t,\ell}^2 C_\ell)$; spatial cross-attention terms become linear in $K_{t,\ell}$, while the feed-forward branch remains dense. Merge and broadcast cost $\mathcal{O}(N_\ell C_\ell)$. The planner scores $\mathcal{O}(N_\ell)$ candidates and sorts them in $\mathcal{O}(N_\ell \log N_\ell)$ without all-pairs feature matching. We report both realized FLOPs and end-to-end latency to include dense computation and merge–broadcast overhead.

4 EXPERIMENTS

Three questions organize this section: whether QUADMERGE improves generation quality at closely matched realized compute, whether its planner converts token reduction into lower wall-clock cost and how this compares with reducing denoising steps, and which components account for the quality difference. PixArt-α is the primary diffusion-transformer benchmark; Stable Diffusion 2.1 tests transfer to a UNet.

4.1 EXPERIMENTAL SETUP

Models, data, and protocol. On PixArt-α Chen et al. (2024) we generate 5,000 images from COCO val2014 captions Lin et al. (2014) with 20 denoising steps. Native resolution binning puts sampling on the 1024×1024 bin, a 64×64 token grid ($N = 4096$); decoded images are resized to 768×768 for evaluation, and all token, latency, and FLOP counts refer to that grid. On Stable

Table 1: **Quality and realized PixArt compute at common target ratios.** Both backbones use $r \in \{.30, .50, .70\}$. PixArt quality in (a) is the equal-quota $\beta_4 \approx .5$ evaluation. ToDo is a clean-room fixed-grid K/V control at native 2×2 stride. (b) Recorded sequence length, full-transformer TFLOPs, and synchronized A100-80GB seconds per image (one warm-up, five timed runs). Bold marks the best compressed value in each ratio group in (a) and the lowest compressed latency in (b).

(a) Generation quality

		PixArt-α 64^2 grid, 20 steps				SD 2.1 512^2, 20-step DDIM			
r	Method	FID ↓	CLIP ↑	PSNR ↑	LPIPS ↓	FID ↓	CLIP ↑	PSNR ↑	LPIPS ↓
–	Baseline	36.60	31.06	–	–	28.11	30.43	–	–
2×2	ToDo	65.83	30.01	17.71	.391	–	–	–	–
.30	ToMe	49.08	30.60	**18.68**	**.318**	30.36	**30.09**	14.52	.462
	Importance	37.65	30.84	17.73	.321	**30.21**	30.08	14.54	**.461**
	Ours	**36.69**	**30.91**	17.45	.322	30.31	30.08	**14.68**	.465
.50	ToMe	76.99	29.71	**17.48**	**.396**	31.35	30.04	14.40	.497
	Importance	42.49	30.52	15.97	.429	30.78	30.09	14.39	.502
	Ours	**38.93**	**30.69**	16.01	.410	**30.35**	**30.14**	**14.83**	**.485**
.70	ToMe	102.4	28.54	**16.68**	**.452**	34.24	30.04	14.37	.541
	Importance	56.91	29.65	14.88	.521	32.29	30.01	14.26	.548
	Ours	**48.99**	**30.20**	15.70	.455	**31.74**	**30.16**	**14.92**	**.496**

(b) PixArt $\bar{K}$, TFLOPs, latency

r	Method	$\bar{K}$	TFLOPs	Lat.↓
–	Baseline	4096.0	129.80	2.550
.30	ToMe	2868.0	113.10	2.622
	Importance	2868.0	113.10	2.690
	Ours	2868.7	113.16	**2.500**
.50	ToMe	2048.0	104.11	2.478
	Importance	2048.0	104.11	2.551
	Ours	2049.4	104.27	**2.341**
.70	ToMe	1229.0	96.87	2.364
	Importance	1229.0	96.87	2.439
	Ours	1275.8	97.45	**2.216**

Diffusion 2.1 Rombach et al. (2022) we use the same captions at 512×512 with 20-step DDIM sampling Song et al. (2020). Within a benchmark every method shares caption ordering, latent seeds, and sampler settings. The merge ratio r is the target fraction of spatial tokens removed before the compressed operation. Because TAC, matcher capacity, and spatial packing can change the realized sequence length, we record $K_{t,\ell}$ at every active step and compare the resulting FLOPs rather than treating equal configured ratios as equal compute.

Baselines. ToMe applies bipartite soft matching **?**Bolya & Hoffman (2023). Importance uses the classifier-free-guidance Ho & Salimans (2022) residual for global source–destination matching Wu et al. (2025). Importance+TAC adds our timestep-adaptive compression to that global matcher, separating the schedule from the spatial planner. ToDo Smith et al. (2024) is a clean-room PixArt control that downsamples self-attention keys and values on a fixed 2×2 grid while keeping queries dense. Two no-merging controls shorten the trajectory instead of the token sequence. On PixArt, QUADMERGE uses $\mathcal{L} = \{1, 2, 4\}$ with equal quotas ($\beta_4 \approx .5$) and starts merging at step 4. Blocks 7–20 are patched; self- and cross-attention use the compressed spatial sequence and the MLP remains dense. The Stable Diffusion configuration uses $\mathcal{L} = \{1, 2\}$ and starts at step 10.

Metrics and selection rule. FID Heusel et al. (2017) measures distribution-level quality against the caption-matched COCO reference and CLIP score Radford et al. (2021); Hessel et al. (2021) text–image alignment; PSNR, SSIM Wang et al. (2004), and LPIPS Zhang et al. (2018) measure consistency with the uncompressed sample from the same prompt and seed. Because a smooth output can score well on all three while losing distributional fidelity, we select on FID and CLIP and treat consistency as diagnostic. Additional protocol details and complete per-ratio tables are provided in the supplementary material.

4.2 QUALITY–EFFICIENCY TRADE-OFF

A moderate budget preserves distribution quality. At $r = .30$, QUADMERGE changes FID by only $+0.09$ ($36.60 \to 36.69$) and CLIP by -0.15, whereas Importance changes FID by $+1.05$ and ToMe by $+12.48$. Relative to Importance on the same prompts, QUADMERGE improves FID by 0.96 and CLIP by 0.07. This is the quality-oriented operating point: Table 1(b) shows that the three methods are within 0.06 TFLOPs, so the quality gap is not explained by extra compute.

The advantage widens under aggressive compression. Among the merge methods, QUADMERGE obtains the best FID and CLIP in every PixArt group and is the only one keeping CLIP above 30 at all three displayed settings. From $r = .30$ to .70, its FID degrades by 12.30 ($36.69 \to 48.99$), compared with 19.26 for Importance and 53.32 for ToMe. At $r = .70$ the FID

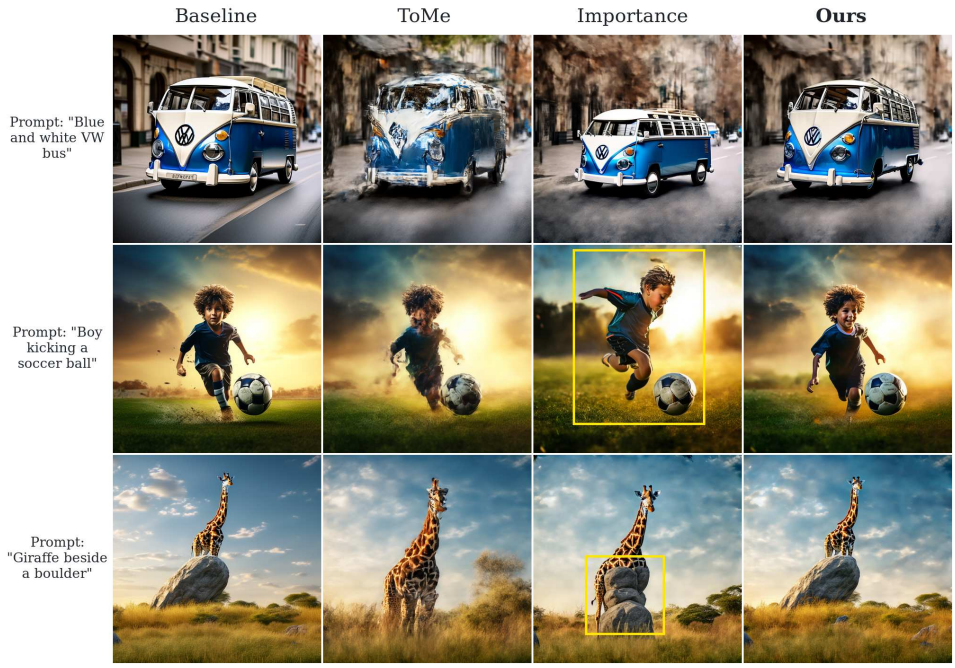


Figure 3: **The two global-matching baselines fail in complementary ways at $r = .50$.** Each row is one prompt, and all four columns share its seed and sampler. ToMe smears texture and breaks object geometry, whereas Importance keeps the subject sharp but can redraw global composition. QUADMERGE (Ours) better preserves both composition and fine structure.

advantage over Importance is 7.92. These high-ratio results characterize robustness rather than the recommended deployment setting; the joint quality and latency results favor $r \approx .3$–$.5$.

Two failure modes explain why consistency must be read with quality. In the matched samples of Figure 3, ToMe merges by feature similarity alone and oversmooths: texture is smeared and geometry breaks, yet the blurred output remains close to the low-frequency structure of the uncompressed sample and therefore attains the highest PixArt PSNR at every ratio, including 18.68 at $r = .30$ while its FID is 49.08. Importance retains sharper subjects but can alter global composition, producing the weakest paired consistency at high ratios (14.88 PSNR and .521 LPIPS). QUADMERGE obtains the best FID and CLIP in every PixArt group. Its LPIPS is .322, .410, and .455; ToMe remains slightly closer on LPIPS because the blurred output stays near the uncompressed low frequencies, so a high consistency score does not imply a better sample. The same dissociation appears in the ablation, where a random importance map retains 16.21 PSNR while FID rises to 112.89. We therefore select on FID and CLIP and treat consistency as diagnostic.

Realized compute confirms the fairness of the main comparison. Packing slack is negligible at $r = .30$ and .50: QUADMERGE differs from the global matchers by at most 1.4 mean tokens and 0.16 TFLOPs. At $r = .70$ in Table 1(b), dyadic packing retains 46.8 more tokens on average, and the 0.58-TFLOP difference is 0.6% of the global matchers' compute. This accounting also shows that the quality-oriented $r = .30$ setting reduces full-transformer compute by 12.8%, despite leaving half of the blocks and every MLP dense.

At closely matched compute, spatial planning has lower overhead than global matching. Table 1(b) isolates planning and execution overhead once realized compute is aligned. Global importance matching is 5.5% slower than uncompressed inference at $r = .30$ (2.690 versus 2.550) and only reaches parity near $r = .50$ (2.551). By avoiding the all-pairs matcher, QUADMERGE stays below the baseline at every displayed ratio and is 7.1%, 8.2%, and 9.1% faster than Importance at $r = .30$, .50, and .70. At $r = .50$ it reduces full-transformer compute by 19.7% and end-to-end latency by 8.2% (2.550 → 2.341), while improving FID from 42.49 to 38.93 relative to Importance.

The wall-clock gain is smaller than the FLOP reduction because half of the transformer blocks, every MLP, text encoding, VAE decoding, and memory-bound operations remain unaffected.

Transfer to a UNet backbone. On Stable Diffusion the three methods are within 0.15 FID at $r = .30$, where Importance is nominally best (30.21 versus 30.31). From $r = .50$ onward, QUADMERGE leads every reported metric; at $r = .70$ it improves FID by 0.55 over Importance and 2.50 over ToMe, and it is the only compressed method with CLIP at or above 30.16. The UNet also converts token reduction into a $2.2\times$ measured speedup ($1.582 \to 0.727$ s per image at $r = .30$). Latency remains nearly constant across merge ratios (0.727 versus 0.738 s at $r = .30$ and $.70$), while FID changes by 1.43, further supporting $r \approx .3$–$.5$ as the useful operating range.

4.3 ALTERNATIVE ACCELERATION CONTROLS

Table 2: **Token merging composes with a shortened trajectory.** The first block is the uncompressed baseline at three step counts; the second is QUADMERGE at $\beta_4 = .25$ in two modes. Step reduction is the stronger single lever on this backbone, but the two axes are independent: QUADMERGE at $r = .30$ on a 14-step trajectory matches the 14-step baseline in FID at 26% less nominal compute. Bold marks the lowest latency and compute.

Method	Steps	r	FID↓	Lat. (s)↓	TFLOPs↓
Baseline	20	–	36.66	2.348	129.8
	17	–	36.78	2.031	110.3
	14	–	37.25	1.714	90.9
QUADMERGE	20	.30	36.93	2.322	91.4
	14	.30	37.20	**1.694**	**66.8**

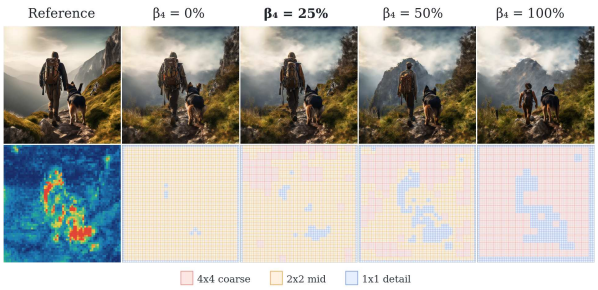
Fixed-grid downsampling is not a substitute for content-adaptive allocation. Table 1(a) places a clean-room ToDo control Smith et al. (2024) at its native 2×2 stride directly under the uncompressed baseline: 65.83 FID and 30.01 CLIP on the same PixArt 5k set. It is better than ToMe at $r = .50$ (76.99 FID) but substantially worse than Importance (42.49) and QUADMERGE (38.93 FID, 30.69 CLIP). Although the operators and compute budgets differ, the gap shows that imposing local structure through a fixed grid is not sufficient; the resolution pattern must adapt to prompt-conditioned importance.

Token merging composes with a shorter trajectory. Whether to shorten the trajectory or reduce per-step cost is a model- and deployment-dependent choice Du et al. (2025). In Table 2, 20-step QUADMERGE at $r = .30$ matches the nominal compute of the 14-step baseline (91.4 versus 90.9 TFLOPs) with a better FID (36.93 versus 37.25). The same ratio on a 14-step trajectory reaches 37.20 FID, matching that baseline (37.25) at 26% less nominal compute (66.8 versus 90.9 TFLOPs) and the lowest latency (1.694 s). Step reduction remains a strong single lever, while token reduction is the remaining axis once the step count is fixed by a solver (Lu et al., 2022) or distilled sampler (Salimans & Ho, 2022; Song & Dhariwal, 2024; Luo et al., 2023).

4.4 ABLATION AND ANALYSIS

Multi-resolution allocation provides a controllable quality–consistency trade-off. Figure 4 varies the coarse-budget fraction at a fixed final token count, isolating where the removal budget is placed. Raising β_4 allows 4×4 groups to absorb more budget in low-importance regions and frees 1×1 singletons for salient content, improving FID from 52.41 at $\beta_4 = 0$ to 48.99 at $\beta_4 \approx .5$. Increasing coarseness also moves the output farther from the uncompressed sample (PSNR $16.76 \to 14.42$, LPIPS $.411 \to .517$ across $\beta_4 = 0 \to 1$), and the fully coarse $\beta_4 = 1$ setting reaches the subject and degrades both FID and CLIP (59.09 and 29.463). Table 1 uses equal quotas ($\beta_4 \approx .5$), the best FID in this sweep (48.99). The fine-heavy $\beta_4 = .25$ setting has the highest CLIP (30.263) and closer paired consistency.

(a) Resolution plans



(b) Component ablation

Variant	FID↓	CLIP↑	PSNR↑	LPIPS↓
Baseline (no merging)	36.60	31.061	–	–
QUADMERGE				
β_4=0 (2×2 only)	52.41	30.259	**16.76**	**.411**
β_4=.25	50.08	**30.263**	16.26	.431
$\beta_4 \approx .5$	**48.99**	30.204	15.70	.455
β_4=1 (4×4 only)	59.09	29.463	14.42	.517
Importance	56.91	29.650	14.88	.521
Importance + TAC	56.30	29.658	14.76	.523
QUADMERGE w/o TAC	57.29	30.029	15.74	.468
Random importance map	112.89	28.440	16.21	.462

Figure 4: **Spatial-allocation ablation at $r = .7$.** (a) Reference output, CFG importance, and resolution plans as β_4 varies at fixed $K/N = .30$; intermediate coarse budgets preserve fine tokens around salient content. (b) Results on 5k matched captions. The β_4 rows are QUADMERGE; the random-map control changes only the spatial signal. Bold marks the best compressed value in each column.

Spatial planning and timestep adaptation reinforce one another. In the $r = .7$ equal-quota ablation, removing TAC worsens FID from 48.99 to 57.29, an 8.30-point change, and the planner without TAC (57.29) is close to plain Importance (56.91). Adding TAC to global Importance improves FID only from 56.91 to 56.30, with CLIP essentially unchanged ($29.650 \rightarrow 29.658$). Equal-quota QUADMERGE therefore outperforms Importance+TAC by 7.31 FID and 0.546 CLIP. Temporal reallocation alone does not recover the benefit of prompt-conditioned multi-resolution planning. The two TAC instantiations share the target mean $r = .70$ but have different feasible intervals, so we interpret this as evidence for their interaction rather than as an identical-schedule decomposition.

Guidance-derived importance is essential to spatial allocation. Replacing the CFG map with reproducible random values while holding the planner and budget fixed increases FID from 48.99 to 112.89, whereas PSNR changes from 15.70 to 16.21. The guidance signal, rather than the dyadic partition alone, therefore determines where coarse representations can be placed without severely damaging distribution quality; paired consistency does not expose this failure.

The allocation framework is robust to local scoring variants. A separate 1k control at $r = .50$ changes only the local risk and aggregation choices. The full configuration obtains 66.85 FID and 30.574 CLIP, uniform aggregation obtains $66.49/30.629$, and average-risk scoring obtains $65.44/30.646$. These differences are small at this sample size and do not establish an independent ranking among the local variants. We retain max-risk scoring and importance-normalized aggregation for their conservative regional interpretation; the strongest evidence supports the CFG-conditioned multi-resolution allocation and its interaction with the denoising trajectory.

5 CONCLUSION

We presented QUADMERGE, a training-free method that treats diffusion token reduction as budgeted spatial-resolution allocation. A CFG-derived importance field becomes a $1\times1/2\times2/4\times4$ partition; conservative scoring and importance-weighted aggregation keep fine tokens on prompt-sensitive structure, while TAC and cached plans decide when to compress. On PixArt-α, this yields the best FID and CLIP among merge methods at every displayed ratio, a nearly lossless $r = .30$ operating point ($+0.09$ FID), and a widening margin under aggressive compression, while staying below uncompressed latency. The same allocation transfers to Stable Diffusion 2.1. Ablations show that guidance, multi-resolution planning, and the timestep schedule must be coupled; a fixed local grid is not a substitute. Moderate compression ($r \approx .3$–$.5$) is the useful operating range.

Limitations. The planner is restricted to a dyadic lattice of square groups, and the CFG signal is available with a one-step lag. Realized FLOP reductions do not fully translate into wall-clock savings, and broader prompts, resolutions, and video remain to be validated.

AI USE STATEMENT

In this work, we used generative AI tools to search for and summarize relevant literature and to edit the manuscript for clarity and readability. We have not used generative AI tools to generate synthetic data, develop theoretical models, formulate or prove mathematical claims, propose hypotheses, design the method or experiments, implement the method, translate text, clean datasets, conduct qualitative analysis, or interpret experimental results. Suggested references were checked against the cited papers, and the edited prose was checked against our technical claims and experimental results. We take responsibility for the final content of this work, including any text produced with the aid of generative AI.

REFERENCES

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *2023 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 4599–4603. IEEE, 2023.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformers for photorealistic text-to-image synthesis. In *International conference on learning representations*, volume 2024, pp. 57611–57640, 2024.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zhenbang Du, Yonggan Fu, Lifu Wang, Jiayi Qian, Xiao Luo, and Yingyan Celine Lin. 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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Haipeng Fang, Sheng Tang, Juan Cao, Enshuo Zhang, Fan Tang, and Tong-Yee Lee. Attend to not attended: Structure-then-detail token merging for post-training dit acceleration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18083–18092. IEEE, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

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

Senmao Li, Taihang Hu, Joost Van De Weijer, Fahad S Khan, Tao Liu, Linxuan Li, Shiqi Yang, Yaxing Wang, Ming-Ming Cheng, and Jian Yang. Faster diffusion: Rethinking the role of the encoder for diffusion model inference. *Advances in neural information processing systems*, 37: 85203–85240, 2024.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17489–17499. IEEE, 2023.

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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

Wenbo Lu, Shaoyi Zheng, Yuxuan Xia, and Shengjie Wang. ToMA: Token merge with attention for diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024a.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15762–15772. IEEE, 2024b.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306. IEEE, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4172–4182, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

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

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Ethan Smith, Nayan Saxena, and Aninda Saha. Todo: Token downsampling for efficient generation of high-resolution images. *arXiv preprint arXiv:2402.13573*, 2024.

Junhyuk So, Jungwon Lee, and Eunhyeok Park. Frdiff: Feature reuse for universal training-free acceleration of diffusion models. In *European Conference on Computer Vision*, pp. 328–344. Springer, 2024.

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

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, volume 2024, pp. 15078–15097, 2024.

Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. *arXiv preprint arXiv:2201.02767*, 2022.

Hongjie Wang, Difan Liu, Yan Kang, Yijun Li, Zhe Lin, Niraj K Jha, and Yuchen Liu. Attention-driven training-free efficiency enhancement of diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16080–16089. IEEE, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Haoyu Wu, Jingyi Xu, Hieu Le, and Dimitris Samaras. Importance-based token merging for efficient image and video generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4983–4995. IEEE, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10799–10808. IEEE, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

A Additional Qualitative Results

The main paper shows matched PixArt samples at $r = .50$ (Figure 3) and the corresponding resolution plans (Figure 4). This section adds further prompts, ratios, the Stable Diffusion transfer, and allocation variants. All panels share the prompt, seed, and sampler within a row.

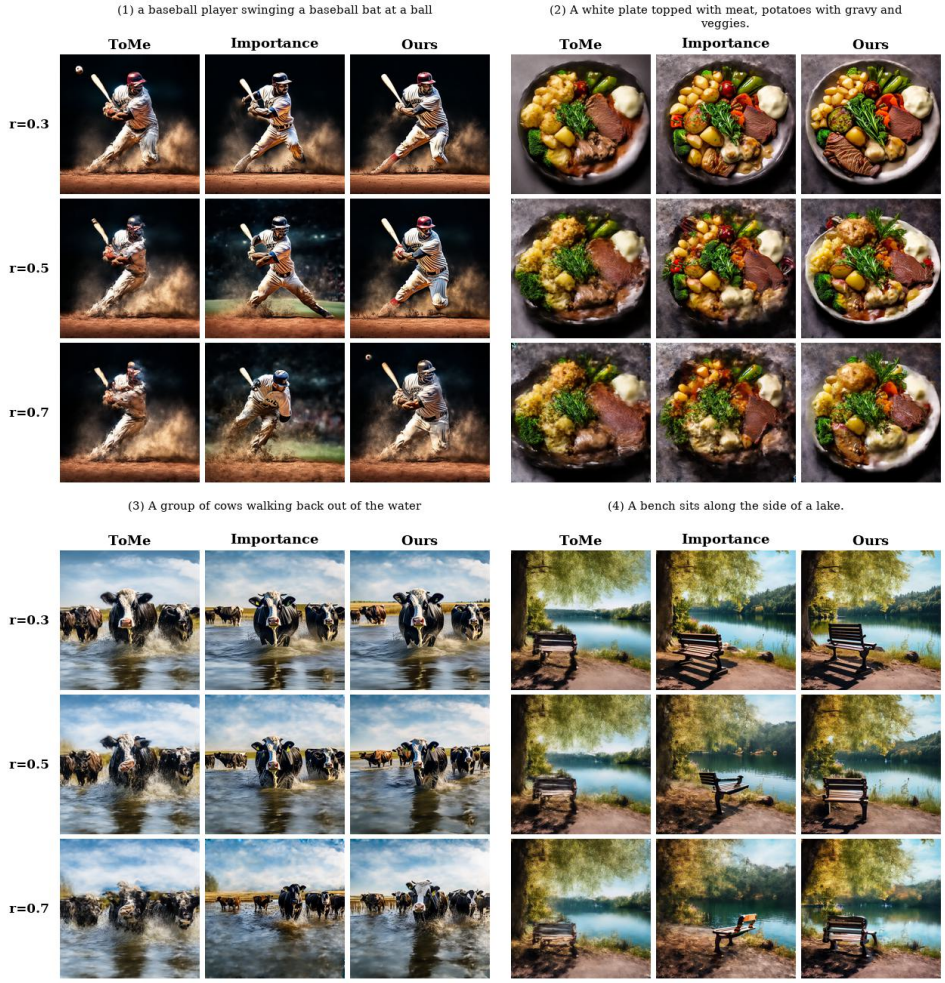


Figure 5: **PixArt comparisons across merge ratios.** Each triplet is one prompt at $r = .30$, $.50$, and $.70$. QuadMerge keeps the batter's pose, the plating, the herd in water, and the lakeside bench as compression increases.

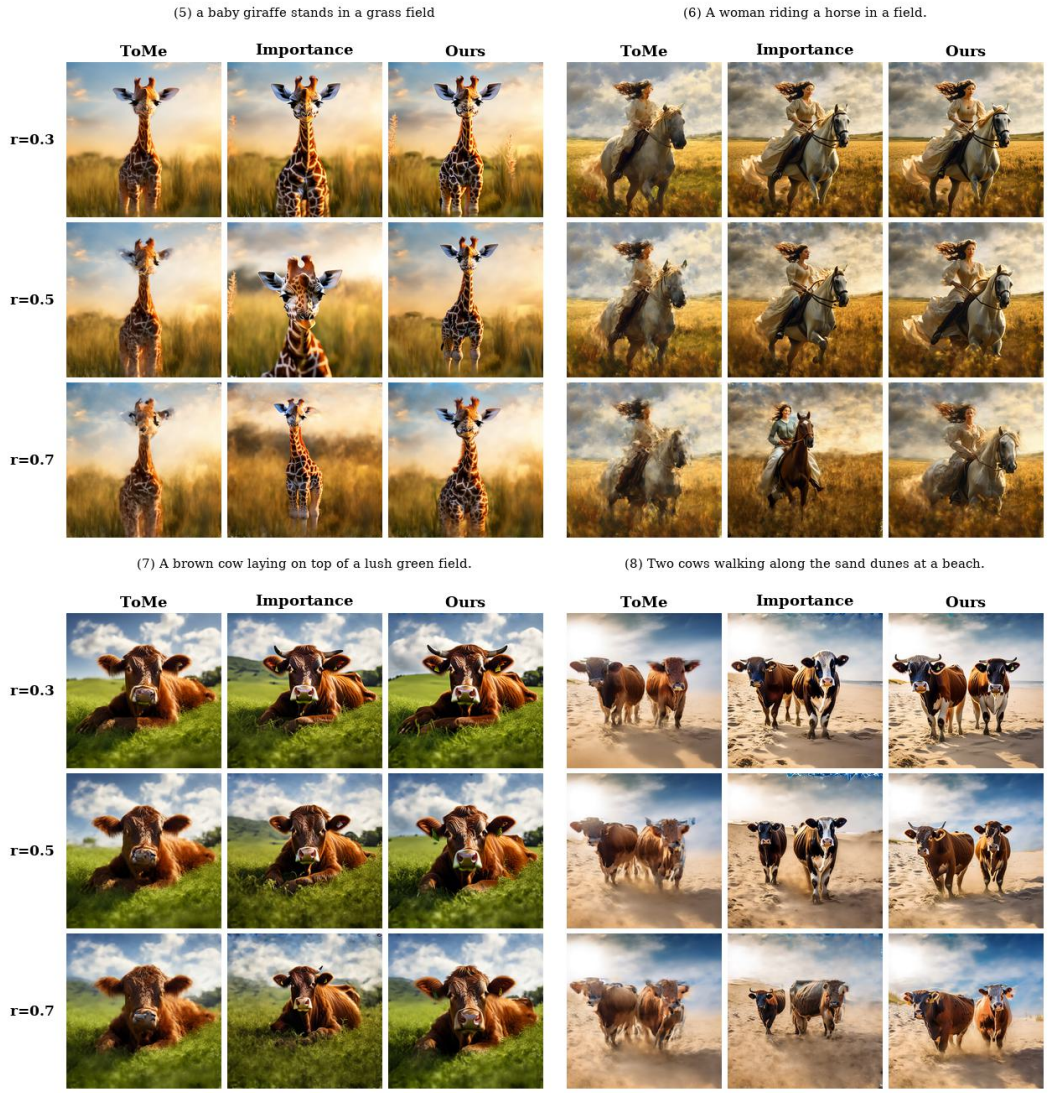


Figure 6: **Additional PixArt prompts at the same three ratios.** QUADMERGE retains giraffe markings, the rider on the horse, and both cows on the beach through $r = .70$.

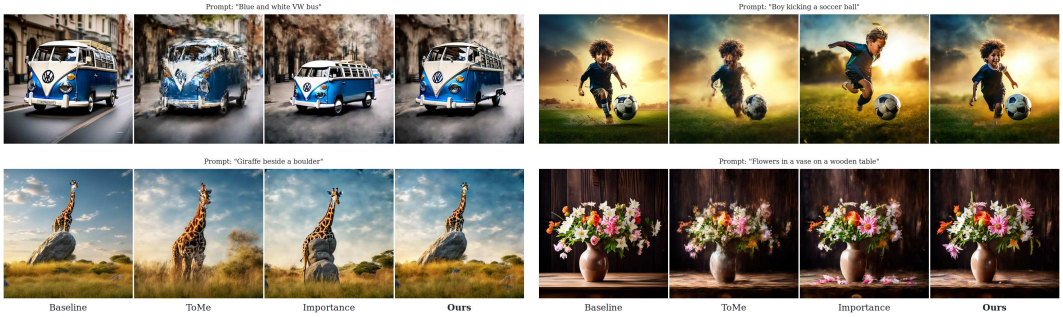


Figure 7: **Extra matched** 2×2 **PixArt comparisons at** $r = .50$**.** Columns are Baseline, ToMe, Importance, and QUADMERGE. Ours stays closest to the uncompressed sample in geometry and texture.

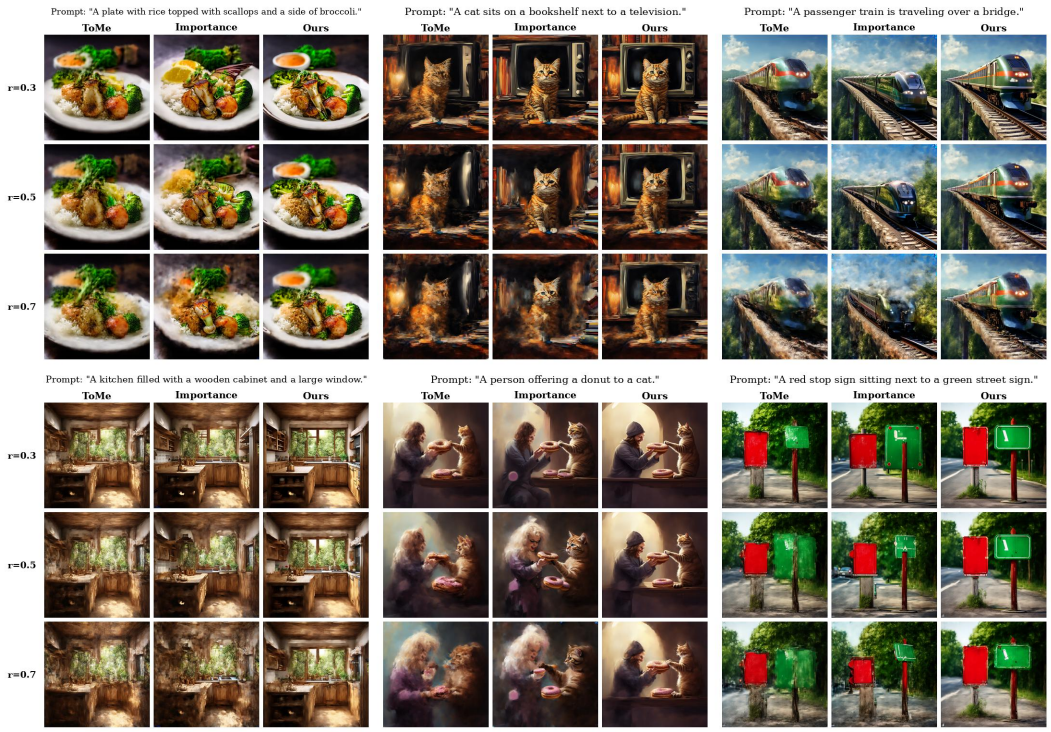


Figure 8: **Further PixArt scenes at $r \in \{.30, .50, .70\}$.** Food, indoor, animal, and signage prompts follow the same pattern: QUADMERGE keeps the train on the bridge, the kitchen layout, and the two signs in place at a matched token budget.



Figure 9: **Stable Diffusion 2.1 transfer.** Each row is one prompt at the same seed. QUADMERGE stays closest to the uncompressed layout—the herd in water, the elephants, and the selfie—matching its lead on every reported metric from $r = .50$ in Table 4.

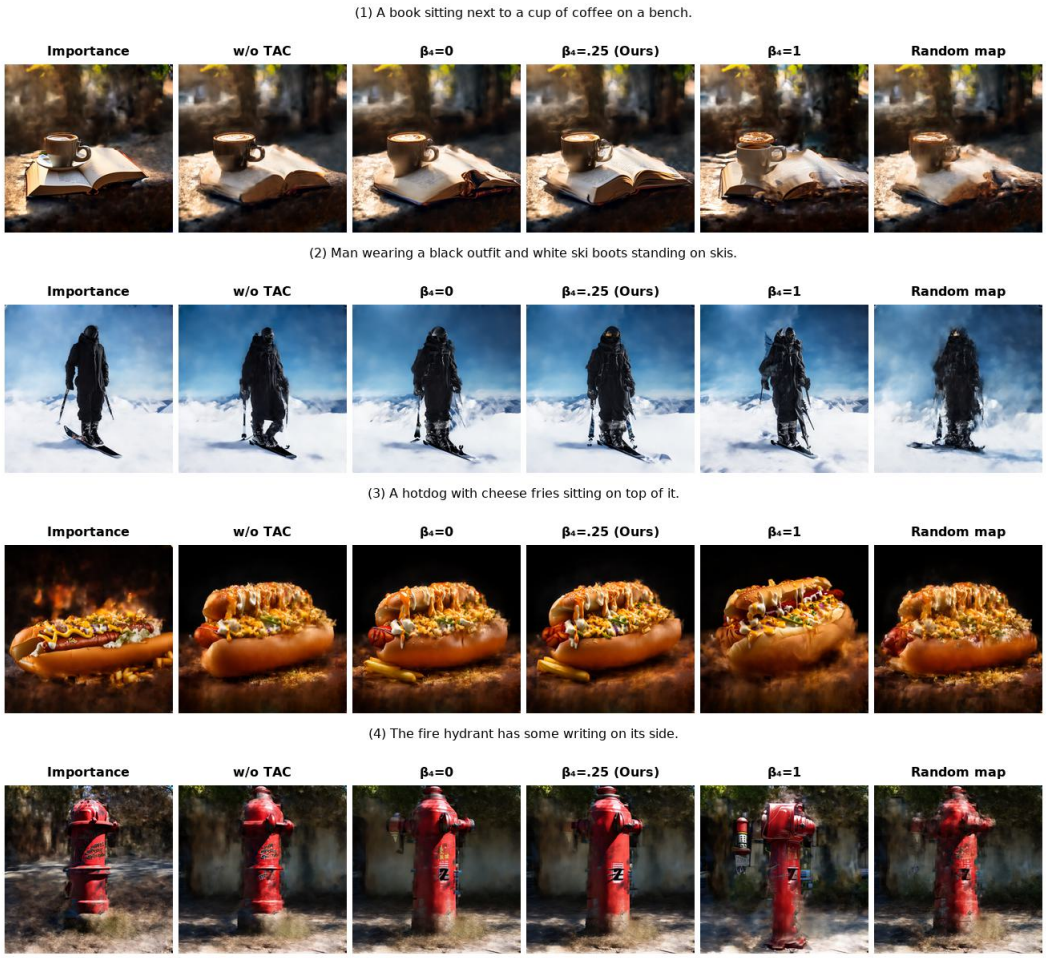


Figure 10: **Allocation variants at $r = .7$.** The displayed $\beta_4 = .25$ setting (Ours) keeps the book and cup, the skier, the hot dog, and the hydrant in place, with fine tokens on the subject.

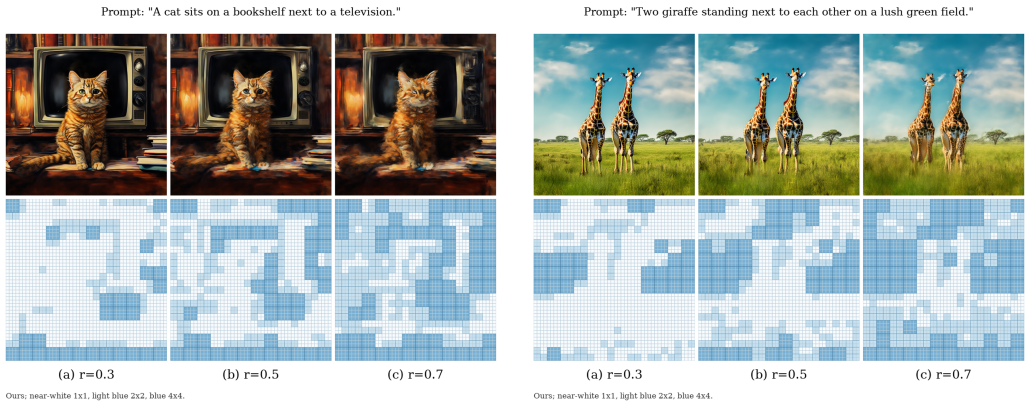


Figure 11: **Resolution plans under increasing compression.** As r grows, QUADMERGE coarsens the background while keeping 1×1 tokens on the cat, television, and giraffes, so the subject remains sharp at every displayed ratio.

B ADDITIONAL QUANTITATIVE RESULTS

On the complete PixArt sweep (Table 3), QUADMERGE attains the best FID and CLIP among merge methods at every listed ratio, including the extra settings $r = .60$ and $.75$. On Stable Diffusion (Table 4), it leads every reported metric from $r = .50$ through $.75$ and is the fastest method at

every ratio, reducing latency from $1.582\,\mathrm{s}$ to 0.727–$0.738\,\mathrm{s}$. Table 5 shows that max-risk scoring and importance-weighted aggregation are stable local choices; the 5k gains come from CFG-guided multi-resolution allocation and its coupling with TAC.

Table 3: **Complete PixArt-α results.** 5k captions, 64×64 grid, 20 steps, equal-quota $\beta_4 \approx .5$. The block $r \in \{.30, .50, .70\}$ repeats Table 1; $r = .60$ and .75 are the additional ratios. QUADMERGE is best on FID and CLIP in every merge group.

r	Method	Quality		Consistency		
		FID↓	CLIP↑	PSNR↑	SSIM↑	LPIPS↓
–	Baseline	36.60	31.06	–	–	–
2×2	ToDo	65.83	30.01	17.71	.6700	.3914
.30	ToMe	49.08	30.60	18.68	.7153	.3184
	Importance	37.65	30.84	17.73	.7084	.3211
	Ours	36.69	30.91	17.45	.7034	.3217
.50	ToMe	76.99	29.71	17.48	.6691	.3955
	Importance	42.49	30.52	15.97	.6372	.4285
	Ours	38.93	30.69	16.01	.6433	.4095
.60	ToMe	91.02	29.14	17.07	.6530	.4243
	Importance	47.87	30.18	15.38	.6110	.4747
	Ours	42.28	30.50	15.75	.6285	.4368
.70	ToMe	102.37	28.54	16.68	.6382	.4522
	Importance	56.91	29.65	14.88	.5868	.5207
	Ours	48.99	30.20	15.70	.6197	.4550
.75	ToMe	107.64	28.28	16.46	.6297	.4706
	Importance	66.96	29.14	14.64	.5725	.5503
	Ours	53.40	30.04	15.66	.6163	.4631

Table 4: **Complete Stable Diffusion 2.1 results.** 5k captions, 512×512, 20-step DDIM. From $r = .50$, QUADMERGE is best on every column; latency is lowest at all ratios.

r	Method	Quality		Consistency			Efficiency
		FID↓	CLIP↑	PSNR↑	SSIM↑	LPIPS↓	Lat.↓
–	Baseline	28.11	30.43	–	–	–	1.582
.30	ToMe	30.36	**30.09**	14.52	.4935	.4619	.761
	Importance	**30.21**	30.08	14.54	.4954	**.4608**	.849
	Ours	30.31	30.08	**14.68**	**.5020**	.4652	**.727**
.50	ToMe	31.35	30.04	14.40	.4808	.4974	.762
	Importance	30.78	30.09	14.39	.4811	.5016	.848
	Ours	**30.35**	**30.14**	**14.83**	**.4980**	**.4847**	**.730**
.60	ToMe	32.40	30.06	14.37	.4752	.5183	.746
	Importance	31.08	30.08	14.31	.4731	.5250	.838
	Ours	**31.04**	**30.16**	**14.91**	**.4962**	**.4922**	**.733**
.70	ToMe	34.24	30.04	14.37	.4705	.5412	.757
	Importance	32.29	30.01	14.26	.4666	.5484	.840
	Ours	**31.74**	**30.16**	**14.92**	**.4949**	**.4957**	**.738**
.75	ToMe	36.83	29.92	14.37	.4670	.5561	.748
	Importance	33.38	29.97	14.24	.4629	.5639	.815
	Ours	**32.19**	**30.16**	**14.92**	**.4943**	**.4976**	**.728**

Table 5: **Local-design controls on 1k matched captions.** PixArt-α at $r = .50$. The three variants remain close, so the displayed max-risk, importance-weighted design is a stable default.

Variant	FID↓	CLIP↑
Full configuration (max-risk, importance-weighted)	66.85	30.574
Uniform aggregation	66.49	30.629
Average-risk scoring	65.44	30.646

Protocol. PixArt-α (`PixArt-XL-2-1024-MS`) generates 5,000 COCO val2014 captions with 20 steps on the 64×64 grid; images are resized to 768×768 for evaluation. Stable Diffusion 2.1 uses the same captions at 512×512 with 20-step DDIM. Caption order, seeds, and sampler settings are shared. FID and CLIP use the caption-matched COCO reference; PSNR, SSIM, and LPIPS compare to the uncompressed sample. Selection is on FID and CLIP.

Planner and TAC. PixArt uses $\mathcal{L} = \{1, 2, 4\}$, max-risk scoring, importance-weighted aggregation, $\beta_4 \approx .5$ in Table 1, a per-step grid offset, and plan reuse over blocks 7–20. TAC starts at step 4 with $\alpha = 1$ and $\delta = .3$. Stable Diffusion uses $\mathcal{L} = \{1, 2\}$, starts at step 10, and compresses self-attention in the first and last UNet blocks. ToDo is a clean-room 2×2 K/V downsample with

dense queries. Latency in Table 1(b) and Table 4 is measured on one A100-80GB (one warm-up, five timed images). Realized $\bar{K}$ and TFLOPs average recorded sequence lengths over eight prompts.