REFADAPT-DIT: ADAPTIVE JOINT ATTENTION FOR REFERENCE-CONDITIONED DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion Transformers (DiTs) have become the standard backbone for high-quality generative modeling, yet deploying them in conditional generation tasks remains computationally prohibitive because bidirectional joint attention repeatedly processes large reference streams. While existing optimization schemes mitigate generic temporal redundancy, they typically rely on coarse-grained static reuse and overlook the distinct dynamics of references and targets. Specifically, we observe that reference representations often evolve slowly along the generation trajectory, while the target often assigns little attention mass to them; reference drift and this target-to-reference exposure jointly shape how strongly stale reference states affect the target. To exploit these patterns, we introduce REFADAPT-DIT, a training-free framework for adaptive control of joint attention between references and targets. Instead of rigid static strategies, REFADAPT-DIT combines consecutive target-Q change with previously observed target-to-reference attention mass to control reference computation adaptively at block granularity. Under ultra-few-step settings, REFADAPT-DIT enables speedups of up to $2.097\times$ on 4-step MiniMax H3 and $3.54\times$ on 8-step Qwen Image Edit, while maintaining comparable visual quality.

1 INTRODUCTION

Diffusion Transformers (DiTs) are now a common backbone for high-fidelity image and video generation (Peebles & Xie, 2023; Wu et al., 2025). Reference-conditioned systems increasingly use joint attention, placing visual references, instructions, and noisy target tokens in a shared Transformer sequence (Xiao et al., 2025; Chen et al., 2025; Ye et al., 2026; Ju et al., 2026; Bian et al., 2026). As these systems expand from single-reference editing to multi-reference image generation and temporally extended video editing (Oshima et al., 2026; Ye et al., 2026; Ju et al., 2026), the reference stream can occupy a large fraction of the sequence. Its cost is paid repeatedly during denoising: although the external reference is fixed, its hidden states and attention K/V continue to evolve through bidirectional interaction with the target.

Training-free acceleration exploits redundancy either across denoising steps, through caching or forecasting (Liu et al., 2025a; Chung et al., 2026; Liu et al., 2025b; Son et al., 2026; Zheng et al., 2026), or within each step, through token and attention reduction (Bolya & Hoffman, 2023; Sun et al., 2025; Wu et al., 2026). Reference-side methods further compress or reuse conditioning features, alter the conditioning dependency, or exploit localized editing redundancy (Tan et al., 2026; Zhang et al., 2025b; Chen et al., 2026b). These methods show that substantial computation can be removed, but they leave a distinct question when the original bidirectional joint attention is preserved: a cached reference can be redundant or predictable yet still be important to the current target. Reference reuse therefore requires estimating not only how stale the reference is, but also the consequence of that staleness for the target.

We first use full-length, full-fresh trajectories to expose a real but nonuniform reference-reuse opportunity; their dense adjacent transitions provide temporal resolution for mechanism analysis. Across H3 video generation and Qwen multi-reference image editing, reference K/V evolves more slowly than the target through broad shallow-to-middle depth regimes, while this asymmetry narrows and

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.

can reverse in deeper blocks. This role-dependent temporal redundancy motivates reuse, but does not by itself make reuse safe. The target reads reference tokens with strongly varying directional exposure across depth and denoising steps, and controlled stale-K/V counterfactuals show that the resulting target perturbation depends on this interaction. Conditioning reference drift on reference-read exposure raises its Spearman association with local target-output impact from 0.564 to 0.826 in H3 and from 0.708 to 0.905 in Qwen. The key distinction is therefore between *reference motion* and the *target consequence* of that motion.

This distinction creates the online design problem. The counterfactual analysis can measure fresh reference drift and stale-reference impact offline, but neither quantity is available before a reuse decision without performing the computation we aim to avoid. REFADAPT-DIT therefore anticipates target consequence from two signals already available from the target path and its recent history: current target-Q change and previously observed directional exposure. The resulting block-local score is evaluated from shallow to deep. Once it exceeds a threshold, REFADAPT-DIT reactivates the cached reference hidden state at that depth and recomputes the remaining reference suffix; shallower blocks continue to use cached reference K/V. Because the policy adapts reference depth rather than token count or generic denoiser reuse, it can be composed with those acceleration mechanisms; we evaluate such combinations directly rather than assuming additive gains.

We evaluate REFADAPT-DIT on 100 MultiBanana cases with Qwen Image Edit and 100 EditVerseBench cases with MiniMax H3, using benchmark-native quality metrics and measured generation efficiency. Our contributions are:

- We identify a depth-dependent reference-reuse opportunity in joint-attention diffusion: reference K/V changes substantially more slowly than the target in broad shallow-to-middle regimes, but not uniformly across depth.
- We show with controlled stale-reference counterfactuals that reference staleness is interaction-conditioned: directional target exposure makes reference drift substantially more predictive of target-side attention error.
- We introduce REFADAPT-DIT, an impact-aware depth-adaptive reuse policy that combines current target-Q dynamics with lagged directional exposure, then reactivates reference computation at a calibrated depth cutoff without retraining.

2 RELATED WORK

Reference-conditioned joint-attention generation. Unified diffusion models formulate editing and reference-guided generation as in-context generation, processing source/reference content and noisy target tokens in a shared Transformer computation (Xiao et al., 2025; Chen et al., 2025; Xia et al., 2025; Oshima et al., 2026; Ye et al., 2026; Ju et al., 2026; Bian et al., 2026). This design supports multi-reference images and video context, but repeatedly updates the reference stream during denoising. We study efficiency within this existing bidirectional joint-attention formulation rather than introducing a new conditioning architecture.

Training-free diffusion acceleration. Temporal reuse methods cache intermediate activations or adapt reuse using trajectory evolution, forecasting, module relations, spectral structure, or block/subspace dynamics (Ma et al., 2024; Wimbauer et al., 2024; Zhao et al., 2025; Lyu et al., 2025; Liu et al., 2025a; Kahatapitiya et al., 2025; Chu et al., 2025; Bu et al., 2026; Liu et al., 2025b; Son et al., 2026; Zheng et al., 2026; Chen et al., 2026a; Gu et al., 2026; Cui et al., 2026; Chung et al., 2026). A complementary family reduces within-step work through token merging or attention reduction (Bolya & Hoffman, 2023; Sun et al., 2025; Liu et al., 2026; Wu et al., 2026; Zhang et al., 2025a). These methods primarily estimate generic computational redundancy—whether a state, block, or token can be reused, predicted, or removed. REFADAPT-DIT instead asks whether stale *reference* information is exposed to and consequential for the current target. Since this criterion controls reference refresh rather than whole-denoiser reuse or token count, the mechanisms can be combined; we report composed variants in the main comparison.

Efficient reference-side computation. Closest to our setting, OminiControl2 compresses conditional tokens and reuses conditional features, while EasyControl changes the conditioning dependency so conditional K/V can be reused (Tan et al., 2026; Zhang et al., 2025b). RegionE focuses

on instruction-based single-image editing and exploits spatially localized edited/unedited trajectories together with region-instruction caching (Chen et al., 2026b). In contrast, REFADAPT-DIT preserves the original bidirectional joint-attention computation and adapts *when* reference states are refreshed according to their predicted target impact. This refresh criterion applies to both multi-reference image editing and reference-conditioned video generation without making the reference branch independent of the target.

3 FROM REFERENCE DYNAMICS TO TARGET-SIDE STALENESS IMPACT

Setting. We ask three linked questions: where reference states change slowly, how target queries read those reusable rows, and whether exposure-conditioned staleness better predicts the resulting local target-attention perturbation. We group each model's native token partition into two operational roles: R contains reusable cached reference-conditioning rows, and Y the remaining active-generation rows. With $\tau_i \in \{\tau_{\text{ref}}, \tau_{\text{target}}\}$, the row sets are

$$\mathrm{R} = \{i \mid \tau_i = \tau_{\text{ref}}\}, \qquad \mathrm{Y} = \{i \mid \tau_i = \tau_{\text{target}}\}.$$

The superscript $\mathrm{Y} \to \mathrm{R}$ means target queries look up reference keys (exposure, not causal flow). Here t indexes denoising steps and $b \in \{0, \ldots, B-1\}$ Transformer blocks. Sections 3.1 and 3.2 use full-length, full-fresh trajectories for temporal resolution; Section 3.3 uses an earlier anchor for a read-only counterfactual. These are mechanism diagnostics; distilled few-step deployment is evaluated in Section 5. Except for the three-prompt profiles in Figure 1, all are within-run diagnostics.

3.1 REFERENCE–TARGET DYNAMICS: POTENTIAL REUSE OPPORTUNITY

For any fixed-scope tensor slice X, define its consecutive relative *drift* as

$$D_{t,b}(X) = \frac{\|X_{t,b} - X_{t-1,b}\|_F}{\|X_{t-1,b}\|_F + \varepsilon}. \tag{1}$$

where token scope and, when applicable, attention head are fixed; $X = Z^{\tau}$, $Z \in \{K, V\}$, and $\tau \in \{\tau_{\text{ref}}, \tau_{\text{target}}\}$. Thus D is adjacent-step motion, not cache error or anchor distance. Figure 1 visualizes the matched log-ratio

$$\rho^{Z}_{t,b} = \log_{10} \frac{D_{t,b}(Z^{\tau_{\text{ref}}}) + \varepsilon}{D_{t,b}(Z^{\tau_{\text{target}}}) + \varepsilon}, \tag{2}$$

so $\rho < 0$ means slower reference motion. In Figure 1, the shallow-to-middle heatmaps are predominantly negative and the absolute Ref profiles are small and flat. The median Ref/Tgt K/V ratio is 0.352 in H3 blocks 6–25 and $0.433/0.381$ for Qwen K/V across depth. Because the comparison is role matched, slower Ref motion indicates potential reuse, not cache safety; the gap narrows or reverses deeper, making one depth-independent rule too coarse.

An H3 timing probe further finds median per-block Spearman 0.675 between target block-input change and Ref K/V change, with 22/50 blocks reaching 0.8. Because the tensors differ, this motivates a target-side runtime cue but neither reconstructs Ref drift nor certifies reuse.

Each heatmap cell is the mean of $\rho^{Z}_{t,b}$ over N prompts. The right profiles retain absolute scale by averaging reference drift over transitions within each prompt and then over prompts:

$$\bar{D}^{Z,\text{ref}}_{b} = \frac{1}{N} \sum_{n=1}^{N} \left[\operatorname*{mean}_{t} D^{(n)}_{t,b}(Z^{\tau_{\text{ref}}}) \right]. \tag{3}$$

where n indexes prompts. The red curve is the mean and the pink envelope is the prompt range; prompts are listed in Appendix D.1.

3.2 DIRECTIONAL EXPOSURE: TARGET QUERIES ATTENDING TO REFERENCE KEYS

Drift alone does not show how strongly the target reads reference rows. For head h, let $\mathcal{Q}_{\mathrm{Y}}$ be sampled target-query rows and $\mathcal{K}_{\mathrm{R}}$ reusable reference-key rows. For post-softmax attention $A = \mathrm{Softmax}(QK^{\top}/\sqrt{d_h})$, define their directional mass

$$M^{\mathrm{Y}\to\mathrm{R}}_{t,b,h}(A) = \frac{1}{|\mathcal{Q}_{\mathrm{Y}}|} \sum_{q \in \mathcal{Q}_{\mathrm{Y}}} \sum_{k \in \mathcal{K}_{\mathrm{R}}} A^{t,b,h}_{qk}. \tag{4}$$

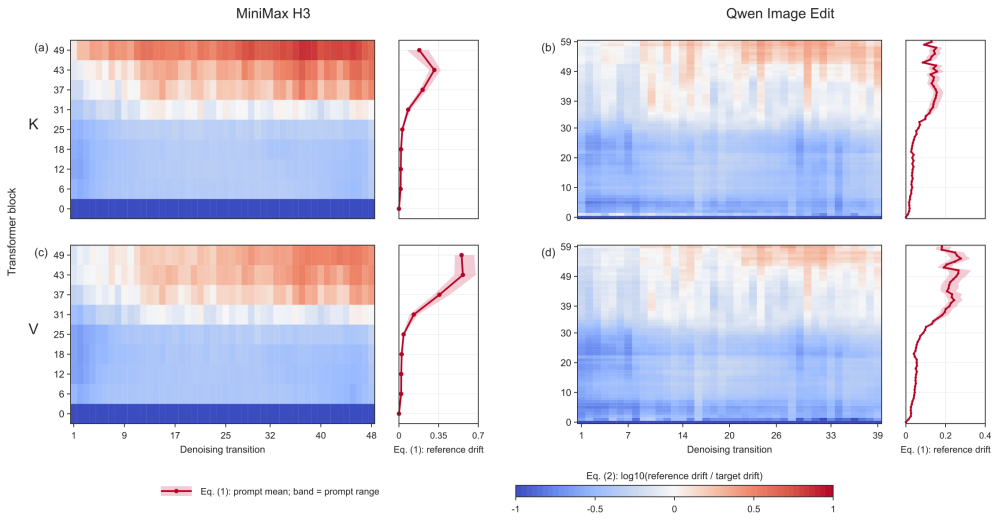


Figure 1: Matched reference–target K/V dynamics. Heatmaps show transition (horizontal), block (vertical), and prompt-averaged ρ (color; negative means slower Ref motion). Right profiles show mean Ref drift horizontally and block vertically; pink spans prompts. H3 uses the nine shared blocks.

Table 1: Trajectory-wide directional attention allocation in full-fresh joint attention.

Model (steps × blocks)	Ref mass	Tgt mass	Tgt/Ref
MiniMax H3 (49 × 50)	9.51%	90.49%	9.52×
Qwen Image Edit (40 × 60)	14.66%	85.34%	5.82×

Thus M measures how much target attention is routed through reference values: it is exposure, not by itself a causal effect or reuse criterion. A^{fresh} uses current Q/K over all keys; the runtime controller applies the same definition to executed attention.

For the summaries in Table 1 and Figure 2, we average the per-head quantity,

$$\bar{M}^{\mathrm{Y}\to\mathrm{R}}_{t,b} = \frac{1}{H}\sum_{h=1}^{H} M^{\mathrm{Y}\to\mathrm{R}}_{t,b,h}(A^{\text{fresh}}). \tag{5}$$

Trajectory means in Table 1 are 9.51% Ref mass for H3 and 14.66% for Qwen; these describe exposure, not reuse thresholds.

To test adjacent-step similarity directly, we average the directional Ref-mass change between neighboring steps at each fixed block:

$$\bar{\Delta}^{\mathrm{Y}\to\mathrm{R}}_{b} := \frac{100}{T-1}\sum_{t=2}^{T}\left|\bar{M}^{\mathrm{Y}\to\mathrm{R}}_{t,b} - \bar{M}^{\mathrm{Y}\to\mathrm{R}}_{t-1,b}\right| \quad \text{(percentage points)}. \tag{6}$$

Figure 2 shows that $\bar{\Delta}_b$ is at most 0.70 points in H3 and 1.44 in Qwen: blockwise Ref mass is stable on average, not guaranteed for every head or transition.

3.3 Counterfactual Local Perturbation from Stale Reference K/V

Let $a < t$ be an earlier synthetic anchor in a fixed three-step schedule. For reference tensor $Z \in \{K, V\}$, define its anchor-to-current endpoint drift

$$D_{t|a,b}(Z^{\mathrm{R}}) = \frac{\left\| Z^{\mathrm{R}}_{t,b} - Z^{\mathrm{R}}_{a,b} \right\|_F}{\left\| Z^{\mathrm{R}}_{a,b} \right\|_F + \varepsilon}. \tag{7}$$

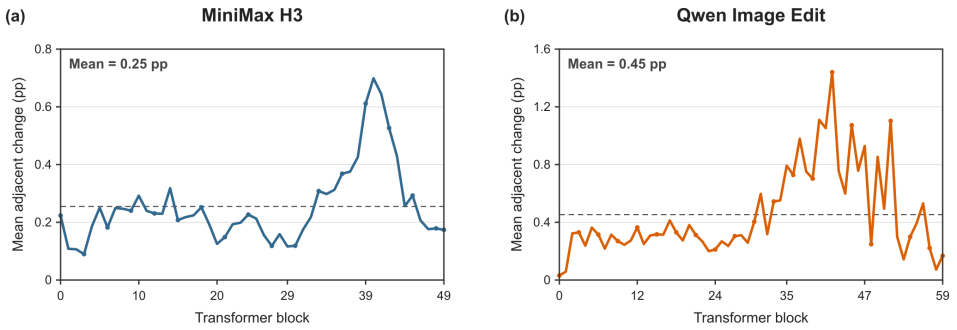


Figure 2: Adjacent-step directional Ref-mass stability. Horizontal: block; vertical: mean change $\bar{\Delta}_b$ in percentage points (shared scale). Mass is exposure, not causal effect or a standalone reuse rule.

Unlike Equation (1), this is anchor-to-current, not adjacent-step, change. Applying this definition separately to head-resolved K and V slices yields their joint endpoint scalar $E^{KV}_{t|a,b,h}$; this diagnostic-only quantity is not used by the deployed controller.

Next, let $\mathrm{Attn}(Q,K,V) = \mathrm{Softmax}(QK^\top/\sqrt{d_h})V$. Holding current target queries and target-side K/V fixed, replace only reference K/V by their anchor values:

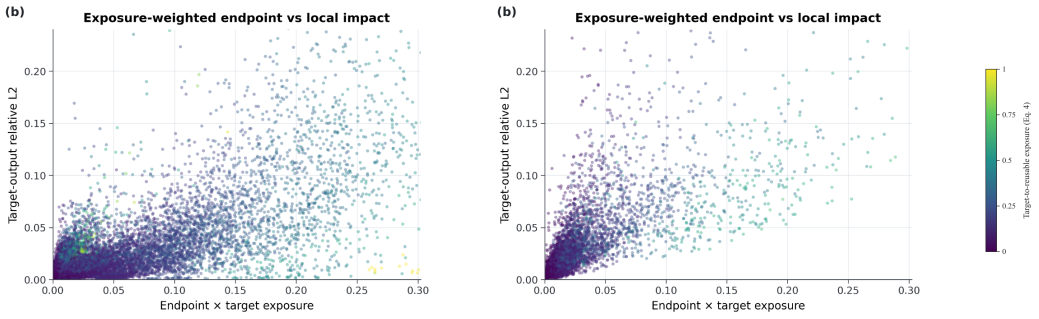
$$O^{\mathrm{fresh}}_{t,b,h} = \mathrm{Attn}\Big(Q^{t,b,h}_{\mathrm{Y}}, [K^{t,b,h}_{\mathrm{R}}; K^{t,b,h}_{\mathrm{Y}}], [V^{t,b,h}_{\mathrm{R}}; V^{t,b,h}_{\mathrm{Y}}]\Big), \tag{8}$$

$$O^{\mathrm{staleKV}}_{t|a,b,h} = \mathrm{Attn}\Big(Q^{t,b,h}_{\mathrm{Y}}, [K^{a,b,h}_{\mathrm{R}}; K^{t,b,h}_{\mathrm{Y}}], [V^{a,b,h}_{\mathrm{R}}; V^{t,b,h}_{\mathrm{Y}}]\Big). \tag{9}$$

The resulting local counterfactual perturbation is

$$I^{KV}_{t|a,b,h} = \frac{\left\| O^{\mathrm{staleKV}}_{t|a,b,h} - O^{\mathrm{fresh}}_{t,b,h} \right\|_F}{\left\| O^{\mathrm{fresh}}_{t,b,h} \right\|_F + \varepsilon}. \tag{10}$$

In Figure 3, each (t,b,h) point has horizontal coordinate $E^{KV}M(A^{\mathrm{fresh}})$, vertical coordinate I^{KV}, and color equal to the same exposure M. The Frobenius ratio is relative L2 on the target attention output; the stale branch is read-only, so I^{KV} is neither a perceptual metric nor a final-error bound.



(a) MiniMax H3 (b) Qwen Image Edit

Figure 3: Exposure-conditioned endpoint score versus local perturbation. Each point is one (t,b,h) comparison: horizontal $E^{KV}M^{\mathrm{Y}\to\mathrm{R}}$, vertical I^{KV}, and color current per-head exposure M to the operational Ref set defined above. This diagnostic does not define the deployed controller score.

Raw E^{KV} has Spearman $0.564/0.708$ with I^{KV} on H3/Qwen; exposure weighting raises this to $0.826/0.905$ in every measured block. Thus equal Ref drift can have different local effects depending on target exposure. The product is a diagnostic, not an error model or calibrated probability.

4 REFADAPT-DIT: DEPTH-ADAPTIVE REFERENCE REUSE

The probes above suggest that reuse should be depth aware and conditioned on target–reference interaction. Because fresh Ref drift is unavailable during reuse, REFADAPT-DIT combines observable target-Q change with recent directional exposure to choose a depth cutoff; a bounded-age refresh limits staleness. Figure 4 separates the cached state, online cues, and action, because each block decision must be made before current Ref rows exist.

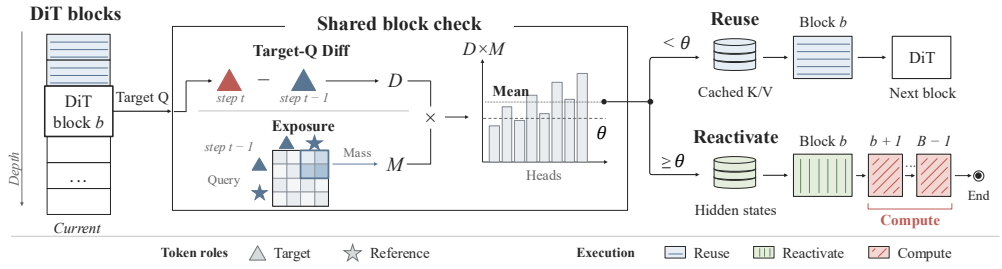


Figure 4: REFADAPT-DIT at one step. Per-head target-Q change D and lagged reference exposure M form $D \times M$; the first block whose head mean reaches θ sets the depth cutoff. Earlier blocks REUSE, the boundary REACTIVATEs, and deeper blocks COMPUTE.

Execution and cache state. A *reuse* block skips reference rows while the target attends to cached reference K/V; *reactivation* restores a cached boundary state before deeper blocks compute and update the reference stream. A *full refresh* starts at block 0. Each block b stores

$$\mathcal{Z}_b = \left(\bar{H}_b^{\mathrm{R}}, \bar{K}_b^{\mathrm{R}}, \bar{V}_b^{\mathrm{R}}, a_b\right), \tag{11}$$

where $\bar{H}_b^{\mathrm{R}}$ is the cached block-input state, $(\bar{K}_b^{\mathrm{R}}, \bar{V}_b^{\mathrm{R}})$ its K/V, and a_b its last materialization step. During reuse, cached K/V serve target attention without propagating a new Ref output. Restoring $\bar{H}_b^{\mathrm{R}}$ at a boundary gives deeper blocks a coherent Ref stream; computing only a suffix therefore yields depth-dependent cache ages without disconnected per-block refreshes.

Runtime cues observable under reuse. Fresh-reference diagnostics are unavailable under reuse, but target Q remains active. The timing probe in Section 3.1 motivates this target-side cue without treating it as reconstructed Ref drift. We apply Equation (1) to each post-RoPE target-query head, obtaining $D_{t,b}(Q^{\mathrm{Y},h})$. Directional interaction uses Equation (4) on executed attention A^{run}. Because block b has not run when deciding step t, the available cue is $M_{t-1,b,h}^{\mathrm{Y}\to\mathrm{R}}(A^{\mathrm{run}})$. It is recorded without another attention evaluation; a full step 0 initializes all state. Thus target-Q change supplies current dynamics. The adjacent-step stability in Section 3.2 supports treating the last executed mass as a recent interaction cue when current Ref rows are unavailable, not as an estimate of current fresh attention.

Interaction-conditioned score. Target-Q drift measures current change, while lagged mass measures recent Ref interaction. Their zero-parameter product transfers the diagnostic's change–interaction factorization into the per-head reactivation score

$$g_{t,b,h} = D_{t,b}(Q^{\mathrm{Y},h}) \cdot M_{t-1,b,h}^{\mathrm{Y}\to\mathrm{R}}(A^{\mathrm{run}}). \tag{12}$$

It is small if either factor is weak and controls reactivation; it neither reconstructs unavailable fresh Ref drift/current exposure nor estimates counterfactual error. The product is formed per head before aggregation, preserving whether change and exposure coincide within the same head.

Depth cutoff and bounded-age refresh. REFADAPT-DIT scans shallow to deep. The first block whose head-mean score reaches θ is the reactivation boundary b_t:

$$\frac{1}{H}\sum_{h=1}^{H} g_{t,b_t,h} \geq \theta, \qquad b_t \text{ is the first block satisfying this condition.} \tag{13}$$

If none crosses, the entire step reuses. Before b_t, blocks use cached K/V; at b_t, the target still uses cached K/V while $\bar{H}^{\mathrm{R}}_{b_t}$ is restored, so new Ref K/V begin at $b_t + 1$ and the deeper suffix recomputes. After $\Delta_{\max}$ steps without a full refresh, the controller recomputes from block 0, updates all caches, and resets their ages. The first-crossing rule keeps reference computation as one contiguous suffix with a single restored boundary, avoiding incompatible independent block decisions.

5 Experiments

5.1 Experimental Setup

We evaluate the same training-free controller design on fixed 100-case Qwen Image Edit (Wu et al., 2025)/MultiBanana (Oshima et al., 2026) image and MiniMax H3/EditVerseBench (Ju et al., 2026) video sets, with model, sampler, precision, resolution, batch size, and hardware fixed within each setting. Baselines are TeaCache, SeaCache, TaylorSeer, ToMe, and H3-only AsymRnR.

We use native metrics. MultiBanana reports IA, RC, BSM, PR, VQ, and official score $3\,\mathrm{IA} + 3\,\mathrm{RC} + \mathrm{BSM} + \mathrm{PR} + \mathrm{VQ}$; EditVerseBench reports EditQ, Pick Score, frame/video text alignment, and CLIP/DINO temporal consistency. The main experiments use distilled few-step checkpoints on NVIDIA H20 GPUs with 96 GB memory per GPU: distilled 8-step Qwen Image Edit at 1024×1024 on one GPU, and the LightX2V release of distilled 4-step MiniMax H3 at 1344×768 (768p) on the same GPU type. Reported speedup uses measured generation latency relative to full computation; all timings come from real executions rather than offline schedule simulation. Unlike the full-length diagnostics, these few-step settings provide a stricter error-recovery test, with fewer later steps to correct a reuse error; full protocol details are in Appendix B. All methods use the same cases and setting-specific configuration. Composed ResComp rows are reported separately so standalone reuse gains remain explicit.

5.2 Main Comparison

Qwen image editing. Table 2 reports all MultiBanana metrics; TaylorSeer uses interval 2, and the composed row adds orthogonal resolution compression to RefAdapt-DiT. Colored parentheses in both main tables show changes from full computation. The Type column uses C/F/T/R/H for caching, forecasting, token reduction, reference reuse, and hybrid methods, respectively.

Table 2: Qwen Image Edit on 100 MultiBanana cases (8 steps). MB Score is the official aggregate; time is in seconds. **ResComp** denotes orthogonal resolution compression.

Method	Type	Efficiency		MultiBanana Evaluation					
		Speedup ↑	Time ↓	IA ↑	RC ↑	BSM ↑	PR ↑	VQ ↑	MB Score ↑
Full computation	–	1.00×	60.39	1.286	1.735	1.408	1.878	2.163	14.512
TeaCache (Liu et al., 2025a)	C	1.13×	53.30	1.367 (+0.081)	1.571 (−0.164)	1.306 (−0.102)	1.816 (−0.062)	2.041 (−0.122)	13.977 (−0.535)
SeaCache (Chung et al., 2026)	C	1.60×	37.63	1.347 (+0.061)	**1.959** (+0.224)	1.551 (+0.143)	1.776 (−0.102)	2.122 (−0.041)	15.367 (+0.855)
TaylorSeer (Liu et al., 2025b) (I=2)	F	1.61×	37.58	**1.396** (+0.110)	1.875 (+0.140)	1.438 (+0.030)	**2.083** (+0.205)	**2.271** (+0.108)	**15.605** (+1.093)
ToMe (Bolya & Hoffman, 2023)	T	1.06×	56.74	1.200 (−0.086)	1.180 (−0.555)	1.120 (−0.288)	1.280 (−0.598)	1.180 (−0.983)	10.720 (−3.792)
RefAdapt-DiT	**R**	**2.58×**	**23.43**	1.320 (+0.034)	1.640 (−0.095)	**1.620** (+0.212)	1.720 (−0.158)	2.220 (+0.057)	14.440 (−0.072)
RefAdapt-DiT + ResComp	**H**	**3.54×**	**17.04**	1.180 (−0.106)	1.180 (−0.555)	1.160 (−0.248)	1.920 (+0.042)	2.160 (−0.003)	12.320 (−2.192)

At 8 steps, RefAdapt-DiT is the fastest standalone method in Table 2: $2.58\times$ speedup versus at most $1.61\times$ for the baselines, with an MB Score close to full computation (14.440 versus 14.512). Adding ResComp reaches $3.54\times$ but lowers the MB Score to 12.320, exposing a separate, more aggressive operating point rather than conflating resolution compression with the standalone controller gain. The component scores make that tradeoff explicit: standalone RefAdapt-DiT improves IA and VQ over full computation and has the best BSM (1.620; next best 1.551), while RC and PR decrease modestly. Thus the near-matched aggregate does not hide a uniform degradation across all five dimensions. TaylorSeer attains a higher aggregate MB Score, but at $1.61\times$ rather than $2.58\times$ speedup; RefAdapt-DiT therefore occupies a distinct higher-speed operating regime rather than dominating every quality dimension.

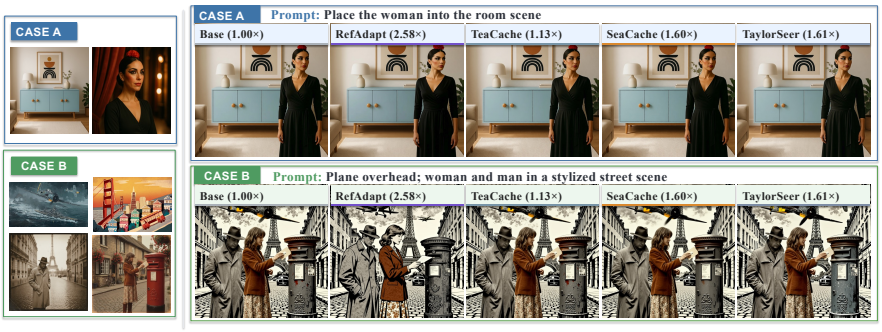


Figure 5: Qwen qualitative comparison for two multi-reference cases. Columns show full computation and four accelerated methods.

In Case A of Figure 5, the 2.58× REFADAPT-DIT output preserves subject identity and room composition. In Case B, it appears to retain the reference's line-art style more clearly than the lower-speed outputs shown, while keeping both subjects and the aircraft.

MiniMax H3 video generation.

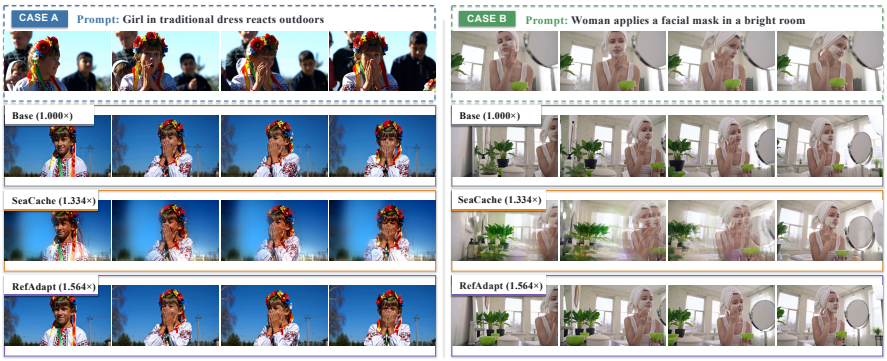


Figure 6: H3 qualitative comparison over sampled frames. Rows show full computation, SeaCache, and REFADAPT-DIT.

Figure 6 shows REFADAPT-DIT following Case A's motion without freezing the trajectory while retaining reference appearance. In this sampled Case B example, REFADAPT-DIT keeps cleaner contours and scene structure while the SeaCache output shows visible ghosting.

Table 3: MiniMax H3 on 100 EditVerseBench cases (4 steps). Quality follows the official protocol; time is in seconds. **ResComp** denotes orthogonal resolution compression.

Method	Type	Efficiency		VLM Eval.	Video Qual.	Text Align.		Temporal Cons.	
		Speedup ↑	Time ↓	EditQ ↑	Pick ↑	Frame ↑	Video ↑	CLIP ↑	DINO ↑
Full computation	–	1.000×	200.80	7.280	**20.528**	28.243	24.994	0.990	**0.990**
TeaCache (Liu et al., 2025a)	C	1.341×	149.69	7.340 (+0.060)	20.509 (−0.019)	28.294 (+0.051)	25.051 (+0.057)	0.990 (+0.000)	**0.990** (+0.000)
SeaCache (Chung et al., 2026)	C	1.334×	150.52	7.060 (−0.220)	20.389 (−0.139)	28.197 (−0.046)	24.800 (−0.194)	0.988 (−0.002)	0.986 (−0.004)
TaylorSeer (Liu et al., 2025b)	F	1.333×	150.67	7.361 (+0.081)	20.506 (−0.022)	28.305 (+0.062)	25.062 (+0.068)	0.990 (+0.000)	**0.990** (+0.000)
ToMe (Bolya & Hoffman, 2023)	T	1.207×	166.37	5.028 (−2.252)	19.851 (−0.677)	27.755 (−0.488)	24.614 (−0.380)	0.988 (−0.002)	0.985 (−0.005)
AsymRnR (Sun et al., 2025)	T	1.044×	192.29	6.407 (−0.873)	20.196 (−0.332)	28.341 (+0.098)	25.003 (+0.009)	0.980 (−0.010)	0.977 (−0.013)
REFADAPT-DIT	**R**	**1.564×**	**128.38**	7.413 (+0.133)	20.437 (−0.091)	28.053 (−0.190)	25.041 (+0.047)	**0.991** (+0.001)	**0.990** (+0.000)
REFADAPT-DIT + ResComp	**H**	**2.097×**	**95.75**	**7.565** (+0.285)	20.444 (−0.084)	**28.414** (+0.171)	**25.206** (+0.212)	**0.991** (+0.001)	**0.990** (+0.000)

In Table 3, on 4-step H3, REFADAPT-DIT is the fastest standalone method (1.564×) and has the highest EditQ among standalone accelerators (7.413 versus at most 7.361), while alignment and tem-

poral consistency remain near full computation. With ResComp it reaches $2.097\times$ while retaining the overall quality profile, demonstrating that the reuse controller composes with an orthogonal optimization even when only four denoising steps are available. Relative to TeaCache, the fastest cache baseline at $1.341\times$, standalone REFADAPT-DIT provides a 16.6% larger speedup. CLIP/DINO temporal consistency remains at 0.991/0.990; the small alignment shifts therefore do not indicate a broad quality collapse.

5.3 ABLATION STUDIES

Runtime-controller ablation. The 50-step Ref2VA trajectory is used only to expose enough decisions for signal selection, not to estimate few-step deployment gain, and is separate from the 4-step main benchmark. We compare consecutive target-Q change, directional exposure, their REFADAPT-DIT product, and anchor-relative target-Q change times exposure; Fixed R3 is a non-adaptive baseline. All adaptive variants share the depth-wise executor in Section 4, the sequential stopping rule, bounded-age fallback, threshold, and other settings. Only the trigger cue changes. Because this executor can reuse a shallow prefix and recompute a deeper suffix within one step, we report measured generation speedup rather than whole-step refresh counts.

Table 4: MiniMax H3 controller ablation on a separate 50-step Ref2VA trajectory. Adaptive variants share the executor and differ only in trigger cue; Fixed R3 is non-adaptive. Speedup uses the depth-wise executor; PSNR/SSIM are same-seed diagnostics from the signal-matched closed-loop replay and indicate trajectory fidelity, not absolute perceptual quality.

Configuration	Controller cue	Speedup ↑	PSNR ↑	SSIM ↑
Full computation	–	1.000×	–	1.0000
Fixed R3	fixed interval	1.464×	18.00	0.8823
Target-Q	consecutive target-Q	1.459×	29.81	0.9879
Exposure	directional exposure	1.459×	**32.24**	**0.9958**
REFADAPT-DIT	target-Q × exposure	**1.482**×	31.24	0.9937
Anchor-Q × Exposure	anchor-Q × exposure	1.376×	29.78	0.9864

Across the depth-wise speed ablation and signal-matched fidelity replay in Table 4, REFADAPT-DIT provides the best overall operating point: it attains the highest measured speedup, $1.482\times$, while maintaining essentially optimal trajectory fidelity (31.24 dB PSNR and 0.9937 SSIM), within only 1.00 dB and 0.0021 of the numerical maxima and markedly above Fixed R3 (18.00 dB/0.8823). Thus its acceleration does not require leaving the high-fidelity regime. The two cues are complementary: Target-Q tracks local target change, exposure tracks whether the target reads Ref, and their product requires both. The weaker anchor-relative result ($1.376\times$, 29.78 dB/0.9864) further favors local step-to-step change over displacement accumulated across a variable lag; additional diagnostics are in Appendix C.

6 CONCLUSION

We propose REFADAPT-DIT, an interaction-conditioned controller that combines consecutive target-Q change with lagged target-to-reference attention mass to select a reference reactivation depth, with bounded-age full refresh. On the fixed 100-case MultiBanana and EditVerseBench sets, it achieves $2.58\times$ Qwen Image Edit and $1.564\times$ MiniMax H3 generation speedups while keeping benchmark quality close to full computation. Because ResComp reduces a different part of the workload than reference-refresh control, the composed rows' $3.54\times$ and $2.097\times$ speedups provide empirical evidence of complementarity rather than assuming orthogonality. Thus interaction-aware reuse removes substantial redundant work even in few-step generation. Full-length counterfactual analysis identifies the mechanism, while ultra-few-step closed-loop experiments test realized generation outcomes under a reduced error-recovery budget. Our evaluation covers two representative joint-attention systems and benchmark-native image/video tasks; future work will extend REFADAPT-DIT to broader architectures and sampling regimes, automate controller calibration for new operating points, and jointly optimize reference reuse with complementary acceleration techniques.

Reproducibility Statement

The method and runtime controller are specified in Sections 3 and 4 and Appendix A, while the datasets, distilled checkpoints, hardware, metrics, timing definitions, and ablations are documented in Section 5 and Appendix B.

AI Use Statement

Generative AI tools were used only to improve manuscript readability, LaTeX organization, and language editing. They were not used for research tasks such as method or experiment design, implementation, data processing, or result interpretation. All AI-assisted edits were reviewed by the authors, who verified that they did not alter the technical content and take full responsibility for the final manuscript.

References

Yuxuan Bian, Xin Chen, Zenan Li, Tiancheng Zhi, Shen Sang, Linjie Luo, and Qiang Xu. Video-As-Prompt: Unified semantic control for video generation. In *International Conference on Learning Representations*, 2026.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4599–4603, 2023.

Jiazi Bu, Pengyang Ling, Yujie Zhou, Yibin Wang, Yuhang Zang, Dahua Lin, and Jiaqi Wang. DiCache: Let diffusion model determine its own cache. In *International Conference on Learning Representations*, 2026.

Guantao Chen, Shikang Zheng, Yuqi Lin, and Linfeng Zhang. Forecast the principal, stabilize the residual: Subspace-aware feature caching for diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23632–23641, 2026a.

Pengtao Chen, Xianfang Zeng, Maosen Zhao, Mingzhu Shen, Peng Ye, Bangyin Xiang, Zhibo Wang, Wei Cheng, Gang Yu, and Tao Chen. RegionE: Adaptive region-aware generation for efficient image editing. In *International Conference on Learning Representations*, 2026b.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, Hui Ding, Zhe Lin, and Hengshuang Zhao. UniReal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12501–12511, 2025.

Huanpeng Chu, Wei Wu, Guanyu Feng, and Yutao Zhang. OmniCache: A trajectory-oriented global perspective on training-free cache reuse for diffusion transformer models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16302–16312, 2025.

Jiwoo Chung, Sangeek Hyun, MinKyu Lee, Byeongju Han, Geonho Cha, Dongyoon Wee, Youngjun Hong, and Jae-Pil Heo. SeaCache: Spectral-evolution-aware cache for accelerating diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14283–14294, 2026.

Hanshuai Cui, Zhiqing Tang, Zhifei Xu, Zhi Yao, Wenyi Zeng, and Weijia Jia. BWCache: Accelerating video diffusion transformers through block-wise caching. In *International Conference on Learning Representations*, 2026.

Lihui Gu, Jingbin He, Lianghao Su, Kang He, Wenxiao Wang, and Yuliang Liu. ScalingCache: Extreme acceleration of DiTs through difference scaling and dynamic interval caching. In *International Conference on Learning Representations*, 2026.

Xuan Ju, Tianyu Wang, Yuqian Zhou, He Zhang, Qing Liu, Cherry Zhao, Zhifei Zhang, Yijun Li, Yuanhao Cai, Shaoteng Liu, Daniil Pakhomov, Zhe Lin, Soo Ye Kim, and Qiang Xu. EditVerse: Unifying image and video editing and generation with in-context learning. In *International Conference on Learning Representations*, 2026.

Kumara Kahatapitiya, Haozhe Liu, Sen He, Ding Liu, Menglin Jia, Chenyang Zhang, Michael S. Ryoo, and Tian Xie. Adaptive caching for faster video generation with diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15240–15252, 2025.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7353–7363, 2025a.

Haosong Liu, Yuge Cheng, Wenxuan Miao, Zihan Liu, Aiyue Chen, Jing Lin, Yiwu Yao, Chen Chen, Jingwen Leng, Minyi Guo, and Yu Feng. Astraea: A token-wise acceleration framework for video diffusion transformers. In *International Conference on Learning Representations*, 2026.

Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Junjie Chen, and Linfeng Zhang. From reusing to forecasting: Accelerating diffusion models with TaylorSeers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15853–15863, 2025b.

Zhengyao Lyu, Chenyang Si, Junhao Song, Zhenyu Yang, Yu Qiao, Ziwei Liu, and Kwan-Yee K. Wong. FasterCache: Training-free video diffusion model acceleration with high quality. In *International Conference on Learning Representations*, 2025.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. DeepCache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15762–15772, 2024.

Yuta Oshima, Daiki Miyake, Kohsei Matsutani, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. MultiBanana: A challenging benchmark for multi-reference text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 448–460, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Byunggwan Son, Jeimin Jeon, Jeongwoo Choi, and Bumsub Ham. Relational feature caching for accelerating diffusion transformers. In *International Conference on Learning Representations*, 2026.

Wenhao Sun, Rong-Cheng Tu, Jingyi Liao, Zhao Jin, and Dacheng Tao. AsymRnR: Video diffusion transformers acceleration with asymmetric reduction and restoration. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 57694–57711, 2025.

Zhenxiong Tan, Qiaochu Xue, Xingyi Yang, Songhua Liu, and Xinchao Wang. OminiControl2: Efficient conditioning for diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4256–4265, 2026.

Felix Wimbauer, Bichen Wu, Edgar Schoenfeld, Xiaoliang Dai, Ji Hou, Zijian He, Artsiom Sanakoyeu, Peizhao Zhang, Sam Tsai, Jonas Kohler, Christian Rupprecht, Daniel Cremers, Peter Vajda, and Jialiang Wang. 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.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report, 2025.

Jianzong Wu, Liang Hou, Haotian Yang, Ye Tian, Pengfei Wan, Di Zhang, and Yunhai Tong. VMoBA: Mixture-of-block attention for video diffusion models. In *International Conference on Learning Representations*, 2026.

Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. DreamOmni: Unified image generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28533–28543, 2025.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

Zixuan Ye, Xuanhua He, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Qifeng Chen, and Wenhan Luo. Unified in-context video editing. In *International Conference on Learning Representations*, 2026.

Yuechen Zhang, Jinbo Xing, Bin Xia, Shaoteng Liu, Bohao Peng, Xin Tao, Pengfei Wan, Eric Lo, and Jiaya Jia. Training-free efficient video generation via dynamic token carving. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. EasyControl: Adding efficient and flexible control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19513–19524, 2025b.

Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast. In *International Conference on Learning Representations*, 2025.

Shikang Zheng, Guantao Chen, Qinming Zhou, Yuqi Lin, Lixuan He, Chang Zou, Peiliang Cai, Jiacheng Liu, and Linfeng Zhang. Let features decide their own solvers: Hybrid feature caching for diffusion transformers. In *International Conference on Learning Representations*, 2026.

A Additional Method Details

A.1 Counterfactual and Exposure Definitions

We use the same two semantic roles and the same stale-reference counterfactual as in Sections 3.2 and 3.3: current target queries and current target-side K/V are held fixed, while only reference-side K/V are replaced by values from an earlier refresh anchor. Stale-K and stale-V diagnostics replace one component at a time. Target exposure is the post-softmax mass from target queries to reference keys defined in Equation (4); reference-query mass on target keys is reported only as a structural diagnostic and is not used as a proxy for target dependence on the reference. The deployed controller uses the one-step-lagged realized runtime mass in Equation (12) rather than the current full-fresh exposure.

A.2 Runtime-Signal Selection and Calibration

The mechanism analysis uses current full-fresh exposure, whereas the deployed controller uses a lagged mass that was already observed in an earlier forward. For reference, the probe also measures a same-step diagnostic in which only the reference keys are replaced by their anchor values,

$$M^{\mathrm{Y}\to\mathrm{R}}_{t,b,h} = \mathrm{Exposure}(Q^{t,b,h}_{\mathrm{Y}}, K^{t,b,h}_{\mathrm{R}}, K^{t,b,h}_{\mathrm{Y}}), \tag{14}$$

$$M^{\mathrm{cached},\mathrm{Y}\to\mathrm{R}}_{t|a,b,h} = \mathrm{Exposure}(Q^{t,b,h}_{\mathrm{Y}}, K^{a,b,h}_{\mathrm{R}}, K^{t,b,h}_{\mathrm{Y}}). \tag{15}$$

For the current H3 short-lag probe, these are nearly identical (Pearson 0.9989, Spearman 0.9991, mean absolute difference 0.00298). This comparison isolates fresh-vs-cached keys at the *same* step; it should not be confused with the deployed one-step lag $M^{\mathrm{Y}\to\mathrm{R}}_{t-1,b,h}$. We therefore treat the deployed lagged mass as a heuristic interaction cue rather than as an exact surrogate for current full-fresh exposure.

The current measured target-side proxy is consecutive post-RoPE target-Q change. For comparison, change to the most recent materialization anchor is

$$c^{\mathrm{anchor}}_{t|a_b,b,h} = \frac{\left\| Q^{t,b,h}_{\mathrm{Y}} - Q^{a_b,b,h}_{\mathrm{Y}} \right\|_F}{\left\| Q^{a_b,b,h}_{\mathrm{Y}} \right\|_F + \varepsilon}. \tag{16}$$

On the current H3 trajectory, consecutive change is the stronger candidate in the existing aligned next-step impact diagnostic. Unlike consecutive change, anchor-relative change accumulates displacement from the most recent block-specific materialization anchor, whose age can vary across both steps and blocks. Under a shared controller threshold, the two signals therefore need not produce the same refresh schedule or speedup. We use consecutive change in the main method and leave target K/V, hidden-state, and anchor-change variants to ablation rather than treating them as equivalent by definition.

The only aggregation required by the main policy is within a block. We use the same mean over heads as in the main text,

$$\mathcal{H}(\{g_{t,b,h}\}_{h=1}^{H}) = \frac{1}{H}\sum_{h=1}^{H} g_{t,b,h}. \tag{17}$$

The reactivation threshold θ and maximum cache age $\Delta_{\max}$ are fixed controller hyperparameters. No cross-block maximum is needed for the depth-cutoff executor because blocks are evaluated sequentially and the first threshold crossing ends reuse for the remaining suffix.

A.3 Algorithm

Algorithm 1 RefAdapt-DiT: Monotone Depth-Adaptive Reference Reuse

Materialize step 0; initialize $\mathcal{Z}_b$ by Equation (11); set $a^{\text{full}} \leftarrow 0$
Store sampled target queries and directional masses for the next step
for denoising step $t = 1, \ldots, N_{\text{step}} - 1$ **do**
  if $t - a^{\text{full}} \geq \Delta_{\max}$ **then**
    Fully recompute the reference stream from block 0; refresh all $\mathcal{Z}_b$ and set $a^{\text{full}} \leftarrow t$
    Record target Q and realized directional masses for the next step
    continue
  end if
  REACTIVATED $\leftarrow$ False
  for block $b = 0, \ldots, B - 1$ **do**
    if REACTIVATED $=$ False **then**
      Form current target Q and compute $D_{t,b}(Q^{\mathrm{Y},h})$ from consecutive target-Q change
      Read $M^{\mathrm{Y}\to\mathrm{R}}_{t-1,b,h}$ from the previous executed attention
      Compute $g_{t,b,h} = D_{t,b}(Q^{\mathrm{Y},h}) M^{\mathrm{Y}\to\mathrm{R}}_{t-1,b,h}$ and $\bar{g}_{t,b} = \mathcal{H}(\{g_{t,b,h}\}_h)$
      if $\bar{g}_{t,b} < \theta$ **then**
        Skip reference rows at block b; target attends to $(\bar{K}^{\mathrm{R}}_b, \bar{V}^{\mathrm{R}}_b)$
      else
        Target still uses block-b cached reference K/V
        Restore $\bar{H}^{\mathrm{R}}_b$ and run reference-query computation with the current target
        Store the reference output as $\bar{H}^{\mathrm{R}}_{b+1}$
        REACTIVATED $\leftarrow$ True
      end if
    else
      Form reference K/V and run the original joint-attention block
      Update $\bar{K}^{\mathrm{R}}_b, \bar{V}^{\mathrm{R}}_b$, store $\bar{H}^{\mathrm{R}}_{b+1}$, and set $a_b \leftarrow t$
    end if
    Record the current target Q and realized runtime mass $M^{\mathrm{Y}\to\mathrm{R}}_{t,b,h}$ for the next step
  end for
end for

Except for the bounded-age fallback that performs a full refresh after $\Delta_{\max}$ steps without a block-0 full refresh, the algorithm uses at most one monotone switch per denoising step. Before the boundary, reference rows are absent and cached K/V are reused. At the boundary, the most recently materialized reference hidden state is restored and the reference query is evaluated against the current target; this updates the reference hidden state for the next block. All deeper blocks then form new reference K/V from that reactivated stream and refresh their caches. As a result, different blocks naturally carry different anchors a_b. The suffix is not identical to a full-fresh forward from block 0 because its starting boundary can be stale; this approximation is included in all reported results.

A.4 Computational Accounting

A full joint-attention block forms four score regions: $Q_{\mathrm{R}}K_{\mathrm{R}}^{\top}$, $Q_{\mathrm{R}}K_{\mathrm{Y}}^{\top}$, $Q_{\mathrm{Y}}K_{\mathrm{R}}^{\top}$, and $Q_{\mathrm{Y}}K_{\mathrm{Y}}^{\top}$. On a cached step, the two reference-query regions are removed. Ignoring head dimensions, score computation changes from

$$\mathcal{O}\big((N_{\mathrm{R}}+N_{\mathrm{Y}})^2\big) \quad \text{to} \quad \mathcal{O}(N_{\mathrm{Y}}(N_{\mathrm{R}}+N_{\mathrm{Y}}))\,, \tag{18}$$

removing approximately $\mathcal{O}(N_{\mathrm{R}}(N_{\mathrm{R}}+N_{\mathrm{Y}}))$ score work. Reference-side QKV projections and other reference-stream operations can also be skipped because the layer-wise reference K/V are cached. These FLOP reductions are theoretical; practical speedups are reported only from measured generation latency under matched software and hardware settings.

B Additional Experimental Details

B.1 Evaluation Protocol

For Qwen Image Edit, we evaluate a fixed 100-case evaluation set from MultiBanana-Benchmark; for MiniMax H3, we evaluate a fixed 100-case evaluation set from EditVerseBench. All runs use NVIDIA H20 GPUs with 96 GB memory per GPU. Qwen uses the distilled 8-step Image Edit checkpoint at 1024×1024, seed 1234, and one GPU. H3 uses the LightX2V release of the distilled 4-step MiniMax H3 checkpoint at 1344×768 (768p), seed 42, and the same GPU type. Within each setting, model, sampler, precision, batch size, and hardware are identical across methods. The separate 50-step replay in Section 5.3 is a controller ablation, not the 4-step main benchmark. Generation latency is measured from real runs; theoretical FLOPs and kernel timings, if reported, are kept separate from measured latency.

B.2 Metrics and Baseline Configuration

For Qwen, the baseline set is TeaCache, SeaCache, TaylorSeer, and ToMe. For MiniMax H3, we use the same set and additionally include AsymRnR. The composed row in each main table combines RefAdapt-DiT with the same orthogonal resolution-compression optimization and is reported separately from the standalone method.

For MultiBanana, we run the benchmark-provided Qwen3-VL-8B-Instruct judge and retain its five 1–10 scores—instruction alignment, reference consistency, background–subject match, physical realism, and visual quality—together with the benchmark's official weighted score $3\,\mathrm{IA} + 3\,\mathrm{RC} + \mathrm{BSM} + \mathrm{PR} + \mathrm{VQ}$. For EditVerseBench, we follow the released evaluation fields: VLM editing quality, Pick Score video quality, frame-level text alignment, video-level text alignment, CLIP temporal consistency, and DINO temporal consistency. We do not collapse the EditVerse metrics into a custom aggregate.

The main tables use generation speedup relative to full computation as the efficiency metric. TaylorSeer uses interval 2 in the Qwen main table. The standalone RefAdapt-DiT rows use the adaptive reference-reuse controller studied in the method section.

B.3 Qwen Performance by Number of References

Table 5 reports the generation speedup as the number of reference images increases. This breakdown is supplementary to the 100-case aggregate in Table 2 and is not used as a separate quality claim.

Table 5: Qwen generation speedup by number of reference images. TaylorSeer uses interval 2. Speedup is relative to full computation at the same reference count.

Method	3 refs	4 refs	5 refs	6 refs	7 refs	8 refs	Overall
ToMe	1.09×	1.08×	1.07×	1.07×	1.06×	1.07×	1.06×
TeaCache	1.13×	1.14×	1.13×	1.14×	1.13×	1.13×	1.13×
SeaCache	1.59×	1.60×	1.60×	1.60×	1.57×	1.62×	1.60×
TaylorSeer (I=2)	1.59×	1.60×	1.60×	1.60×	1.60×	1.61×	1.61×
RefAdapt-DiT	**2.09**×	**2.26**×	**2.38**×	**2.49**×	**2.57**×	**2.62**×	**2.58**×

C Additional Quantitative Analysis

C.1 Mechanism Statistics for H3 and Qwen

Table 6 collects the compact mechanism statistics used in the main-text H3/Qwen comparison. The two settings use different raw token partitions within the shared reference/target roles, so absolute values are reported within each setting rather than pooled across them.

Table 6: Mechanism statistics for the current canonical H3 and Qwen cases. Spearman correlations are within-case diagnostics and are reported separately for each setting.

Diagnostic	H3 Ref2VA	Qwen multi-image
Raw endpoint vs. I^{KV}	0.564	0.708
Endpoint $\times$ exposure vs. I^{KV}	0.826	0.905
Target-Q change vs. I^{KV}_{t+1}	0.607	0.710
Target-Q $\times$ runtime exposure vs. I^{KV}_{t+1}	0.869	0.846
Sensitivity blocks improved by exposure	9/9	7/7

Here I^{KV} denotes the local target-output perturbation produced by replacing current reference-side K/V with stale values from the relevant refresh anchor while holding the target side fixed. "Within-case" means that each Spearman coefficient is computed across aligned internal measurements (e.g., blocks, heads, and anchors) from one canonical trajectory, rather than across independent benchmark inputs. These numbers are therefore mechanism and signal selection diagnostics, not population-level estimates of benchmark behavior.

For Qwen, reversing the order of the same two reference images gives a raw endpoint-to-impact Spearman of 0.685 and an exposure-weighted value of 0.911. The control also shows that the two reference images are not interchangeable: their exposure and individual stale impact differ even when the aggregate reference-side behavior remains similar.

C.2 Reference–Target Dynamics

In the current full video trajectory, the median reference-drift/target-drift ratio over matched K/V measurements in blocks 6–25 is 0.352. The gap narrows around block 31 and reverses in deeper blocks. Target-state change and reference-K/V change also share some timing structure: the median per-block Spearman correlation is 0.675 across 50 blocks, and 22/50 blocks reach at least 0.8. These statistics describe one complete H3 video case and are reported as within-case diagnostics.

C.3 Directional Attention Diagnostics

For target-video queries in the current video case, Ref receives 9.51% of attention mass on average and the remaining 90.49% is assigned to Tgt. Reference-video queries assign 26.14% mass to target-video keys, versus 9.51% in the opposite direction; the directional difference is positive in 97.22% of step–block cells. These values are structural diagnostics and are not used to infer target dependence from reference-query attention.

C.4 Runtime-Signal Diagnostics

Table 7 complements the cross-model summary in Table 6 with H3-only runtime-signal checks that are not already reported there. In particular, it contrasts the deployed consecutive-change design with an anchor-relative alternative.

The anchor-relative signal measures displacement from the most recent block-specific materialization anchor, which may be several denoising steps old, whereas the deployed Target-Q signal measures only the consecutive step-to-step change. The former therefore accumulates motion over a variable lag and has a different numerical scale and trigger history under the same controller threshold. On the current ablation, this makes the anchor-relative variant more conservative about reuse and yields lower measured speedup; this comparison should be read as a fixed-controller ablation rather than as a threshold-matched universal ranking of the two signal families.

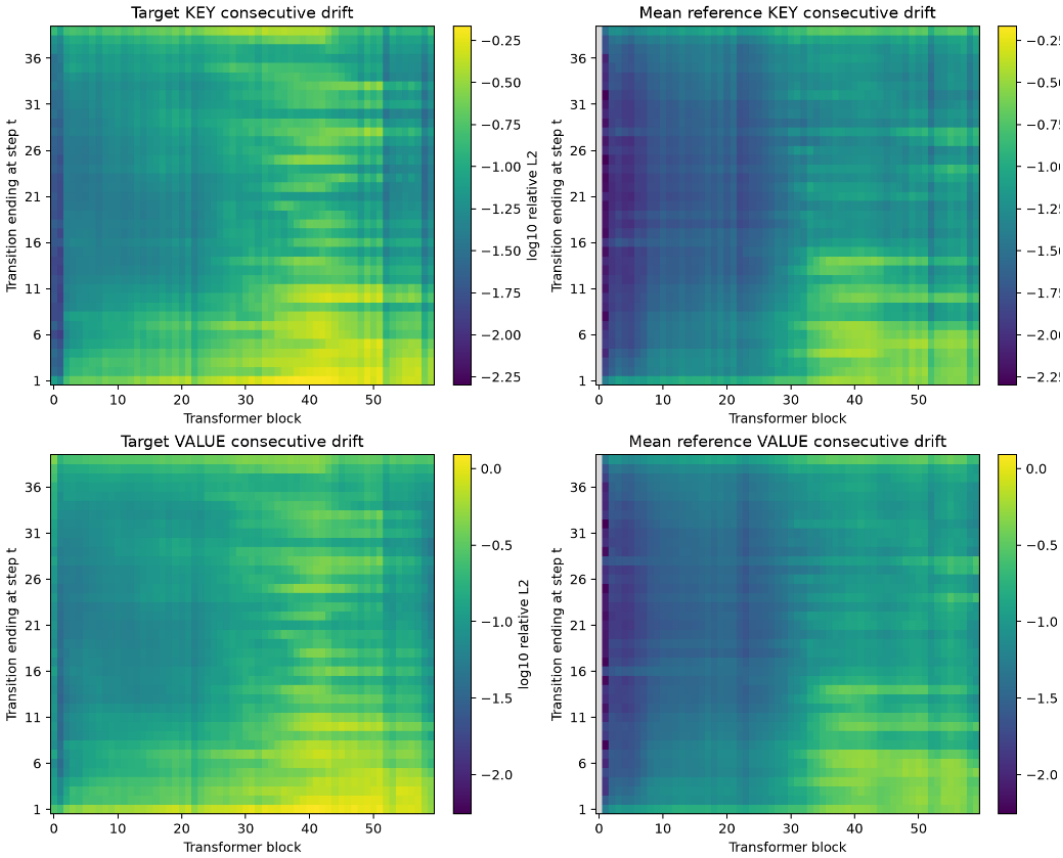


Figure 7: Retained left-half panels from the full-block Qwen reference/target dynamics diagnostics. These four panels report target and mean-reference consecutive drift for K and V across blocks and denoising steps; the dropped right-half panels contained the corresponding log-ratio and per-block timing-coupling summaries.

Table 7: Additional H3 runtime-signal diagnostics from one complete Ref2VA trajectory. Correlations are within-case screening statistics, not confidence estimates over independent inputs.

Signal / target	Pearson ↑	Spearman ↑
Runtime stale-reference exposure vs. I_{t+1}^{KV}	0.479	0.789
Target-Q anchor change vs. I_{t+1}^{KV}	0.390	0.418
Anchor change $\times$ exposure vs. I_{t+1}^{KV}	0.589	0.545

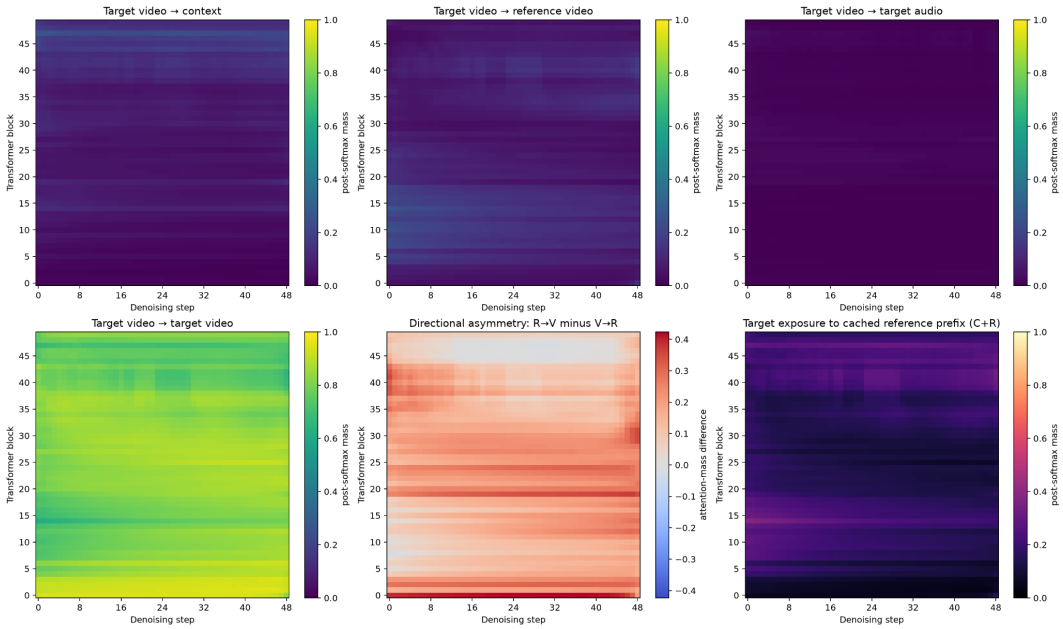
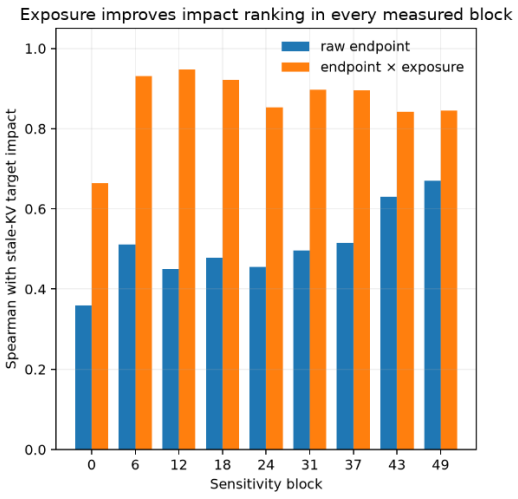
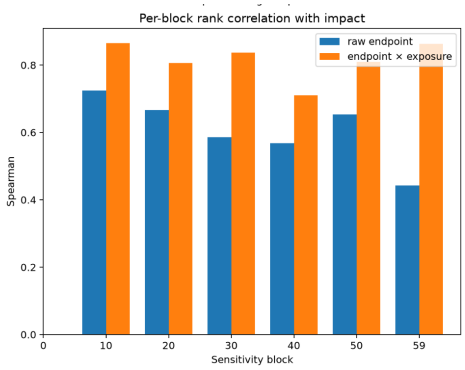


Figure 8: Complete H3 raw-partition attention diagnostics. The main text keeps target exposure to Ref and aggregates the remaining mass as Tgt; the remaining panels report implementation-level allocation and directional asymmetry. Every arrow in the panel titles runs from query role to key role, not along causal flow.



(a) MiniMax H3



(b) Qwen Image Edit

Figure 9: Block-wise support for exposure conditioning. In every measured sensitivity block, endpoint drift weighted by target exposure has higher Spearman association with local stale-KV impact than raw endpoint drift (9/9 H3 blocks and 7/7 Qwen blocks).

C.5 Probe Non-Intrusiveness and Multi-Reference Controls

For the canonical Qwen probe, enabling the observation sidecar leaves the final PNG bit-identical to the probe-off run (matching SHA-256). Probe-on generation takes 247.1 s versus 232.9 s probe-off, corresponding to roughly 6.1% overhead from the optional research logging sidecar. This sidecar is used to collect additional diagnostic traces and is distinct from the runtime controller; it is not part of the production configuration whose speedups are reported in the main text.

D Additional Qualitative Results

D.1 Prompts Used for the Figure 1 Cross-Prompt Profiles

The magnitude profiles in Figure 1 use the following three prompts per model. Within each model, the reference inputs, seed, resolution, and denoising configuration are held fixed; only the prompt changes. The heatmap averages Equation (2) cellwise across these prompts. The profile first averages each prompt over denoising transitions and then averages the prompt-level values as specified in Equation (3).

Model	Case	Prompt
MiniMax H3	`canonical`	Reimagine the animal from the reference video as a giant panda ambling through the snowy forest, same motion and camera path, breath misting, soft wind and distant birdsong
MiniMax H3	`red_fox_autumn`	Reimagine the animal from the reference video as a red fox trotting through an autumn maple forest, preserve the original motion and camera path, rustling leaves, warm sunset light, and natural forest ambience
MiniMax H3	`robot_dog_neon`	Reimagine the animal from the reference video as a small silver robot dog exploring a rain-soaked neon city street, preserve the original motion and camera path, reflected lights, light rain, and distant traffic ambience
Qwen	`canonical`	the woman and man are hugging together
Qwen	`outdoor_cafe`	Place the woman and the man seated across from each other at a small outdoor cafe table, preserving both identities and realistic facial details
Qwen	`umbrella_neon`	Place the woman and the man standing back-to-back under one transparent umbrella on a rainy neon-lit street, preserving both identities and realistic facial details

Table 8: Exact prompts used to construct the three-prompt reference-drift profiles in Figure 1.