MedCache: Training-Free Spatially Aware Caching for Accelerated Medical Video Generation

Ufaq Khan[1,*], **Umair Nawaz**[1], **Sathira Silva**[1], **Numan Saeed**[1], **Muhammad Umer Sheikh**[1], **Muhammad Bilal**[3], **Junaid Qadir**[4], **Lena Maier-Hein**[1,2], **Yutong Xie**[1], **Muhammad Haris Khan**[1]

[1]Mohamed bin Zayed University of Artificial Intelligence, UAE
[2]DKFZ Heidelberg, Helmholtz Imaging, Germany
[3]Birmingham City University, UK
[4]Qatar University, Qatar

Project Page: `ufaqkhan.github.io/MedCache-Website`

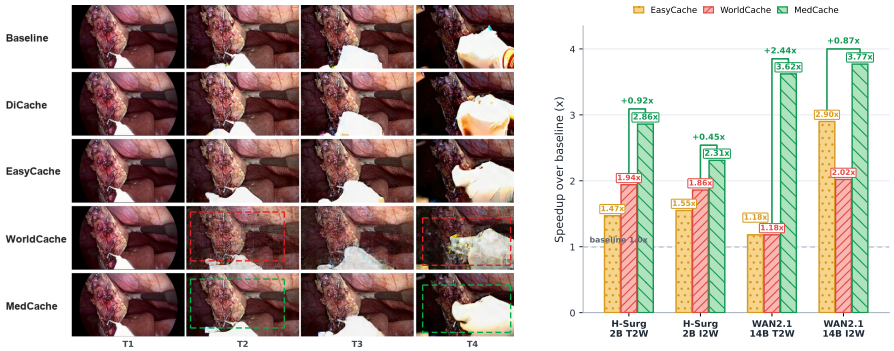


Figure 1: *Left:* under the same Image2World conditioning, WorldCache [Nawaz et al., 2026] introduces local artifacts around the active tool-tissue region, while MedCache remains closer to the uncached baseline. *Right:* speedup over the uncached baseline across Cosmos-H-Surgical 2B [He et al., 2025] and WAN2.1 14B [Wan et al., 2025] Text2World/Image2World settings. MedCache improves over both EasyCache [Zhou et al., 2025] and WorldCache [Nawaz et al., 2026], with gains shown relative to the stronger prior cache in each setting.

Abstract

Video world models are increasingly adapted to medical applications such as surgical simulation, procedural rehearsal, and synthetic data generation, but their inference cost remains a major barrier to practical use. Training-free caching can accelerate existing checkpoints without retraining, yet current methods trade decision quality against decision cost. Lightweight caches reduce the latent change to a single global statistic that weights every region equally, which is a poor fit for medical video, where clinically meaningful change is often confined to a small region, such as a tool-tissue interaction. Probe-based caches recover spatial awareness by partially running the network at every step, but this cost is paid whether or not the step is skipped, eroding the speedup. Our key observation is that the spatial signal needed for selective reuse is already present in the latent. We introduce **MedCache**, a training-free and probe-free cache for rectified-flow video generation. MedCache computes a spatial risk map directly from the latent with inexpensive arithmetic and uses it to weight the accumulated drift so that changes in high-risk regions

*Corresponding author: ufaq.khan@mbzuai.ac.ae

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

count more toward the recompute decision. A prompt-derived domain prior sets the operating point, and a lazy consistency guard runs only when reuse is no longer clearly safe. Across three open world-model families (WAN2.1, Cosmos-Predict-2.5, and Cosmos-H-Surgical) and a HunyuanVideo transfer evaluation, MedCache accelerates every model tested and gives the strongest speed-quality trade-off in the medical and image-conditioned settings that motivate its design. On Cosmos-H-Surgical 2B Text2World, it raises Total quality from 0.767 to 0.837, achieving a $2.86\times$ speedup over the uncached baseline. Ablations show that risk-weighted drift is the main source of runtime gain, while the prompt prior and the lazy guard are optional controls that tune the operating point.

1 Introduction

Video world models have advanced quickly enough that medical applications are now within reach. Surgical training, interactive simulation, procedural rehearsal, and synthetic data generation all benefit from the ability to generate plausible medical video under controlled conditions, and a model is practically useful in these settings only if it can generate video quickly enough to support repeated sampling, iterative analysis, and human-in-the-loop workflows. Recent models such as WAN2.1 [Wan et al., 2025], Cosmos-Predict-2.5 [Ali et al., 2025], and medical adaptations such as Cosmos-H-Surgical [He et al., 2025] show that high-fidelity video generation is now feasible in both general and specialized domains. The remaining obstacle is the inference cost. Generating a short clip of roughly 5-6 seconds typically requires 30-50 denoising steps through a multi-billion-parameter diffusion transformer [Peebles and Xie, 2023], and classifier-free guidance (CFG) [Ho and Salimans, 2022], which doubles the number of forward passes. Even on recent high-performance GPUs, video generation is therefore too slow [Ma et al., 2025] for the iterative and interactive workflows that motivate medical use.

This bottleneck is especially restrictive in medical video generation, where models are often specialized to narrow procedural domains or limited datasets [He et al., 2025]. In this setting, distillation-based acceleration [Salimans and Ho, 2022, Meng et al., 2023, Luo et al., 2023] is less attractive because it adds another training stage, while architectural changes such as block skipping [Shen et al., 2024, Chen et al., 2025] require modifying the model or inference stack. Training-free acceleration is therefore particularly relevant.

Inference-time caching is one such approach. It relies on a simple empirical regularity of diffusion and rectified-flow sampling: across adjacent denoising steps, the residual transformation ($\Delta_t = f_\theta(x_t) - x_t$) often changes slowly enough that a previous computation can be reused instead of recomputed from scratch. The difficulty is deciding when that reuse is still reliable, and existing training-free caches expose a clear trade-off between decision quality and decision cost. EasyCache [Zhou et al., 2025] reduces the entire frame to a single scalar that estimates how much the latent has drifted, and skips the next forward pass while that scalar stays below a threshold. The decision is essentially free, but it is also spatially blind: a small change in a clinically critical region, such as a tool-tissue contact, and a large but inconsequential change in the background contribute equally to the same scalar. WorldCache [Nawaz et al., 2026] recovers spatial awareness by partially evaluating the network at every step and using those intermediate features, together with online system identification and latent optical flow, to make a richer decision. The signal is better, but it is no longer free: the partial forward pass is paid on every step, including steps that are eventually skipped, so the decision cost itself eats into the speedup it enables as depicted in Figure 9 of the appendix.

Our key observation is that this trade-off is not fundamental. The information needed to identify spatially risky reuse is already present in the latent representation: boundaries, fine local structure, and localized motion leave measurable signatures that can be extracted with affordable arithmetic, without involving the network at all. This motivates a third design point: keep the inexpensive decision path of scalar drift caching, but replace its spatially uniform statistic with one that emphasizes where reuse errors are most likely to matter.

We instantiate this idea in **MedCache**, a training-free and probe-free cache for rectified-flow video generation. MedCache augments accumulated-drift caching with three lightweight components: a latent-derived spatial risk map that highlights regions where reuse errors are more likely to be consequential, a prompt-derived domain prior that adjusts skip aggressiveness to the medical context,

and a lazy Laplacian consistency check that is evaluated only when the current pass is no longer clearly safe. Together, these components give MedCache a fast path comparable to EasyCache [Zhou et al., 2025] and a spatial selectivity comparable to WorldCache [Nawaz et al., 2026], without the unconditional probe cost that WorldCache pays at every step. Across three open world-model families in Text2World and Image2World settings, MedCache achieves up to **2.86**$\times$ wall-clock speedup over the uncached baseline on Cosmos-H-Surgical 2B while also improving quality over the strongest training-free baseline, as depicted in Figure 1.

We make three contributions. ***First***, we identify that medical video caching does not require probing the generator: the latent itself contains enough spatial signal to make selective skip decisions cheaply, which closes the gap between scalar-drift and probe-based caches. ***Second***, we introduce MedCache, a training-free probe-free cache built around a latent-derived spatial risk map, a prompt-derived domain prior, and a lazy consistency check, where each component plays a distinct role in the speed-quality trade-off. ***Third***, we evaluate MedCache across three main open world-model families (WAN2.1, Cosmos-Predict-2.5, Cosmos-H-Surgical) in both Text2World and Image2World settings, and additionally include a HunyuanVideo [Kong et al., 2024] transfer evaluation on VBench [Huang et al., 2024]. We also analyze robustness to every exposed constant, aggregate drift concentration, risk-map staleness, and cache fidelity inside annotated instrument regions.

2 Related Work

Video world models and medical video generation. Recent progress in open video world models has made it possible to study inference efficiency under realistic deployment settings rather than only in small-scale prototypes. The Cosmos family [Agarwal et al., 2025, Ali et al., 2025] frames world modeling as a unified generative platform for physical AI. WAN2.1 [Wan et al., 2025] likewise demonstrates that large rectified-flow video generators can achieve strong motion quality and broad applicability at open-model scale. In parallel, medical video generation has evolved from early proof-of-concept systems toward more specialized surgical and endoscopic models. Endora [Li et al., 2024], MedSora [Wang et al., 2024], and SurGen [Cho et al., 2024] show that domain-adapted generative models can capture procedural dynamics and anatomical appearance with increasing fidelity. At the same time, these works make clear that generation cost remains a practical limitation, especially in medical settings where repeated sampling, large-scale evaluation, and interactive simulation are often required.

Training-free acceleration for diffusion and video generation. Recent studies show how to reduce the cost of diffusion inference without retraining the underlying generator. Early approaches such as DeepCache [Ma et al., 2024] exploit temporal redundancy across denoising steps by reusing intermediate features. TeaCache [Liu et al., 2025] uses timestep-aware variation in model output to guide caching, and FasterCache [Lv et al., 2025] further leverages redundancy within classifier-free guidance to improve efficiency. EasyCache [Zhou et al., 2025] also offers a minimal, practical, and training-free design where it relies on an accumulated drift signal, incurs little decision overhead, and can be integrated into existing samplers with few modifications. However, the reuse rule in EasyCache is fundamentally global and therefore cannot distinguish between benign changes and errors occurring in spatially important regions. ErtaCache [Peng et al., 2026] further analyzes cache-induced degradation as feature shift and step amplification errors, and mitigates them through offline residual profiling, timestep adjustment, and residual error rectification. Similarly, WorldCache [Nawaz et al., 2026] enriches the decision process through content-aware signals, built on top of DiCache [Bu et al., 2026] by probing the model, improving selectivity but incurring additional per-step overhead. Our method is motivated by the gap between these two design points: retaining the simplicity of lightweight caching while adding the spatial awareness that is particularly important in medical video.

Adaptive computation in diffusion models. Early-exit methods reduce cost by removing computation within the network according to timestep-dependent schedules [Moon et al., 2024]. Such approaches reduce the amount of computation performed inside a forward pass. However, caching addresses a complementary question, i.e., whether a full forward pass is needed at all at a given step. Distillation-based acceleration [Salimans and Ho, 2022] is also being used in the literature, but it operates under a different regime as it requires additional training.

3 Preliminaries and Core Hypothesis

The central question in training-free caching is not whether adjacent denoising steps are redundant, but how to decide when reuse is safe. Existing methods make this decision in two different ways. Lightweight drift-based caches like EasyCache [Zhou et al., 2025] use a global scalar statistic and are therefore inexpensive, but they discard where the latent change occurs. Probe-based caches like DiCache [Bu et al., 2026] or WorldCache [Nawaz et al., 2026] recover content awareness by partially executing the network before making the reuse decision, but this introduces a non-negligible decision cost at every pass.

Core hypothesis. We hypothesize that medical video generation admits a third design point: Cache safety can be made spatially selective without probing the generator. The reason is that clinically consequential reuse errors are rarely uniformly distributed over the frame. They are typically concentrated around anatomical interfaces, tool-tissue contact, fluid boundaries, and motion-sensitive local structures, as also exemplified in Figure 5 of the Appendix. These cues are already partially encoded in the latent representation. Therefore, a cache decision that estimates spatial risk directly from the latent should preserve the low overhead of scalar drift caching while improving fidelity in medically important regions.

Sampling and Cache Reuse. Let x_t denote the latent at denoising time t, and let c denote the conditioning signal. A rectified-flow video generator parameterizes a velocity field $v_t = f_\theta(x_t, t, c)$, which is integrated over T denoising steps to transform an initial noise sample x_T into a sample x_0. Under classifier-free guidance (CFG) [Ho and Salimans, 2022, Luo et al., 2023], the conditional and unconditional predictions are combined as $v_t = v_t^u + s\big(v_t^c - v_t^u\big)$, where v_t^c and v_t^u denote the conditional and unconditional branches, respectively, and s is the guidance scale.

Because CFG doubles the internal evaluations, we distinguish *visible denoising steps* from *internal forward passes*. We index internal passes by n, with even n corresponding to the conditional branch and odd n to the unconditional branch. Cache state must remain branch-consistent, so the two branches are tracked separately. Following prior training-free caching methods, we define the residual update at internal pass n as $\Delta_n = f_\theta(x_n) - x_n$. If pass n is skipped, the sampler reuses the most recent residual from the same branch, $\hat{y}_n = x_n + \Delta_{n-2}$, that is the reuse rule on which both EasyCache and MedCache operate.

Training-Free Caching. Inference-time caching exploits the empirical fact that adjacent denoising passes often induce similar residual updates. EasyCache [Zhou et al., 2025] expresses this idea with a scalar accumulated-error estimate:

$$\hat{E}_n = \hat{E}_{n-2} + k_n \cdot \frac{\|x_n - x_{n-2}\|_1}{\|f_\theta(x_{n-2})\|_1 + \epsilon}, \qquad \text{skip if } \hat{E}_n < \tau, \tag{1}$$

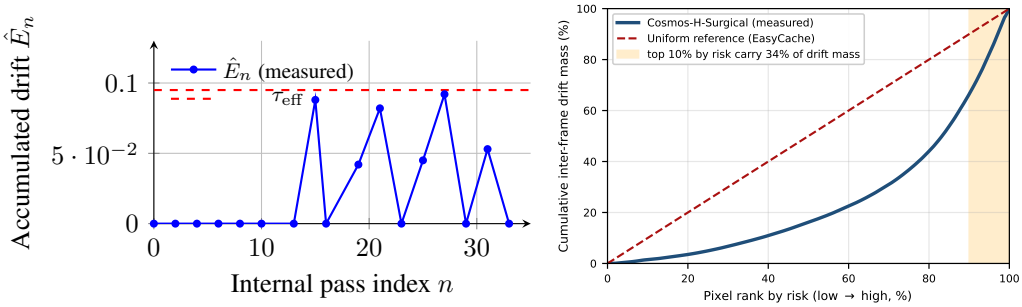


Figure 2: **Why scalar drift is insufficient for medical video caching. Left (a):** accumulated-drift trajectory of an EasyCache-style configuration, showing the characteristic rise-and-reset pattern of scalar thresholding; the longest skip streak spans 16 internal passes. **Right (b):** cumulative drift mass ranked by risk for one representative decision, where the top 10% highest-risk pixels account for 34% of the drift mass. This panel is an illustration rather than a dataset average; Appendix B.4 reports the dataset-level analysis. Together, the plots show that scalar drift discards spatial structure while providing no localized quality signal.

where τ is the skip threshold and k_n is a local sensitivity factor estimated from the most recent forced-compute step. This rule is attractive because it is training-free, simple to implement, and nearly cost-free on skipped passes. Its limitation is equally direct: the decision statistic is a single global scalar, so any spatial variation in reuse risk is discarded before the skip decision is made. In medical video, where the most important changes are often localized, this makes the cache both hard to tune and easy to oversimplify. Therefore, the next subsection isolates the two constraints that motivate MedCache.

3.1 Design Constraints from Medical Video Generation

Constraint 1: the cost of reuse error is spatially asymmetric. In medical video, the regions that matter most are often small: tool boundaries, tissue interfaces, fluid boundaries, and small anatomical motions. A small reuse error at a tool-tissue contact can matter more than a larger error in the background, yet a spatially averaged drift statistic weighs both equally. Latent drift is also unevenly distributed: for the representative decision in Figure 2b, the top 10% of risk-ranked positions hold 34% of the drift mass, and across the full Cosmos-H-Surgical Image2World evaluation, the risk-ranked positions hold more than their uniform share of drift at every support size we tested (Appendix B.4). The medical motivation is therefore not that a small region must contain most of the frame-wide drift, but that the consequences of cache errors are spatially asymmetric.

Constraint 2: the decision path must remain inexpensive. A richer decision signal can be obtained by probing the generator itself, as in WorldCache [Nawaz et al., 2026], but this changes the cost structure because part of the network must still be executed even on passes that are eventually skipped. If the first d blocks of an L-block backbone are always evaluated, the expected per-pass cost at skip rate r_{skip} is $\frac{d}{L} + (1 - r_{\text{skip}})\left(1 - \frac{d}{L}\right)$, which yields the speedup bound

$$\text{Speedup}_{\max} = \frac{1}{\frac{d}{L} + (1 - r_{\text{skip}})\left(1 - \frac{d}{L}\right)}. \tag{2}$$

Thus, a better decision signal is only useful if it does not erase the benefit of reuse through fixed per-pass overhead.

4 MedCache

The starting point is the trade-off established in Sec. 3: a scalar reuse criterion is inexpensive but spatially blind, whereas a richer probe-based criterion is more selective but introduces a fixed per-pass cost. Figure 3 shows the MedCache framework, which is designed to address this trade-off directly. MedCache has one core component and two optional controls. The core is a *latent-derived spatial risk (LSR) map* that identifies regions where reuse error is more likely to matter. MedCache uses it to replace the uniform drift term with a *risk-weighted drift* term (Secs. 4.1-4.2). On top of this, a *prompt-derived domain prior* adjusts the operating point by scaling the skip threshold (Sec. 4.3), and a *lazy high-frequency guard* acts as a fidelity safeguard near the reuse boundary (Sec. 4.4). All components operate at latent resolution inside the existing branch-wise caching framework. Algorithm 1 (Appendix E) gives the complete per-pass decision.

4.1 Latent-Derived Spatial Risk Estimation

The central novelty in MedCache is a spatial risk map $r \in [0, 1]^{B \times 1 \times H \times W}$, computed directly from the current latent state. The purpose of r is not to estimate semantics explicitly, but to identify spatial locations where a reused residual is more likely to produce visible or clinically relevant degradation. To construct r, we first collapse the latent over channels and time:

$$\bar{x} = \frac{1}{CT} \sum_{c,t} x_{b,c,t,h,w}. \tag{3}$$

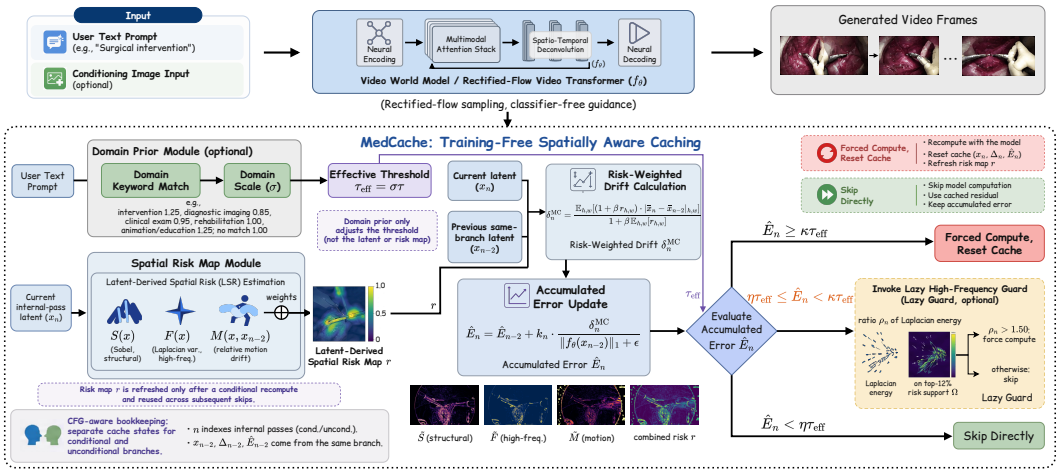


Figure 3: MedCache adds a training-free cache-decision layer to rectified-flow video generation. A prompt-derived domain prior sets the effective threshold, a latent-derived spatial risk map reweights drift estimation, and a lazy high-frequency guard is evaluated only in borderline cases. The controller chooses between direct skip, guarded skip, and forced recomputation with cache refresh.

This produces one spatial map per sample while preserving the local structure needed for a cache decision. From $\bar{x}$, we derive three complementary spatial energies:

$$S(x) = |\nabla_x \bar{x}| + |\nabla_y \bar{x}|, \qquad \text{(structural boundary energy)}, \tag{4}$$

$$F(x) = \mathrm{Var}_{3\times 3}\big[(\Delta * \bar{x})\big], \qquad \text{(local high-frequency energy)}, \tag{5}$$

$$M(x, x_{n-2}) = \frac{|\bar{x} - \bar{x}_{n-2}|}{|\bar{x}| + |\bar{x}_{n-2}| + \epsilon}, \qquad \text{(relative motion drift)}. \tag{6}$$

Here, ∇_x and ∇_y denote Sobel gradients and $\Delta * (\cdot)$ denotes the discrete Laplacian. The three cues target different reuse failures. S marks persistent interfaces, such as organ boundaries and tool-tissue contours, where a reused residual can displace or blur an edge. F marks thin or rapidly varying detail that first-order edges under-represent, such as fine sutures, narrow interfaces, and endoscopic texture. M marks local change within an otherwise stable scene, such as instrument motion, tissue deformation, or moving fluid. S and F describe the current spatial state at two derivative orders, while M describes how that state changes. A static boundary can have high S but low M, whereas a fine texture can have high F without a strong coarse edge, and a locally moving instrument can have high M while most of the frame is unchanged. A single cue can therefore miss failures that the others capture. The self-normalized form in Eq. 6 reduces sensitivity to variation in latent magnitude across denoising stages.

Each component is min-max normalized over the spatial domain and then combined linearly:

$$r = \mathrm{clip}_{[0,1]}\big(\alpha_S \tilde{S} + \alpha_F \tilde{F} + \alpha_M \tilde{M}\big), \qquad (\alpha_S, \alpha_F, \alpha_M) = (1.00,\ 0.75,\ 1.25). \tag{7}$$

The motion term is weighted slightly more strongly than the other two components, while the high-frequency term is weighted slightly less strongly to reduce redundancy with the structural term. Because each cue is min-max normalized before Eq. 7, the sum combines dimensionless spatial rankings rather than quantities with incompatible units. The non-negative sum acts as a union-of-evidence rule: any cue can raise a location's risk, the cues cannot cancel one another, and clipping limits extreme responses. The map ranks locations by the risk of applying a reused residual. The resulting map is computed entirely with low-cost operations at latent resolution and requires neither a learned saliency head nor a model-side probe. In our implementation, r is refreshed on every conditional cache failure and then reused for the corresponding CFG pair and subsequent skips until the next refresh. This keeps the current spatial signal without incurring the cost of recomputing it on every internal pass. Appendix B.6 measures how far the cached map drifts from a freshly computed one during a skip streak.

4.2 Risk-Weighted Accumulated Drift

MedCache retains the accumulated-error structure of Eq. 1, but changes how the latent change is aggregated before it enters the accumulator. EasyCache uses a spatially uniform drift statistic, so that all locations contribute equally. MedCache instead defines a risk-weighted drift term,

$$\delta_n^{\mathrm{MC}} = \frac{\mathbb{E}_{h,w}[(1+\beta r_{h,w}) \cdot |\bar{x}_n - \bar{x}_{n-2}|_{h,w}]}{1+\beta\,\mathbb{E}_{h,w}[r_{h,w}]}, \qquad \beta = 6.0. \tag{8}$$

The numerator increases the influence of high-risk regions while still retaining information from the rest of the frame. The denominator keeps the overall scale comparable to the unweighted statistic, so that Eq. 8 reduces to the uniform drift case when r is spatially uniform.

The accumulated error estimate is then updated by substituting δ_n^{MC} for the uniform latent-change term in Eq. 1. The resulting recursion has the same structure as the baseline cache, but is sensitive to *where* changes occur rather than only to their global average magnitude.

4.3 Prompt-Derived Domain Prior

The latent does not tell the cache how much reuse error a given type of content can tolerate. To inject this information without adding a learned controller, MedCache uses a prompt-derived threshold prior. It takes this optionally from the prompt by scaling the skip threshold, $\tau_{\text{eff}} = \sigma(p)\,\tau$, where $\sigma(p)$ comes from a fixed keyword lookup with five groups: INTERVENTION (e.g., incision, suture, electrocautery; $\sigma = 1.25$), DIAGNOSTIC-IMAGING (e.g., ultrasound, MRI, CT, X-ray; 0.85), CLINICAL-EXAM (e.g., palpation, auscultation, examination; 0.95), REHABILITATION (e.g., physiotherapy, gait training, mobility exercise; 1.00), and ANIMATION/EDUCATION (e.g., lecture, anatomy diagram, instructional; 1.25). The prior acts only through the threshold, and it does not enter the risk map, the risk-weighted drift, or the guard. If no keyword matches, $\sigma = 1$ and MedCache is the complete prompt-agnostic controller, and thus a keyword miss neither disables caching nor alters the risk map.

4.4 Lazy High-Frequency Consistency Guard

The risk-weighted accumulator is still a heuristic, so MedCache adds a direct consistency check for passes that are no longer clearly safe. The key design choice is that this check is not evaluated unconditionally. Instead, MedCache partitions the current pass into three regimes according to the accumulated error:

$$\text{decision at pass } n = \begin{cases} \text{force compute}, & \hat{E}_n \geq \kappa\tau_{\text{eff}}, \\ \text{evaluate guard}, & \eta\tau_{\text{eff}} \leq \hat{E}_n < \kappa\tau_{\text{eff}}, \\ \text{skip directly}, & \hat{E}_n < \eta\tau_{\text{eff}}, \end{cases} \qquad \kappa = 1.10,\ \ \eta = 0.20. \tag{9}$$

The factor κ defines a narrow over-threshold band beyond which MedCache does not attempt reuse. The factor η determines when the guard starts to operate within the nominally safe region. With $\eta = 0.20$, the guard is inactive only in the lowest fifth of the safe band.

When the guard is evaluated, MedCache forms the cache-projected output $\hat{y}_n = x_n + \Delta_{n-2}$, comparing its Laplacian energy to that of the most recent computed output over the highest-risk spatial support. Let Ω denote the top-$K = 12\%$ support of the current risk map. We compute

$$\rho_n = \frac{\sum_{(h,w)\in\Omega}(\Delta * \hat{y}_n)^2}{\sum_{(h,w)\in\Omega}(\Delta * y_{n-2})^2 + \epsilon}, \qquad \text{force compute if } \rho_n > 1.50. \tag{10}$$

If the guard does not trigger, the pass is skipped. The guard is therefore spatially restricted and invoked only after the inexpensive drift test has placed the current pass outside the clearly safe regime. Its role is not to replace the accumulator but to prevent the decision to reuse from being made solely by a scalar heuristic as the pass approaches the decision boundary.

4.5 CFG-Aware Bookkeeping

A final implementation detail is required for correct behavior under classifier-free guidance. Since each visible denoising step consists of two internal forward passes, all cache counters must be defined

over the internal pass index rather than over visible steps. Otherwise, the warm-up, cutoff, or reset logic can be prematurely triggered and remove valid reuse opportunities in the second half of the trajectory. Consequently, if the sampler uses T visible denoising steps, we set the internal horizon to $n_{\max} = 2T$ and define all scheduling windows in the same indexing scheme. In our implementation, this corresponds to using $W = 2W_0$ for warm-up and applying the terminal always-compute region over the final internal passes. We use $W_0 = 6$. Although this bookkeeping correction is simple, it affects both speedup and fidelity and is therefore treated as part of the method rather than as an implementation footnote.

Table 1: **Text2World results across small-scale and large-scale base models.** Among accelerated methods, the best value is shown in **bold**, and the second-best is underlined.

Model	Method	Quality Metrics ↑					Latency (s) ↓	SpeedUp ↑
		IQ	SC	BC	MS	Total		
Small-scale models - Text2World								
Cosmos-Predict-2.5 2B	Baseline	0.751	0.959	0.961	0.991	0.916	271.00	1.00×
	FasterCache	0.383	0.917	0.943	0.936	0.795	**102.00**	**2.65**×
	EasyCache	0.691	0.878	0.908	0.975	0.863	198.00	1.37×
	DiCache	0.725	**0.948**	**0.957**	0.977	**0.902**	205.00	1.32×
	WorldCache	0.701	0.932	0.943	0.987	0.891	128.00	2.12×
	MedCache	**0.747**	0.918	0.942	**0.990**	0.899	108.00	2.51×
Cosmos-H-Surgical 2B	Baseline	0.625	0.906	0.934	0.982	0.862	272.00	1.00×
	FasterCache	0.348	**0.893**	**0.971**	0.962	0.794	130.00	2.09×
	EasyCache	0.462	0.774	0.899	0.933	0.767	185.00	1.47×
	DiCache	0.445	0.868	0.930	0.981	0.806	250.00	1.09×
	WorldCache	0.467	0.877	0.929	0.984	0.814	140.00	1.94×
	MedCache	**0.549**	0.884	0.928	**0.986**	**0.837**	**95.00**	**2.86**×
Large-scale models - Text2World								
WAN2.1 14B	Baseline	0.713	0.947	0.968	0.990	0.904	663.52	1.00×
	EasyCache	**0.713**	**0.933**	**0.963**	**0.987**	**0.898**	560.62	1.18×
	FasterCache	0.711	0.929	0.961	0.986	0.897	345.58	1.92×
	DiCache	0.713	0.933	0.960	0.987	0.898	562.77	1.18×
	WorldCache	0.713	0.933	0.960	0.987	0.898	561.13	1.18×
	MedCache	0.704	0.925	0.949	0.987	0.891	**183.09**	**3.62**×
Cosmos-Predict-2.5 14B	Baseline	0.726	0.929	0.941	0.987	0.896	767.97	1.00×
	FasterCache	0.478	0.916	0.936	0.886	0.804	442.36	1.74×
	EasyCache	0.706	0.928	**0.944**	0.986	0.891	434.60	1.78×
	DiCache	0.723	**0.928**	0.942	0.985	0.895	700.64	1.10×
	WorldCache	0.715	0.916	0.937	0.984	0.888	365.05	2.10×
	MedCache	**0.725**	0.923	**0.944**	**0.987**	**0.895**	**331.40**	**2.32**×

5 Results

Experimental Setup. We evaluate MedCache on three open rectified-flow video generators spanning both general-domain and medical-domain generation: WAN2.1 [Wan et al., 2025], Cosmos-Predict-2.5 [Ali et al., 2025], and Cosmos-H-Surgical [He et al., 2025]. Experiments are organized by model scale, where the corresponding checkpoints are publicly available. Since MedCache is training-free, all methods are evaluated directly on the released checkpoints without modifying model weights. Although Cosmos-H-Surgical has been officially released for image-conditioned prediction, the inherited unified Cosmos-Predict2.5 architecture supports zero-frame conditioning. We therefore report text-only results (Text2World) as a controlled out-of-distribution cache stress test, applying identical inputs and sampling configurations to all methods.

We consider two generation settings: **Text2World** and **Image2World**. In Text2World, the model generates a video from a text prompt obtained from the test set of [Wang et al., 2025], also known as Med-VBench. In Image2World, generation is conditioned on an input image. For this setting, we used 300 samples from the conditional-generation portion of SurgSigma [Zeng et al., 2026], selected according to the subset of CholecT50 [Nwoye and Padoy, 2022]. We compare MedCache against the uncached **Baseline** and four training-free baselines: **FasterCache**, **DiCache**, **EasyCache**, and **WorldCache**.

For evaluation, we report four quality metrics: *Image Quality (IQ)*, *Subject Consistency (SC)*, *Background Consistency (BC)*, and *Motion Smoothness (MS)*, following [Wang et al., 2025]. Their arithmetic mean is reported as *Total*. We also report *Latency* and *SpeedUp* relative to the uncached baseline. Additional details on benchmark construction, hardware, latency measurement, and hyperparameter selection are provided in Appendices A and B.

Text2World. Table 1 shows that MedCache gives the strongest overall speed-quality trade-off on the Cosmos checkpoints. On Cosmos-Predict-2.5 2B, it improves over EasyCache from 0.863 to 0.899 in Total while reducing latency from 198.00 s to 108.00 s; on Cosmos-H-Surgical 2B, it improves

Table 2: **Image2World results across small-scale and large-scale base models.** Best accelerated value is shown in **bold** and the second-best is underlined.

Model	Method	Quality Metrics ↑					Latency (s) ↓	SpeedUp ↑
		IQ	SC	BC	MS	Total		
Small-scale models - Image2World								
Cosmos-Predict-2.5 2B	Baseline	0.640	0.718	0.850	0.959	0.792	52.14	1.00×
	FasterCache	0.377	0.510	0.708	0.920	0.629	29.24	1.78×
	EasyCache	0.620	0.720	0.851	0.960	0.788	43.59	1.20×
	DiCache	**0.633**	0.714	0.845	0.959	0.788	36.61	1.42×
	WorldCache	0.581	0.717	0.848	0.964	0.778	27.50	1.90×
	MedCache	0.628	**0.723**	**0.862**	**0.968**	**0.795**	**22.91**	**2.28**×
Cosmos-H-Surgical 2B	Baseline	0.604	0.881	0.948	0.968	0.850	48.65	1.00×
	FasterCache	0.414	0.514	0.720	0.901	0.637	23.57	2.06×
	EasyCache	0.590	0.882	0.947	0.970	0.847	31.45	1.55×
	DiCache	0.577	0.861	0.934	0.965	0.834	29.25	1.66×
	WorldCache	0.587	0.864	0.933	**0.973**	0.839	26.12	1.86×
	MedCache	**0.592**	**0.884**	**0.950**	**0.973**	**0.850**	**21.05**	**2.31**×
Large-scale models - Image2World								
WAN2.1 14B	Baseline	0.719	0.876	0.956	0.987	0.885	524.44	1.00×
	FasterCache	0.718	0.877	0.956	0.987	0.884	247.14	2.12×
	EasyCache	0.709	**0.880**	0.955	**0.987**	0.883	181.03	2.90×
	DiCache	**0.719**	0.876	**0.957**	0.987	**0.885**	257.32	2.04×
	WorldCache	0.717	0.876	0.956	0.986	0.884	259.49	2.02×
	MedCache	0.717	0.876	0.953	0.985	0.883	**138.97**	**3.77**×
Cosmos-Predict-2.5 14B	Baseline	0.677	0.865	0.937	0.980	0.865	207.26	1.00×
	FasterCache	0.474	0.512	0.722	0.886	0.649	130.27	1.59×
	EasyCache	0.671	0.870	0.944	0.981	0.867	102.85	2.02×
	DiCache	0.671	0.860	0.931	0.980	0.861	162.43	1.28×
	WorldCache	0.655	0.825	0.911	0.980	0.843	91.74	2.26×
	MedCache	**0.678**	**0.872**	**0.945**	**0.984**	**0.870**	**76.80**	**2.70**×

from 0.767 to 0.837 while reducing latency from 185.00 s to 95.00 s. The same trend holds at larger scale on Cosmos-Predict-2.5 14B, where MedCache attains the best accelerated Total (0.895) and the lowest latency (331.40 s). On WAN2.1 14B, by contrast, MedCache gives the largest speedup ($3.62\times$) but not the best Total, indicating a more aggressive speed-quality operating point.

Image2World. In Table 2, MedCache consistently gives the lowest latency and highest speedup. It also gives the best accelerated Total on the Cosmos-Predict and Cosmos-H-Surgical blocks. It improves over EasyCache in both Total and latency on Cosmos-Predict-2.5 2B (0.788→0.795, 43.59 s→22.91 s), Cosmos-H-Surgical 2B (0.847→0.850, 31.45 s→21.05 s), and Cosmos-Predict-2.5 14B (0.867→0.870, 102.85 s→76.80 s). On WAN2.1, it trades a small Total-score reduction for a larger runtime gain by achieving the lowest latency (138.97 s) and highest speedup ($3.77\times$). Figure 4 shows a representative rollout: WorldCache introduces local artifacts at the tool-tissue interface, while MedCache stays close to the uncached baseline.

Overall trend. Across both settings, MedCache is strongest on Cosmos-Predict and Cosmos-H-Surgical, and especially in Image2World setting, where localized structure is more strongly constrained. This is consistent with the method design: spatially selective reuse is most beneficial when the underlying dynamics are structured and spatially concentrated. On broader generic checkpoints such as WAN2.1, MedCache still provides strong acceleration, but the gain is concentrated more in runtime than in aggregate quality. On WAN2.1, MedCache transfers the speedup rather than the quality gain: it is much faster than WorldCache ($3.62\times$ vs. $1.18\times$ in Text2World, $3.77\times$ vs. $2.02\times$ in Image2World) at a slightly lower Total (by 0.007 and 0.001).

Table 3: **Component removal** on Cosmos-H-Surgical Image2World. Skip denotes the fraction of conditional-branch passes that reuse the cached residual.

Config	Wall (s)	Speedup	Med-VBench ↑					
			Skip	IQ	SC	BC	MS	Total
Vanilla (uncached)	55.7	1.00×	-	0.610	0.866	0.946	0.952	0.844
A: EasyCache ($\tau = 0.05, W_0 = 5$)	40.1	1.39×	-	0.594	0.865	0.943	0.959	0.840
B: + risk weighted drift	23.5	2.37×	61.1%	0.593	0.862	0.943	0.961	0.840
C: + domain prior	22.1	2.52×	64.0%	0.593	0.865	0.942	0.960	0.840
D: + lazy consistency guard	22.1	2.52×	64.0%	0.612	0.867	0.946	0.960	0.846

Table 4: **Configuration-level efficiency** on Text2World.

Config	Speedup	Skip	Med-VBench ↑				
			IQ	SC	BC	MS	Total
Vanilla (uncached)	1.00×	-	0.545	0.799	0.899	0.933	0.794
B: MedCache, no prior	2.15×	57.9%	0.519	0.793	0.903	0.940	0.789
C: MedCache + prior	2.29×	60.8%	0.521	0.795	0.905	0.939	0.790

Ablations. Tables 3-5 report three ablations, and Table 6 shows a transfer test. First, component removal (Table 3) shows that the main speedup comes from the risk-weighted drift term i.e., replacing the scalar EasyCache statistic increases speedup from $1.39\times$ to $2.37\times$ with essentially unchanged Total. The prompt prior makes a smaller operating-point adjustment, increasing the speedup from $2.37\times$ to $2.52\times$ while keeping the total unchanged. The lazy guard improves fidelity at the same speedup and skip rate, increasing IQ from 0.593 to 0.612. Second, the prompt-derived prior acts as a

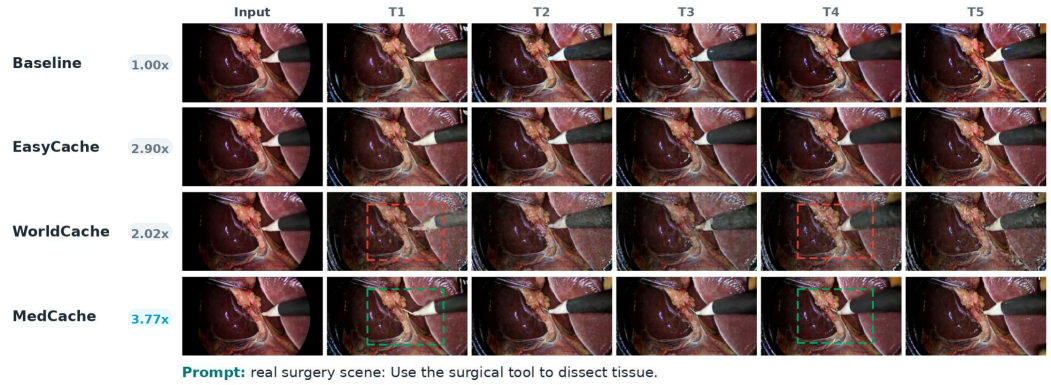


Figure 4: **Qualitative Image2World comparison on a surgical prompt.** Given the same input image and prompt, WorldCache introduces local artifacts around the active tool-tissue interface, while MedCache preserves the corresponding region more coherently over time. The first column is the conditioning image; the remaining columns are generated rollout frames.

content-dependent control on reuse rather than as a fixed global threshold. In Table 4, enabling the prior raises the average skip rate and speedup at essentially unchanged quality. The per-domain analysis (Table 9, Appendix B.1) shows that this change is not uniform: the shift is largest for intervention prompts, smaller for diagnostic-imaging and clinical-exam prompts, zero for rehabilitation, and slightly reversed for animation/education content. Third, risk-map composition (Table 5) has limited effect on runtime, with all variants lying between $2.23\times$ and $2.29\times$, but the full S+F+M map gives the highest Total (0.844), indicating that the high-frequency term mainly improves fidelity. Finally, we evaluate MedCache on HunyuanVideo [Kong et al., 2024] with VBench [Huang et al., 2024] as a transfer test to assess whether the same latent-space decision rule remains effective beyond the medical setting. Here, MedCache achieves $3.44\times$ speedup while retaining baseline quality. We also provide further ablations for different parameters along with the speed-quality tradeoff in Figure 6 and Figure 7 of the appendix, respectively.

Table 5: **Risk map composition** ablation on Cosmos-H-Surgical 2B Image2World.

					Med-VBench ↑				
Config	Speedup	Skip	w_S	(w_M, w_F)	IQ	SC	BC	MS	Total
Vanilla (uncached)	1.00×	-	-	–	0.610	0.866	0.946	0.952	0.844
M-only (motion)	2.28×	61.4%	0.0	(1.25, 0.0)	0.593	0.864	0.943	0.960	0.840
S-only (structural)	2.23×	60.0%	1.0	(0.0, 0.0)	0.592	0.866	0.944	0.959	0.840
S+M (no HF)	**2.29**×	60.0%	1.0	(1.25, 0.0)	0.593	0.866	0.944	0.959	0.841
S+F+M (full, ours)	2.24×	60.0%	1.0	(1.25, 0.75)	0.601	0.867	0.946	0.960	0.844

Table 6: **Generalization to a generic benchmark** on HunyuanVideo evaluated with VBench.

Method	IQ	SC	BC	MS	Total	Latency	SpeedUp
Baseline	0.661	0.960	0.971	0.993	0.896	984.04	1.00×
EasyCache	0.662	0.962	0.972	0.993	0.897	485.68	2.03×
WorldCache	0.637	0.959	0.971	0.993	0.890	334.78	2.93×
MedCache	**0.664**	0.961	**0.974**	**0.995**	**0.899**	**286.54**	**3.44**×

6 Conclusion

We presented **MedCache**, a training-free caching method for video generation. MedCache is motivated by the observation that reuse errors in medical video are often spatially localized, while existing methods either ignore this structure or recover it via an expensive decision path. By deriving a spatial reuse signal directly from the latent, MedCache makes the skip decision more selective while retaining the simplicity of lightweight caching. Across Text2World and Image2World, it improves the speed-quality trade-off, with the clearest gains on Cosmos-Predict and Cosmos-H-Surgical models and in image-conditioned settings. The operating point is stable across 40 perturbed configurations, and inside annotated instrument regions MedCache stays substantially closer to the uncached output than the probe-based WorldCache.

Acknowledgments and Disclosure of Funding

This work was supported by Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) through internal university funds. No external funding was received for this work. All experiments were conducted on MBZUAI computing infrastructure. We thank the developers of the Cosmos,

WAN2.1, and HunyuanVideo models for releasing their checkpoints, and the creators of CholecT50, SurgΣ, and MedGen/Med-VBench for making their data and benchmarks publicly available.

Competing interests. The authors declare no competing interests.

References

Niket Agarwal et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Arslan Ali et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

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

Guanjie Chen, Xinyu Zhao, Yucheng Zhou, Xiaoye Qu, Tianlong Chen, and Yu Cheng. Towards stabilized and efficient diffusion transformers through long-skip-connections with spectral constraints. In *ICCV*, 2025.

Joseph Cho, Samuel Schmidgall, Cyril Zakka, Mrudang Mathur, Dhamanpreet Kaur, Rohan Shad, and William Hiesinger. Surgen: Text-guided diffusion model for surgical video generation. *arXiv preprint arXiv:2408.14028*, 2024.

Yufan He et al. Cosmos-h-surgical: Learning surgical robot policies from videos via world modeling. *arXiv preprint arXiv:2512.23162*, 2025.

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

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Chenxin Li et al. Endora: Video generation models as endoscopy simulators. In *MICCAI*, 2024.

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 *CVPR*, 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.

Zhengyao Lv, 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 *ICLR*, 2025.

Wenping Ma, Xiaoting Yang, Licheng Jiao, Lingling Li, Xu Liu, Fang Liu, Puhua Chen, Yuting Yang, Mengru Ma, Long Sun, et al. Video diffusion generation: comprehensive review and open problems. *Artificial Intelligence Review*, 2025.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *CVPR*, 2024.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023.

Taehong Moon, Moonseok Choi, EungGu Yun, Jongmin Yoon, Gayoung Lee, Jaewoong Cho, and Juho Lee. A simple early exiting framework for accelerated sampling in diffusion models. In *ICML*, 2024.

Umair Nawaz, Ahmed Heakl, Ufaq Khan, Abdelrahman Shaker, Salman Khan, and Fahad Shahbaz Khan. Worldcache: Content-aware caching for accelerated video world models. In *European Conference on Computer Vision*, 2026.

Chinedu Innocent Nwoye and Nicolas Padoy. Data splits and metrics for method benchmarking on surgical action triplet datasets. *arXiv preprint arXiv:2204.05235*, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Xurui Peng, Chenqian Yan, Hong Liu, Rui Ma, Fangmin Chen, Xing Wang, Zhihua Wu, Songwei Liu, and Mingbao Lin. ERTACache: Error rectification and timesteps adjustment for efficient diffusion. In *ICLR*, 2026.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Mingzhu Shen, Pengtao Chen, Peng Ye, Guoxuan Xia, Tao Chen, Christos-Savvas Bouganis, and Yiren Zhao. Md-dit: Step-aware mixture-of-depths for efficient diffusion transformers. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Rongsheng Wang, Junying Chen, Ke Ji, Zhenyang Cai, Shunian Chen, Yunjin Yang, and Benyou Wang. Medgen: Unlocking medical video generation by scaling granularly-annotated medical videos. *arXiv preprint arXiv:2507.05675*, 2025.

Zhenbin Wang, Lei Zhang, Lituan Wang, Minjuan Zhu, and Zhenwei Zhang. Optical flow representation alignment mamba diffusion model for medical video generation. *arXiv preprint arXiv:2411.01647*, 2024.

Zhitao Zeng et al. SurgΣ: A spectrum of large-scale multimodal data and foundation models for surgical intelligence. *arXiv preprint arXiv:2603.16822*, 2026.

Xin Zhou et al. Less is enough: Training-free video diffusion acceleration via runtime-adaptive caching. *arXiv preprint arXiv:2507.02860*, 2025.

Appendix for MedCache

A Implementation and Experimental Details

A.1 Base Models

We evaluate MedCache on three open rectified-flow video generators: WAN2.1 [Wan et al., 2025], Cosmos-Predict-2.5 [Ali et al., 2025], and Cosmos-H-Surgical [He et al., 2025]. These models span three regimes relevant to this work: a strong open general-domain generator, a large open world-model family, and a medically adapted checkpoint in which localized procedural fidelity is especially important. Experiments are organized by model scale where corresponding checkpoints are publicly available. Since MedCache is training-free and does not modify model weights, all methods are evaluated directly on the publicly released checkpoints. Moreover, for generalization, we also conducted transfer evaluation on the HunyuanVideo [Kong et al., 2024] model using VBench [Huang et al., 2024] as shown in Table 6.

A.2 Benchmarks

We study two generation settings: **Text2World** and **Image2World**.

Text2World. In Text2World, the model generates a temporally coherent video sequence from a text prompt obtained from the test set of [Wang et al., 2025]. We evaluated it on its 200-prompt benchmark, which was used to evaluate the Text2World settings. The benchmark spans intervention, diagnostic imaging, clinical examination, rehabilitation, and instructional content, and each sample is stored as a JSON record containing the prompt, guidance scale, seed, and requested frame count.

Image2World. In Image2World, generation is conditioned on an input image. For this setting, we used 300 samples from the conditional-generation portion of SurgSigma [Zeng et al., 2026], selected according to the conditional video generation subset, which is based on CholecT50 [Nwoye and Padoy, 2022].

A.3 Baselines

We compare MedCache with five inference configurations. The first is the uncached **Baseline**, which runs the original sampler without reuse. The remaining four are training-free acceleration baselines: **FasterCache** [Lv et al., 2025], **DiCache** [Bu et al., 2026], **EasyCache** [Zhou et al., 2025], and **WorldCache** [Nawaz et al., 2026]. These baselines span a representative range of decision rules, from lightweight global-drift heuristics to richer probe-based pipelines.

A.4 Evaluation Metrics

For each generated video, we report four quality dimensions: *Image Quality (IQ)*, *Subject Consistency (SC)*, *Background Consistency (BC)*, and *Motion Smoothness (MS)*, following [Wang et al., 2025]. These metrics capture complementary aspects of generation quality, ranging from frame-level appearance to temporal coherence. We summarize them with a *Total* score computed as their arithmetic mean. To evaluate efficiency, we report average *Latency* in seconds and *SpeedUp* relative to the uncached baseline. Latency is measured only for the core video-generation stage, excluding all pre-processing and post-processing overhead.

A.5 Sampling Configuration, Hardware, and Latency Measurement

All methods are evaluated under matched sampling and hardware settings within each model block. For the Cosmos-Predict-2.5 and Cosmos-H-Surgical experiments, including the component-removal and risk-map-composition ablations, we use $T = 35$ denoising steps with classifier-free guidance scale 7.0 on a single NVIDIA H200 GPU. These runs follow the same timing protocol as the original supplementary setup.

For WAN2.1 experiments, we use the public WAN2.1 sampling configurations used throughout the main tables. WAN2.1 Text2World uses 50 denoising steps, guidance scale 5.0, and resolution

Table 7: **Hyperparameter sensitivity on Cosmos-Predict-2.5 2B Text2World.** Each row sweeps one hyperparameter around the frozen configuration. Arrows give values from the lowest to the highest setting. Total reports the range over the sweep; PSNR/SSIM are measured against the same-seed uncached output. The L_8 joint-variation study uses $\tau \in \{0.0855, 0.1045\}$, $\beta \in \{4.5, 7.5\}$, $\kappa \in \{1.05, 1.15\}$, $\eta \in \{0.15, 0.25\}$, $K \in \{9\%, 15\%\}$, $\rho \in \{1.35, 1.65\}$, and $W_o \in \{5, 7\}$.

Sweep	Range	Reuse (%)	Speedup	Total	PSNR/SSIM
τ	$0.075 \rightarrow 0.115$	31.14	$2.408 \rightarrow 2.762\times$	0.8548-0.8592	$26.724/0.9082 \rightarrow 22.366/0.8318$
κ	$1.00 \rightarrow 1.20$	34.43	$2.472 \rightarrow 2.766\times$	0.8539-0.8616	$25.657/0.8894 \rightarrow 22.873/0.8393$
β	$0 \rightarrow 9$	41.43	$2.628 \rightarrow 2.618\times$	0.8557-0.8570	essentially flat

832×480. WAN2.1 Image2World uses 40 denoising steps, guidance scale 7.0, and resolution 832×480. WAN2.1 14B experiments are run with distributed model-parallel inference on 8 AMD Instinct MI210 GPUs. The WAN2.1 14B sensitivity analysis in Sec. B.3 is also conducted on the same AMD MI210 8-GPU setting. Within each WAN2.1 block, all methods use the same hardware, prompt or image inputs, sampler settings, and output resolution.

Under classifier-free guidance, all cache bookkeeping is performed over internal CFG passes rather than visible denoising steps. Thus, a run with T visible denoising steps has $2T$ internal cache-decision passes, one for the conditional branch and one for the unconditional branch.

A.6 Hyperparameters and Selection Protocol

MedCache hyperparameters were selected once on a held-out Text2World development split and then frozen for all reported experiments. The spatial risk map uses weights $(\alpha_S, \alpha_F, \alpha_M) = (1.00, 0.75, 1.25)$, risk amplification $\beta = 6.0$, over-threshold factor $\kappa = 1.10$, lazy-guard fraction $\eta = 0.20$, and top-risk support $K = 12\%$. The domain-prior threshold scale $\sigma \in \{0.70, 0.85, 0.95, 1.00, 1.25\}$ is assigned by keyword matching at inference time and is not optimized per evaluation sample. No hyperparameter is adjusted after the development split is closed.

To check that the method is not implicitly tuned to the development distribution, we also report out-of-distribution evaluations with the same frozen configuration. First, all Image2World results use the Text2World-selected hyperparameters without retuning under a different conditioning modality. Second, the WAN2.1 14B Image2World sensitivity study in Sec. B.3 uses the same default configuration and varies only one scalar hyperparameter at a time.

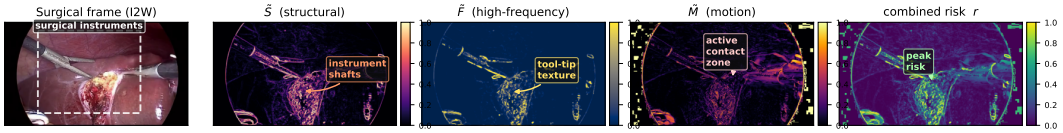


Figure 5: **From when to skip to where reuse is risky.** From left to right: generated frame, structural-boundary energy $\tilde{S}$, local high-frequency energy $\tilde{F}$, relative-motion energy $\tilde{M}$, and the combined risk map r. Existing training-free caches decide *when* to skip from a global drift scalar. In this example, the highest responses fall near the instrument-tissue interaction. The map ranks locations by the risk of applying a reused residual and is not a semantic or instrument segmentation. This motivates MedCache to replace a global drift average with a latent-derived spatial reuse signal.

A.7 Sensitivity and Interaction Analysis

We also examine whether the gains of MedCache depend on precise tuning, complementing the WAN2.1 sweeps of Sec. B.3 with a different backbone and task. All experiments use a fixed, domain-balanced bank of 20 Cosmos-Predict-2.5 2B Text2World prompts that was not used for selection. The prompts, seeds, the sampling schedule, resolution, hardware, and all non-varied settings are held fixed.

Single-hyperparameter sweeps. Table 7 sweeps τ, κ, and β around the frozen configuration. Increasing τ or κ smoothly trades fidelity for speed: reuse and speedup rise, while PSNR/SSIM relative to the uncached output fall, as expected when more computation is skipped. Perceptual

Table 8: **Joint perturbation (L_8) on Cosmos-Predict-2.5 2B Text2World.** Seven hyperparameters are varied simultaneously at two levels each: $\tau \in \{0.0855, 0.1045\}$, $\kappa \in \{1.05, 1.15\}$, $\beta \in \{4.5, 7.5\}$, $\eta \in \{0.15, 0.25\}$, $K \in \{9\%, 15\%\}$, $\rho \in \{1.35, 1.65\}$, and $W_o \in \{5, 7\}$. Runs are grouped by their shared (τ, κ). Each entry gives the range over the two runs in the group, which differ in the remaining hyperparameters.

Runs	τ	κ	Reuse (%)	Speedup	Total	PSNR	SSIM
J1-J2	0.0855	1.05	33.29-33.57	2.452-2.458×	0.8595-0.8600	26.097-26.182	0.8971-0.8986
J3-J4	0.0855	1.15	41.29-41.86	2.628-2.642×	0.8535-0.8543	24.255-24.390	0.8659-0.8682
J5-J6	0.1045	1.15	46.71-47.57	2.773-2.798×	0.8507-0.8514	22.198-22.618	0.8224-0.8337
J7-J8	0.1045	1.05	40.29-40.57	2.603-2.610×	0.8555-0.8558	24.022-24.124	0.8642-0.8666

quality nonetheless remains within a narrow band throughout. The two hyperparameters produce the same qualitative trend because both act on a single effective reuse boundary rather than serving as independent corrective mechanisms. In contrast, the response to β is essentially flat in both reuse and quality, indicating that its default is not sharply tuned.

Joint variation. One-at-a-time sweeps cannot reveal interactions, so we additionally vary seven hyperparameters simultaneously using an eight-run L_8 orthogonal design, with each set to a low and a high level around its default (Table 8). The outcome is organized almost entirely by the reuse boundary: runs that share the same (τ, κ) are nearly indistinguishable despite changes in the remaining hyperparameters. Conservative boundaries remain closest to the uncached output, aggressive boundaries increase reuse and speed, and no interaction collapse is observed. Across all 40 evaluated configurations, including contrasts of the domain prior, every configuration retains at least a $2.4\times$ speedup and stays within 0.01 Total of the frozen setting.

Seed robustness. Repeating the evaluation with three random seeds on a fixed six-context subset changes speedup by less than 1%.

Summary. The speed-quality behavior of MedCache is governed by a low-dimensional reuse boundary (τ, κ), whereas the remaining hyperparameters are stable over the tested ranges. Consequently, a single frozen configuration transfers across backbones and tasks without per-setting tuning.

B Additional Ablations and Analyses

This section provides additional discussion for the ablations.

B.1 Domain Prior

The purpose of the domain prior is to modulate reuse aggressiveness according to prompt content. Under the rule $\tau_{\text{eff}} = \sigma\tau$, values $\sigma < 1$ reduce the effective skip threshold and therefore make reuse more conservative, while values $\sigma > 1$ make reuse more permissive. Table 4 reports the aggregate configuration-level effect, and Table 9 breaks this effect down by content class.

The per-domain view in Table 9 shows that the prior does not act as a uniform global shift, and instead, it moves different content classes to different operating points. The largest change occurs for intervention prompts, which reuse noticeably more often once the prior is applied. Diagnostic-imaging and clinical-exam prompts shift in the same direction by smaller margins, rehabilitation prompts are unaffected because their scale is neutral, and animation/education prompts move slightly in the opposite direction.

Taken together, these results support the interpretation that the prompt prior is a lightweight, content-aware control on reuse aggressiveness rather than a source of quality gains. Its role is to allocate the skip budget differently across content types, while fidelity in spatially sensitive regions remains governed by the risk-weighted drift and the lazy guard, which the prior does not modify. As Table 4 shows, the additional reuse comes without a loss in aggregate quality, and the core MedCache decision rule is unchanged.

Table 9: **Per-domain skip shift.** "NoPrior" fixes $\sigma = 1.0$ for all prompts. "WithPrior" applies the prompt-derived scale. Δskip denotes the change induced by the prior in percentage points.

Domain	σ	NoPrior skip	WithPrior skip	Δskip
Intervention	1.25	48.6%	58.1%	$\uparrow 9.5$ pp
Diagnostic imaging	0.85	59.0%	62.9%	$\uparrow 3.9$ pp
Clinical exam	0.95	60.0%	61.9%	$\uparrow 1.9$ pp
Rehabilitation	1.00	59.0%	59.0%	≈ 0 pp
Animation/education	1.25	62.9%	61.9%	$\downarrow 1.0$ pp
All (mean)	N/A	57.9%	60.8%	$\uparrow 2.9$ pp

B.2 Risk Map Composition

Table 5 ablates the composition of the latent risk map on Cosmos-H-Surgical 2B Image2World. A first observation is that speedup is relatively insensitive to the exact composition: all tested variants lie in a narrow range of $2.23\times$ to $2.29\times$, and the skip rate varies only between 60.0% and 61.4%. This indicates that the detailed choice of risk-map components is not the main determinant of runtime.

The quality metrics separate the variants more clearly. Among the reduced maps, the S+M variant gives the highest speedup at $2.29\times$, but its Total is 0.841. The full S+F+M map gives a slightly lower speedup at $2.24\times$, but increases Total to 0.844, matching the uncached baseline and exceeding all reduced variants. This difference is small in absolute terms, but it is consistent across the reported metrics.

These results suggest a clear functional split among the components. Motion provides the strongest single signal for runtime reduction, structure complements motion by preserving sensitivity to spatial boundaries, and the high-frequency term improves fidelity without increasing speed. This is exactly the role intended for the full MedCache risk map: not to maximize skip rate at all costs, but to preserve localized detail while maintaining strong acceleration.

B.3 Sensitivity to τ, β, and κ

We report a local sensitivity analysis of the three main scalar MedCache hyperparameters on a fixed 20-prompt WAN2.1 14B Image2World subset. The defaults ($\tau = 0.095, \beta = 6.0, \kappa = 1.10$) were selected once on a held-out Text2World development split and then frozen for all reported experiments. Accordingly, the goal of this study is not to retune MedCache for this subset, but to assess whether the default operating point lies in a predictable and non-brittle region of the speed-quality tradeoff. Since the main paper reports WAN2.1 14B Image2World on the full 300-sample benchmark, the absolute values in this subset analysis are not expected to match Table 2 exactly.

Figure 6 shows direct sweeps of τ, β, and κ. Increasing τ produces the expected shift toward more aggressive reuse and higher speedup, with a corresponding reduction in Total. By contrast, the β sweep lies in a relatively flat region, indicating that MedCache is not highly sensitive to moderate changes in risk amplification around the default. The κ sweep traces the clearest speed-quality tradeoff frontier: larger κ yields progressively higher speedup at the cost of lower Total. Taken together, these curves indicate that the default configuration is not an isolated tuning point. In particular, $\beta = 6.0$ lies in a broad stable region, while $\kappa = 1.10$ provides a balanced operating point between conservative and aggressive reuse. The default $\tau = 0.095$ corresponds to the more aggressive regime used in the main experiments; smaller τ improves Total on this subset at the cost of noticeably lower speedup.

Figure 7 summarizes the same sweeps in speed-quality space. The Pareto view shows that the κ sweep spans the widest frontier, whereas the β sweep remains tightly clustered around the default point. Overall, these results support the claim that MedCache behaves predictably under moderate hyperparameter variation and that its dominant speed-quality tradeoff is controlled primarily by κ.

B.4 Dataset-Level Drift Concentration

For a dataset-level measure, at each eligible conditional decision, we select the highest-risk positions and measure the fraction of the actual latent drift mass within them, averaging decisions within videos and then across videos; confidence intervals are estimated using 20,000 video-bootstrap replicates. Table 10 shows that on the full 300-video Cosmos-H-Surgical 2B Image2World evaluation (8,100

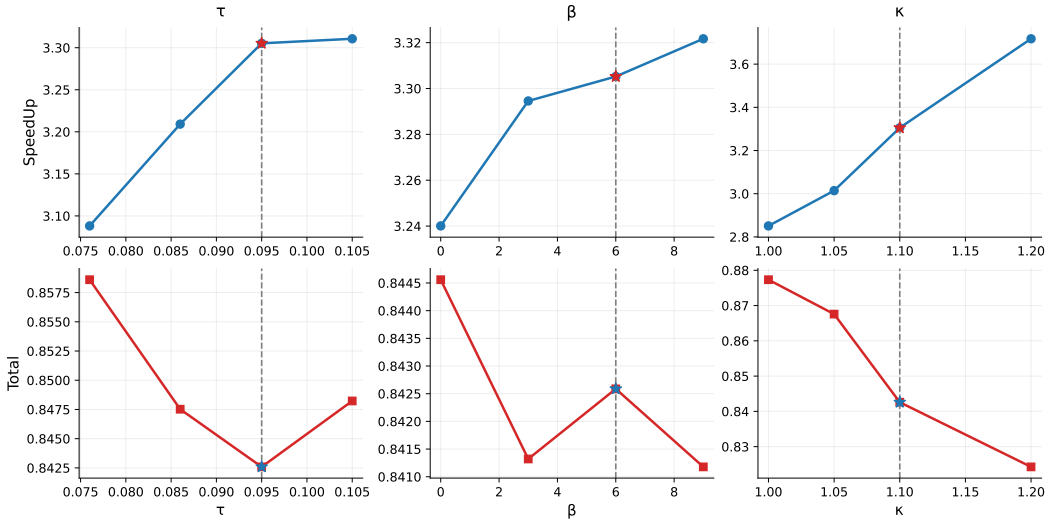


Figure 6: **Sensitivity of MedCache to τ, β, and κ on a fixed 20-prompt WAN2.1 14B Image2World subset.** The dashed line marks the default value used in the main experiments. The defaults were selected on a held-out Text2World development split and then frozen, so this figure should be interpreted as a local robustness analysis rather than a subset-specific retuning study.

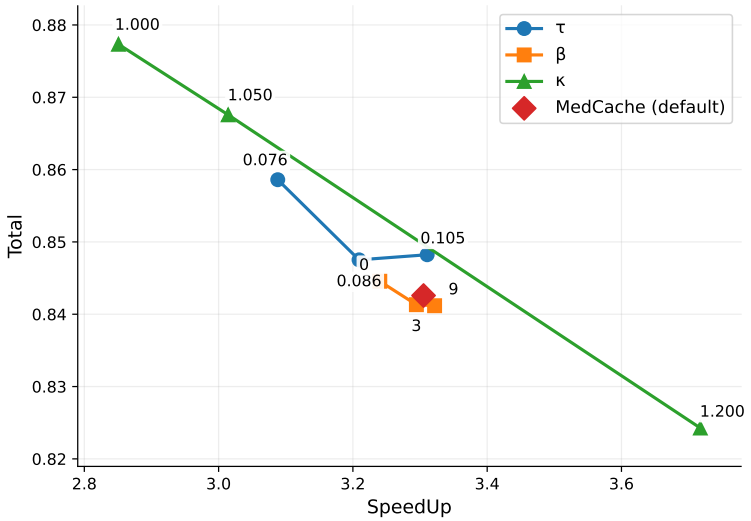


Figure 7: **Speed-quality tradeoff induced by sweeping τ, β, and κ.** Each non-default point is labeled only by its scalar value, while the red diamond denotes MedCache (default). The plot highlights that κ acts as the principal speed-quality tradeoff knob, whereas β lies in a comparatively flat region around the default.

decisions), the risk-ranked support holds more than its uniform share of drift at every support size. On Cosmos-Predict-2.5 2B Text2World (20 videos), the risk-ranked top 10% holds 10.83% [10.65, 11.04]. The map therefore consistently ranks locations that carry disproportionate latent change, on both a general and a medical checkpoint. The medical motivation does not rest on concentration alone but on the asymmetric cost of local error, which Appendix B.7 measures directly.

Table 10: **Drift mass inside the risk-ranked top-q support**, Cosmos-H-Surgical 2B Image2World, 300 videos.

Support q	Uniform	Risk-ranked (95% CI)
1%	1.00%	1.28% [1.26, 1.30]
5%	5.00%	6.21% [6.13, 6.28]
10%	10.00%	12.12% [11.99, 12.25]
12%	12.00%	14.42% [14.27, 14.57]

B.5 Prompt Prior: Coverage and Threshold Scale

Keyword matching is the least general component of MedCache and is not the basis of the spatial or medical-motivation claim. Table 11 explores the reduced three-group vocabulary on every prompt bank. No generic control prompt (animals, architecture, food, people, lifestyle, plants, scenery, vehicles) activates the prior. Specialties outside the vocabulary, such as dermatology and dermoscopy (and ultrasound under the three-group vocabulary), fall back to $\sigma = 1$ and leave the latent controller unchanged. We therefore do not claim semantic generalization to arbitrary specialties.

Table 12 isolates the threshold scale on ten separate Cosmos-H-Surgical 2B Image2World samples, with the model, inputs, denoising schedule, cache configuration, and hardware fixed and the guard disabled. Speedup rises monotonically with σ, from $2.015\times$ at $\sigma = 0.70$ to $2.715\times$ at $\sigma = 1.70$, while Total stays between 0.8454 and 0.8499 (uncached: 0.8517).

Table 11: **Prompts with a non-neutral prior scale** ($\sigma \neq 1$) under the submitted three-group vocabulary.

Prompt bank	Non-neutral / total
Medical Text2World (Med-VBench)	171/200 (85.5%)
Medical Image2World	23/23 (100%)
CholecT50 surgical Image2World	300/300 (100%)
Generic controls	0/16 (0%)

Table 12: **Fixed-σ sweep** on 10 Cosmos-H-Surgical 2B Image2World samples (guard disabled). Latency is per sample, with 95% CI.

σ	Latency (s)	Speedup	Skip	Total
Uncached	52.4 [51.9, 53.1]	$1.000\times$	-	0.8517
0.70	26.0 [24.8, 27.0]	$2.015\times$	54.86%	0.8473
1.00	22.4 [21.6, 23.3]	$2.339\times$	62.29%	0.8454
1.25	21.0 [20.2, 21.9]	$2.495\times$	65.43%	0.8499
1.70	19.3 [18.8, 20.0]	$2.715\times$	68.86%	0.8459

B.6 Risk-Map Staleness

MedCache does not assume stationarity. The map is refreshed after every full conditional computation and reused only during the following skip streak. For analysis, we computed the map that would be obtained at every eligible decision and compared it with the cached map. This diagnostic never entered the controller, did not affect any skip decision or generated video, and its cost is excluded from all latency measurements. Table 14 summarizes which reuse failure each cue targets (Sec. 4.1), whereas in Table 13, map age is the number of conditional decisions since the last refresh; MedCache imposes no age limit, so the maxima are observed values. Cached soft maps stay highly similar to fresh maps, while the lower top-12% IoU shows that locations near the ranking boundary can change during a skip streak. In a refresh-delay stress test on the 20-video Text2World dataset, refreshing only after every second or fourth full computation lowers agreement, as expected, but changes Total by less than 0.0002. We retain refresh after every full computation as the most conservative policy. Finally, the soft multiplier $(1 + \beta r)$ gives every location a base weight of one, so a newly changing

region outside the old high-risk support still contributes to the accumulated drift, and a rejected reuse triggers a full computation that refreshes the map.

Table 13: **Risk-map staleness**: cached versus freshly computed map. Top: default refresh on two checkpoints. Bottom: refresh after every N full computations on the Text2World bank.

Setting	Decisions	Age (mean / 95th / max)	Map cosine	Top-12% IoU	Total
Cosmos-Predict-2.5 2B T2W (20 videos)	540	2.08 / 5 / 11	0.9724 [0.9700, 0.9747]	0.5677 [0.5476, 0.5882]	0.8570
Cosmos-H-Surgical 2B I2W (300 videos)	8,100	3.22 / 8 / 16	0.9906 [0.9904, 0.9909]	0.7107 [0.7056, 0.7159]	-
$N = 1$ (default)	540	2.08 / 5 / 11	0.9724	0.5677	0.8570
$N = 2$, stress test	-	3.16 / 8 / 14	0.9688	0.5242	0.8570
$N = 4$, stress test	-	5.07 / 12 / 20	0.9587	0.4593	0.8568

Table 14: What each cue targets.

Cue	Signal	Reuse failure targeted	Medical example
Structural S	First-order spatial gradients	Displacement or blurring of persistent interfaces	Organ boundaries; tool-tissue contours
High-frequency F	Local variation of the Laplacian response	Loss of thin or rapidly varying detail underrepresented by 1st-order edges	Fine sutures; narrow interfaces; endoscopic texture
Relative motion M	Self-normalized change from the preceding same-branch latent	Local change within an otherwise stable scene	Instrument motion; tissue deformation; moving fluid

B.7 Cache Fidelity in Annotated Instrument Regions

The medical motivation for MedCache is that the consequences of cache error are spatially asymmetric. To measure this directly, we use 260 CholecT50 Image2World cases with non-empty SurgSigma instrument masks and compare each cached video with the uncached video generated from the same prompt, image, and seed. The instrument mask of the conditioning frame defines a fixed region that covers only 7.42% of the frame on average, and its one-pixel morphological gradient defines the boundary band. We report the mean absolute RGB error (0-255) over the first generated second. Table 15 shows that MedCache reduces this error relative to WorldCache by 57.9% inside the instrument region and by 55.7% along its boundary. This is a localized cache-fidelity diagnostic, not a claim of segmentation or clinical correctness.

Table 15: **Deviation from the same-seed uncached output** (MAE, 0-255, lower is better) on 260 Cosmos-H-Surgical 2B Image2World cases over the first generated second. Absolute reduction with paired 95% CI.

Region	WorldCache	MedCache	Reduction	Absolute reduction
Instrument region (7.42%)	32.534	**13.699**	57.9%	18.835 [17.247, 20.540]
Instrument boundary	31.005	**13.723**	55.7%	17.283 [15.954, 18.671]

B.8 Transfer to a Medically Adapted Checkpoint

To examine generalization beyond the reported checkpoints, we also compare the generic WAN2.1-1.3B [Wan et al., 2025] checkpoint with MedGen-1.3B [Wang et al., 2025], a medically adapted model initialized from the same architecture. The architecture, sampler, prompts, resolution, denoising steps, random seed, MedCache configuration, and threshold are held fixed across the two checkpoints, and no checkpoint-specific tuning is performed. Table 16 reports textual prompts at a quality-oriented threshold ($\tau = 0.020$) and at a more aggressive shared threshold ($\tau = 0.040$). MedCache achieves nearly identical acceleration on the generic and the medically adapted checkpoint, and at $\tau = 0.020$ same-seed similarity to the uncached output remains high. At $\tau = 0.020$, WAN2.1 Total is effectively unchanged, while MedGen shows a small observed reduction whose paired 95% confidence interval includes zero, whereas at $\tau = 0.040$, MedGen Total falls by 0.010.

C Additional Motivation for MedCache

This section provides additional empirical evidence for the design constraints summarized in Sec. 3.1.

A scalar accumulator cannot localize where the error is growing. Once latent change is reduced to a single scalar, the decision rule no longer retains information about *where* that change occurred.

Table 16: **Shared-threshold transfer** between WAN2.1-1.3B and MedGen-1.3B (17 held-out prompts). SSIM against the same-seed uncached output; ΔTotal is the paired change in VBench Total with 95% CI.

Model	Setting	Speedup	Skip	SSIM	Total	ΔTotal
WAN2.1-1.3B	Uncached	1.00×	-	-	0.89315	-
	$\tau = 0.020$	1.48×	36.5%	0.954	0.89335	$+0.00019$ $[-0.00074, 0.00115]$
	$\tau = 0.040$	2.05×	59.3%	0.875	0.89242	-0.00073 $[-0.00580, 0.00437]$
MedGen-1.3B	Uncached	1.00×	-	-	0.88449	-
	$\tau = 0.020$	1.47×	34.6%	0.960	0.88280	-0.00169 $[-0.00363, 0.00001]$
	$\tau = 0.040$	2.02×	57.9%	0.895	0.87444	-0.01005 $[-0.01505, -0.00563]$

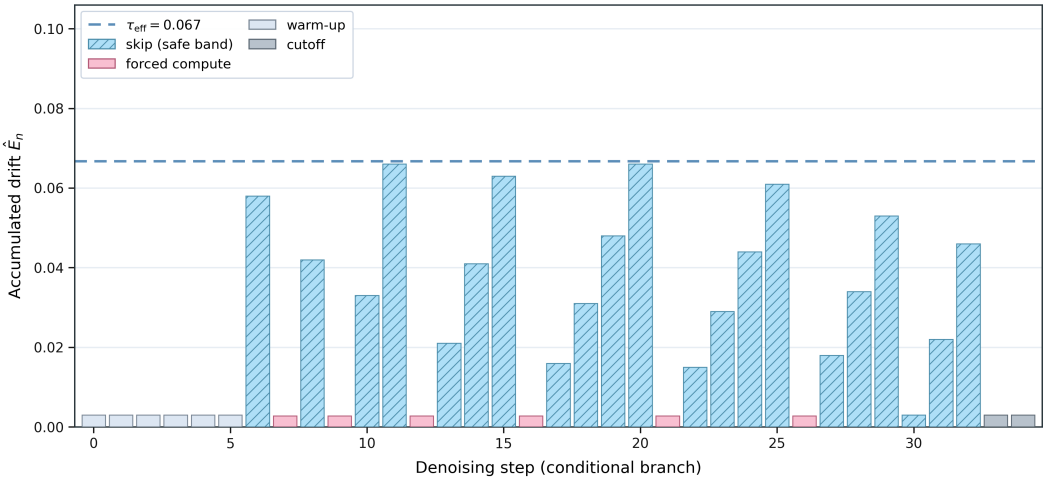


Figure 8: **Empirical MedCache decision trajectory on a representative Image2World generation.** The figure shows the accumulated drift $\hat{E}_n$ on the conditional branch during a real Cosmos-H-Surgical 2B Image2World rollout. After warm-up, MedCache enters a stable skip/compute alternation, with 21 of 35 visible denoising steps skipped and no guard recomputes on this sample.

As a result, the same accumulated value $\hat{E}_n$ can correspond either to broad, low-value variation or to a small but clinically important failure region. The left panel of Figure 2 in the main text shows the characteristic rise/reset pattern of scalar accumulated drift, while the right panel shows that clinically relevant drift mass is spatially concentrated rather than uniformly distributed over the frame. Figure 8 complements this observation by showing how MedCache behaves after replacing the uniform scalar statistic with a spatially weighted one: once the warm-up region is passed, the method quickly reaches a stable reuse regime on the representative sample shown.

Probe cost matters even when skip rate is high. WorldCache [Nawaz et al., 2026] improves on scalar drift by evaluating the first d transformer blocks at every internal pass and then using the resulting probe features, together with online system identification and optical-flow-based alignment, to determine whether cached states remain valid. This yields a more informative signal, but the probe cost is unconditional and cannot be amortized away as the skip rate increases.

Figure 9 shows that skip rate alone is not the right measure of a caching method. WorldCache achieves a higher skip rate, but part of its speed gain is spent on the probe itself. MedCache achieves a higher net speedup with a lower skip rate because its decision path is substantially cheaper.

D Additional Qualitative Results

We provide additional qualitative comparisons in Figures 10 and 11 to complement the main qualitative example in Figure 4. All methods are evaluated under the same conditioning signal and sampling configuration for a given example. The panels show evenly spaced rollout frames, with rows corresponding to the uncached baseline and training-free cache baselines. These examples visualize

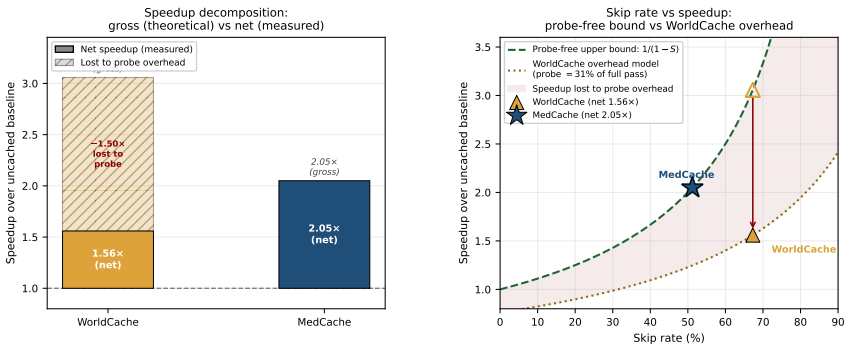


Figure 9: **Probe overhead as a speed tax.** *Left:* WorldCache's $d = 3$ block probe costs 469 ms, which is 31% of a full forward pass, so a theoretical 3.06× gross speedup collapses to 1.56× net speedup after probe cost. MedCache uses only a sub-millisecond risk map and lazy guard, so its gross and net speedups coincide at 2.05×. *Right:* MedCache lies on the probe-free upper bound $1/(1 - S)$, while WorldCache is pulled below it by fixed probe overhead.

the temporal and local structural behavior behind the quantitative speed-quality trends reported in the main tables.

Image2World surgical rollouts. Figure 10 shows additional Image2World examples from surgical scenes with active tool-tissue interaction. These cases are representative of the setting targeted by MedCache: clinically meaningful changes are concentrated around small local regions such as tool tips, tissue boundaries, bleeding regions, and retraction or manipulation sites. In these localized regions, cache-induced errors can appear as texture degradation, deformation of surgical tools, loss of tissue detail, or inconsistent motion over time. We therefore annotate only the frames where there is a clear matched local comparison: red dashed boxes highlight representative WorldCache artifacts, while green dashed boxes mark the corresponding regions in MedCache. The boxes are intended to focus attention on the same tool-tissue contact region across methods.

Text2World surgical rollouts. Figure 11 provides two additional Text2World examples using cardiac and leg surgery prompts. Unlike the Image2World cases, the differences in this example are more distributed across the generated rollout rather than concentrated in one clean local artifact. For this reason, we omit bounding boxes and show the full rollout directly.

E Algorithm Summary

Algorithm 1 summarizes the MedCache decision on the conditional CFG branch. The unconditional branch follows the same rule with its own branch-wise cached state, while sharing the same overall scheduling logic.

F Limitations and Future Work

MedCache is designed for training-free acceleration of rectified-flow video generation, and its strongest gains appear in settings where reuse risk is spatially localized. This also defines its main limitation. When the scene dynamics are globally distributed rather than concentrated in a small region, the advantage of spatially selective reuse is smaller, and the method may move to a more aggressive speed-quality operating point rather than improving both simultaneously. This behavior is visible on some general-domain checkpoints, where MedCache gives large runtime gains but smaller quality advantages than on the medical and image-conditioned settings.

A second limitation is that the domain prior is heuristic. It is lightweight and easy to apply, but it depends on prompt-side keyword matching rather than a learned or formally calibrated estimate of content sensitivity. While the ablations show that it changes reuse behavior in a content-dependent way, its design remains hand-specified and may not transfer optimally to domains with different prompt styles or weaker textual grounding.

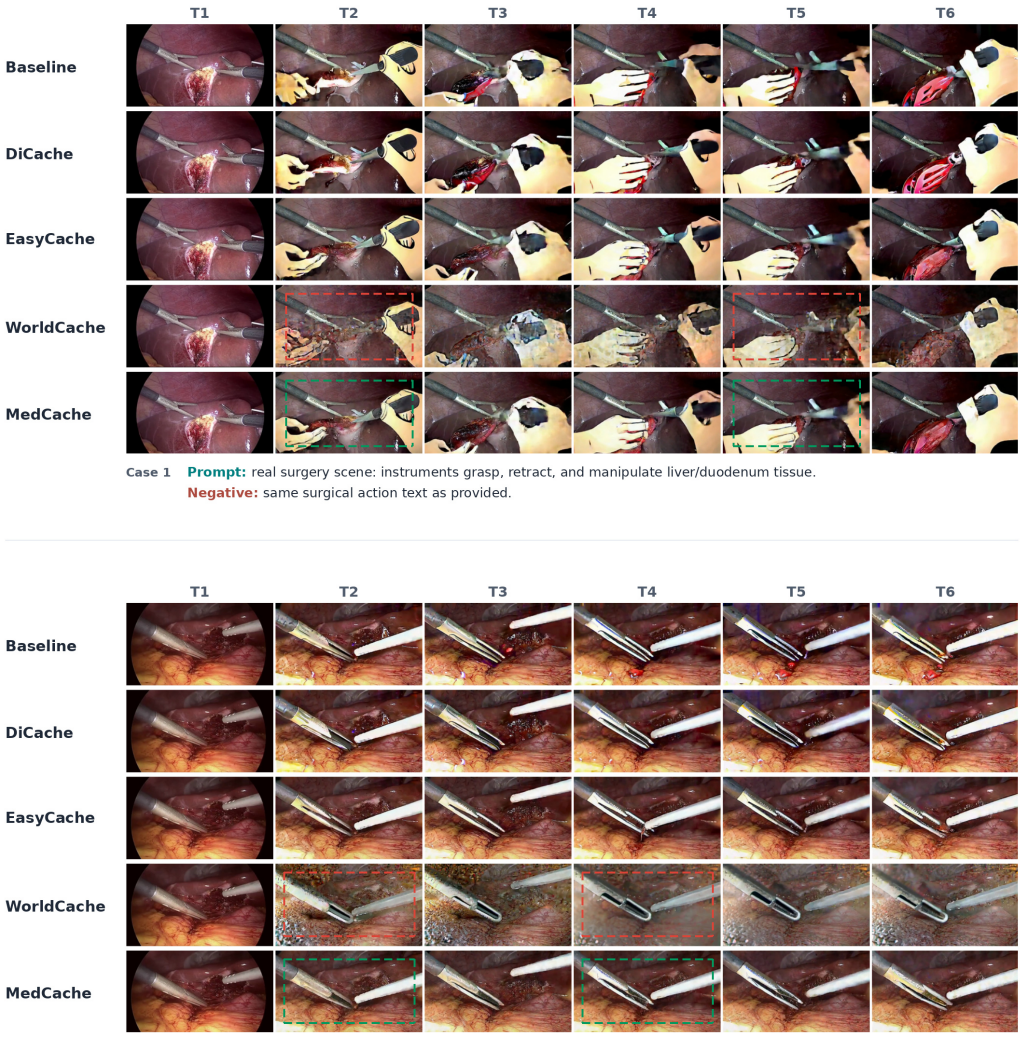
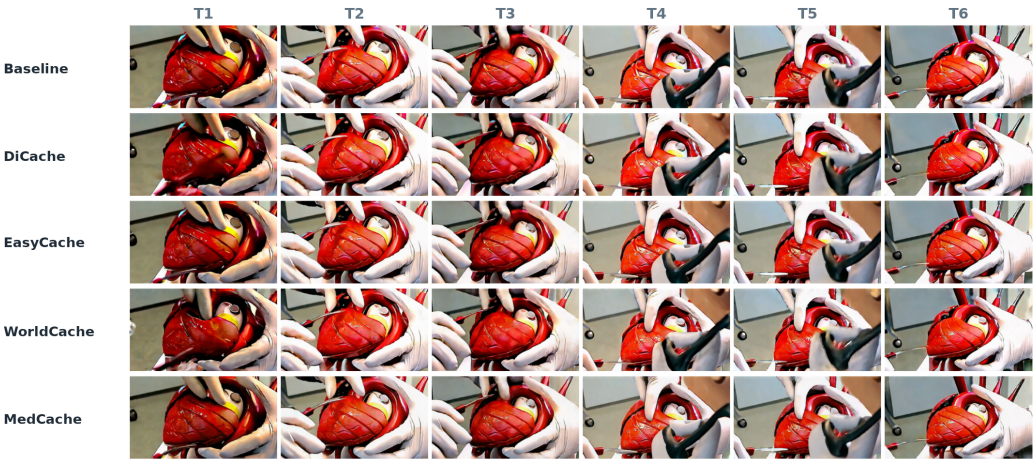


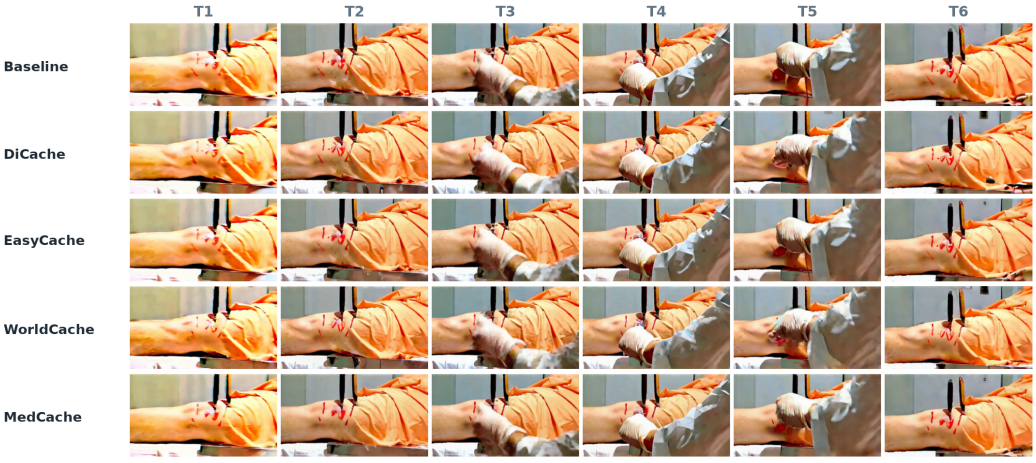
Figure 10: **Additional Image2World qualitative comparisons on surgical prompts.** Given the same conditioning, WorldCache exhibits local texture degradation and structure distortion around active tool-tissue contact regions, while MedCache preserves more coherent surgical structure across time. Red dashed boxes mark representative WorldCache artifacts; green dashed boxes mark the corresponding regions in MedCache.

Third, the current evaluation is strongest on open rectified-flow world models and on medical or medically adapted generation. The conclusions should therefore be interpreted at the level of training-free caching for the model families evaluated here.

Several directions follow naturally from these limitations. One direction is to replace the current prompt prior with a more principled estimate of content sensitivity, while preserving the training-free nature of the method. Another is to study whether latent-derived spatial caching composes effectively with other inference-time accelerations, such as adaptive-depth execution or step reduction. A third direction is to extend the evaluation to broader generic benchmarks and additional world-model families, in order to better characterize when spatially selective reuse provides the largest advantage. More broadly, we view MedCache as one instance of a larger design principle: training-free acceleration benefits from using spatial structure in the latent, provided that the decision path remains substantially cheaper than the computation it seeks to skip.



Case 1 **Prompt:** The short video depicts a surgical procedure focused on the heart, specifically the reconstruction of the aortic valve. Set in an operating room, the surgical team is actively engaged in the operation. The video offers close-up views of the surgical site, highlighting the intricate and precise nature of the procedure. The team utilizes various medical instruments, including forceps and specialized tools designed for delicate cardiac surgery. Gloved hands are seen manipulating the heart and surrounding tissues, indicating the use of a retractor to keep the surgical field open. The heart appears exposed, with the surgical team meticulously working around the aortic valve area. The video captures critical steps of the procedure, including the preparation and placement of a prosthetic valve. The team carefully positions and secures the new valve, ensuring it is properly aligned and functioning. Surrounding tissues and blood vessels are visible, emphasizing the complexity and precision required in such cardiac surgeries. Overall, the video provides a detailed look at the technical aspects of aortic valve reconstruction, showcasing the skill and care involved in performing this life-saving procedure.



Case 2 **Prompt:** The short video showcases a surgical procedure being conducted on a patient's leg. Set in a sterile operating room, the patient's leg is exposed and prepped for surgery, marked with surgical lines that indicate areas of interest or incision points. Throughout the video, various medical instruments are visible, including retractors, forceps, and scalpels. The surgeon's gloved hands are actively engaged in the operation, using retractors to hold the incision open and provide a clear view of the leg's internal structures. These retractors are strategically positioned to keep the incision accessible. The surgeon's hands manipulate tissues and may be suturing or repairing internal structures. The presence of blood and the careful handling of tissues suggest that the procedure is complex and delicate. The surgeon's precise movements, along with the use of specialized surgical tools, underscore the technical skill required for the operation. Overall, the video offers a detailed view of the surgical procedure on the patient's leg, highlighting the meticulous and careful nature of the operation, with a focus on the surgeon's hands and the surgical instruments, emphasizing the technical aspects of the medical procedure

Figure 11: **Additional Text2World qualitative comparisons on surgical prompts.** Across both cases, MedCache maintains temporally coherent rollouts close to the uncached baseline while avoiding stronger texture degradation observed in some cache baselines.

G Broader Impacts

MedCache is an inference-time acceleration method for video world models, with the main intended use case being faster generation in medical and physically structured settings. A potential positive impact is improved usability of medical video generation for applications such as surgical simulation, procedural rehearsal, and synthetic data generation, where repeated sampling and interactive workflows are often limited by inference cost. By reducing runtime without retraining, MedCache can also make experimentation with specialized medical checkpoints more accessible in resource-constrained settings.

At the same time, faster video generation can also amplify misuse. In particular, accelerating medical or procedural video synthesis could reduce the cost of producing misleading, unrealistic, or fabricated

Algorithm 1 MedCache forward pass on one CFG branch

Require: current latent x_n, previous same-branch latent x_{n-2}, cached residual Δ_{n-2}, previous computed output y_{n-2}, accumulated error $\hat{E}_{n-2}$, sensitivity k_n, risk map r
Require: base threshold τ, domain scale σ, risk amplification β, guard parameters η, κ, top-K support, warm-up W, cutoff C

$\tau_{\text{eff}} \leftarrow \sigma\tau$
if $n < W$ **or** $n \geq C$ **then**
    $y_n \leftarrow f_\theta(x_n)$
    $\Delta_n \leftarrow y_n - x_n$
    $\hat{E}_n \leftarrow 0$
    if current branch is conditional **then**
        refresh r using Eqs. 4 to 7
    end if
    return y_n
end if
$\delta_n^{\text{MC}} \leftarrow$ Eq. 8 using r, x_n, and x_{n-2}
$\hat{E}_n \leftarrow \hat{E}_{n-2} + k_n \cdot \delta_n^{\text{MC}} / (\|y_{n-2}\|_1 + \epsilon)$
if $\hat{E}_n \geq \kappa\tau_{\text{eff}}$ **then**
    $y_n \leftarrow f_\theta(x_n)$
    $\Delta_n \leftarrow y_n - x_n$
    $\hat{E}_n \leftarrow 0$
    if current branch is conditional **then**
        refresh r using Eqs. 4 to 7
    end if
    return y_n
end if
$\hat{y}_n \leftarrow x_n + \Delta_{n-2}$
if $\hat{E}_n \geq \eta\tau_{\text{eff}}$ **then**
    form top-K mask Ω from r
    $\rho_n \leftarrow$ Eq. 10 on Ω
    if $\rho_n > 1.50$ **then**
        $y_n \leftarrow f_\theta(x_n)$
        $\Delta_n \leftarrow y_n - x_n$
        $\hat{E}_n \leftarrow 0$
        if current branch is conditional **then**
            refresh r using Eqs. 4 to 7
        end if
        return y_n
    end if
end if
return $\hat{y}_n$

visual content. Even in intended use, generated videos may contain clinically implausible details or omit important procedural structure. As a result, the outputs of the accelerated generative models should not be treated as clinical evidence or as substitutes for expert supervision. More broadly, MedCache is an efficiency method rather than a validation method: it speeds up the generation but does not guarantee the safety or fairness of the underlying model.

We therefore view MedCache as most appropriate for research, simulation, and data-generation settings in which generated output is used with domain oversight and where downstream users understand that acceleration does not imply clinical reliability. Future work should study safeguards for deployment, including content validation, uncertainty-aware filtering, and monitoring for misuse in high-stakes settings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's core contribution accurately, which is that MedCache is a training-free, probe-free caching method for video generation models that makes cache decisions spatially selective using latent-derived risk. They also reflect the empirical scope of the paper by emphasizing the strongest gains in medical and image-conditioned settings, rather than claiming uniform improvement across all models. The claims are supported by the reported Text2World, Image2World, and ablation results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations and Future Work section in the appendix that discusses the main limitations of MedCache. Some of them include: strongest gains occur when reuse risk is spatially localized, the domain prior is heuristic and prompt-based, and the evaluation is done on a specific set of benchmarks. The section also clarifies that performance can differ on broader generic checkpoints and outlines directions for broader evaluation and more principled domain adaptation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results such as theorems, lemmas, or proofs. Its contributions are methodological and empirical, supported by experimental evaluation and ablation studies.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the evaluated models, benchmark settings (Text2World and Image2World), baselines (DiCache, EasyCache, WorldCache, FasterCache), metrics, sampling configuration, hardware, latency measurement protocol, MedCache hyperparameters, and the hyperparameter selection protocol. The method is also described algorithmically in the appendix, including the risk map, drift update, guard, and CFG-aware bookkeeping, so the main experimental results can be reproduced using the released base models and datasets.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The MedCache implementation and evaluation scripts can be accessed via the project page (`https://ufaqkhan.github.io/MedCache-Website`). All evaluated base models (WAN2.1, Cosmos-Predict-2.5, Cosmos-H-Surgical, and HunyuanVideo) and benchmarks (Med-VBench, CholecT50/SurgΣ, VBench) are publicly available, and the sampling configurations, hardware, and frozen hyperparameters are specified in Appendix A.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the experimental setup needed to interpret and reproduce the results, including the evaluated models and benchmarks, baselines, metrics, sampling configuration, hardware, latency measurement protocol, MedCache hyperparameters, and the hyperparameter selection protocol.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars or formal significance tests for the main tables. This is primarily due to the computational cost of repeated evaluation: each run requires multi-step video generation on large rectified-flow backbones, so full multi-seed replication across all models, baselines, and benchmarks was not feasible within the available compute budget. To reduce uncontrolled variation, all methods were evaluated under matched sampling settings, on the same checkpoints, prompts/images, same seed, hardware, and latency protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies the compute setting used for the reported experiments, including single-GPU NVIDIA H200 runs for the Cosmos-Predict-2.5 and Cosmos-H-Surgical experiments, 8-GPU AMD Instinct MI210 distributed inference for the WAN2.1 14B experiments, the matched sampling configurations used within each model block, and the wall-clock latency measurement protocol. The main and supplementary tables report per-sample latency and speedup for each evaluated method.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics, and the research reported in this paper conforms to it. The work focuses on training-free inference acceleration for publicly available video generation models, reports results transparently including limitations and boundary cases, and does not involve deceptive practices, human subjects experimentation, or restricted-access private data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive and negative societal impacts. On the positive side, faster medical video generation can support applications such as surgical simulation, procedural rehearsal, and synthetic data generation. On the negative side, the same acceleration could lower the cost of generating misleading or unrealistic medical video, and incorrect or low-fidelity generations could be harmful if over-trusted in training or evaluation workflows. The paper also notes that the method should be understood as an inference acceleration technique rather than a clinically validated decision tool.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new generative model, dataset, or scraped data collection. MedCache is an inference-time acceleration method evaluated on existing publicly available base models and benchmarks. As such, the paper does not introduce a new high-risk model or dataset release that would require release-specific safeguards beyond those already associated with the underlying models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper uses publicly available models, datasets, and baselines, and credits their original creators through citations. However, we do not explicitly list the licenses, versions, and terms of use for all external assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release a new dataset, model, or other standalone asset. MedCache is a method for inference-time acceleration evaluated on existing publicly available models and datasets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects. All experiments are conducted on existing publicly available models, datasets, and benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects, so IRB approval or equivalent review was not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

Justification: LLMs are not an important, original, or non-standard component of the core method, experiments, or scientific claims in this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.