NEURAL INFORMATION PROCESSING SYSTEMS

Mohamed bin Zayed University of Artificial Intelligence



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] MBZUAI [2] DKFZ Heidelberg, Helmholtz Imaging [3] Birmingham City University [4] Qatar University

جامعة قطر
QATAR UNIVERSITY

dkfz.
GERMAN CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION

BIRMINGHAM CITY University

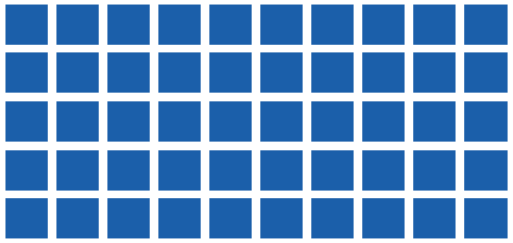
Medical video world models are too slow at inference

- World models such as Cosmos and WAN are being adapted to **surgical simulation, procedural rehearsal and synthetic data**.
- Sampling is the bottleneck: **30–50 denoising steps**, each run twice for classifier-free guidance, through a transformer with billions of parameters.
- Retraining or distilling a medical checkpoint is costly, so we want a **training-free** fix: caching.



Denoising step

50 / 50

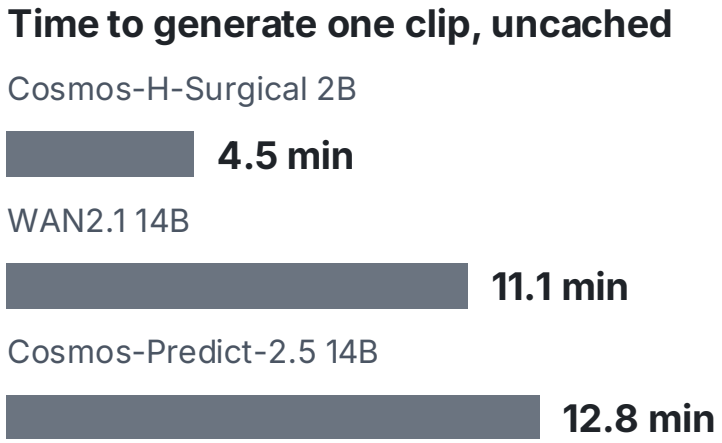


Transformer passes

100 / 100

Every step runs the full transformer twice.

Time to generate one clip, uncached



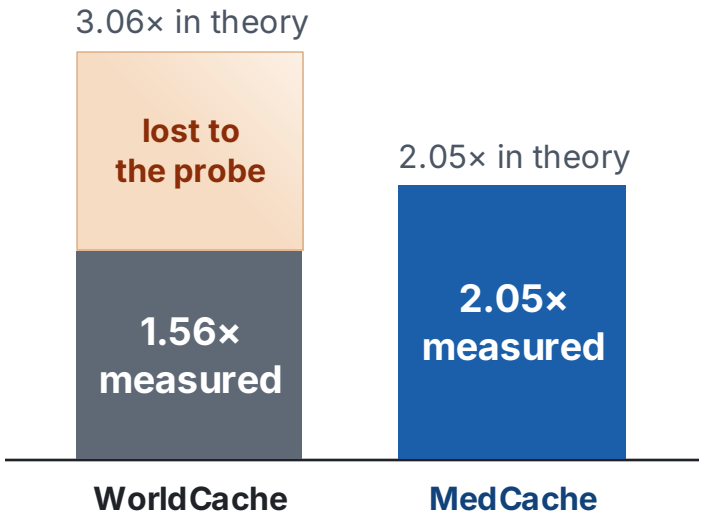
Text2World. Cosmos models on one NVIDIA H200; WAN2.1 on eight AMD MI210.

→ Caching reuses the previous step's output when little has changed. The hard part is deciding *when* reuse is safe.

Existing caches are either spatially blind or costly to run

Method	How it decides	Spatially aware	Cost per pass
EasyCache	One global drift value for the whole latent	**No**	≈ 0
WorldCache	Probe: runs the first 3 transformer blocks on every pass	Yes	**469 ms, 31% of a forward pass**
MedCache (ours)	Risk map computed from the latent itself	**Yes**	**Under 1 ms**

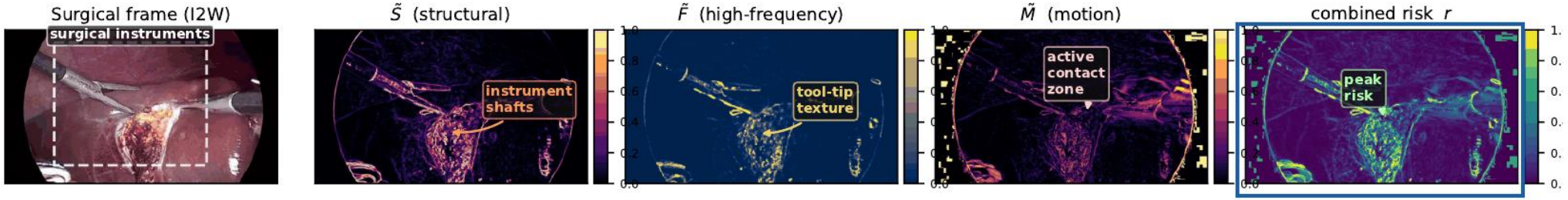
The probe tax: theoretical vs. measured speedup



→ Goal: the spatial awareness of a probe at the cost of a scalar test.

Key idea: the spatial signal is already in the latent

Reuse errors matter in small regions: instruments cover only **7.4%** of the frame on average. Three cheap cues, read from the latent, mark where reuse is risky:



S: structure

Sobel gradients. Organ boundaries and tool contours.

F: fine detail

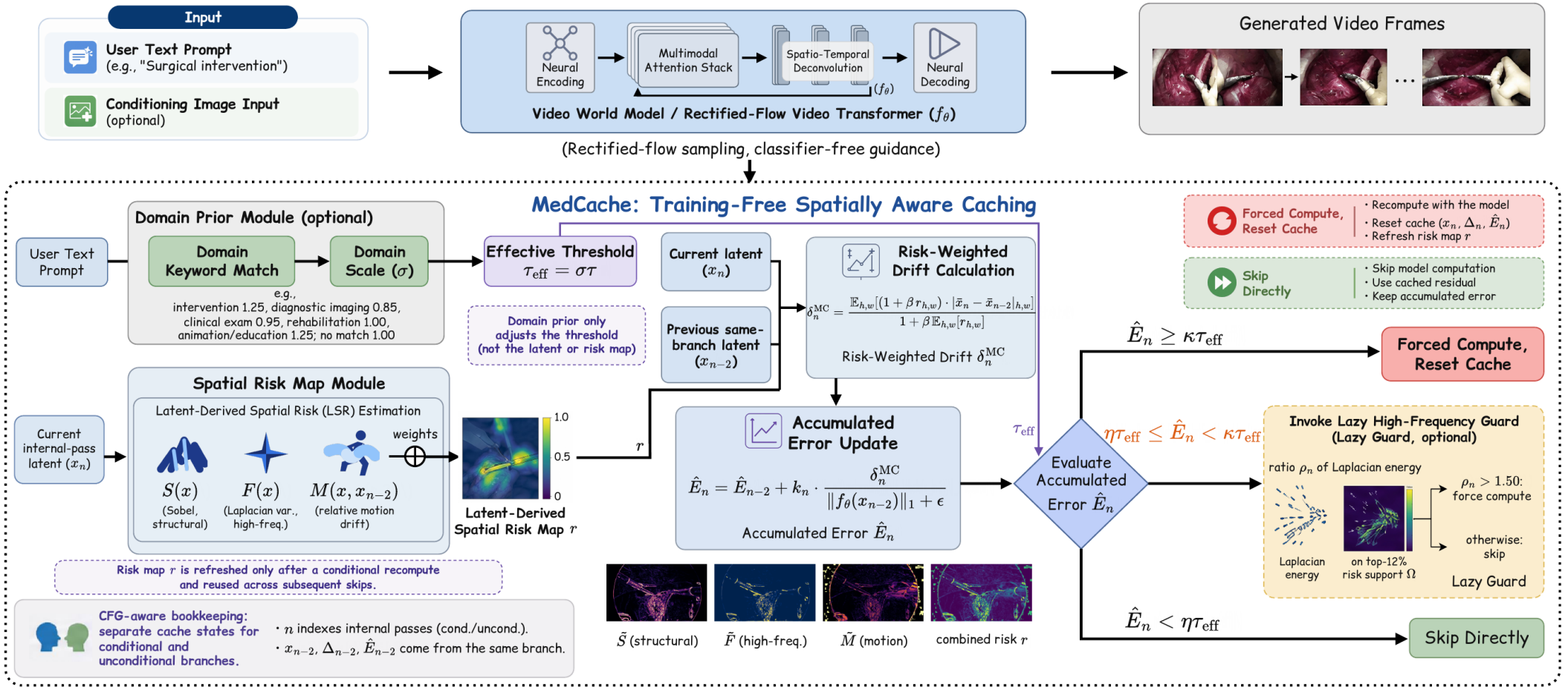
Laplacian variance. Sutures, thin interfaces, texture.

M: local motion

Change against the previous latent. Moving tools, deforming tissue.

Risk map r = clip(S + 0.75·F + 1.25·M) · no learned module, no network probe, under 1 ms

MedCache architecture: a training-free decision layer



Decision rule: skip, check, or recompute

1. Weight the drift by the risk map

$$\delta = E[(1 + \beta \cdot r) \cdot |\Delta x|] \,/\, (1 + \beta \cdot E[r])$$

Change in high-risk regions counts more. With uniform risk this is exactly EasyCache.

2. Accumulate and compare with threshold τ

$$\hat{E} \leftarrow \hat{E} + k \cdot \delta \,/\, \|f(x)\|$$

An optional prompt prior scales τ: ×1.25 for interventions, ×0.85 for diagnostic imaging.

→

Ê below 0.2 τ

Skip: reuse the cached residual

Ê between 0.2 τ and 1.1 τ

Check: lazy guard

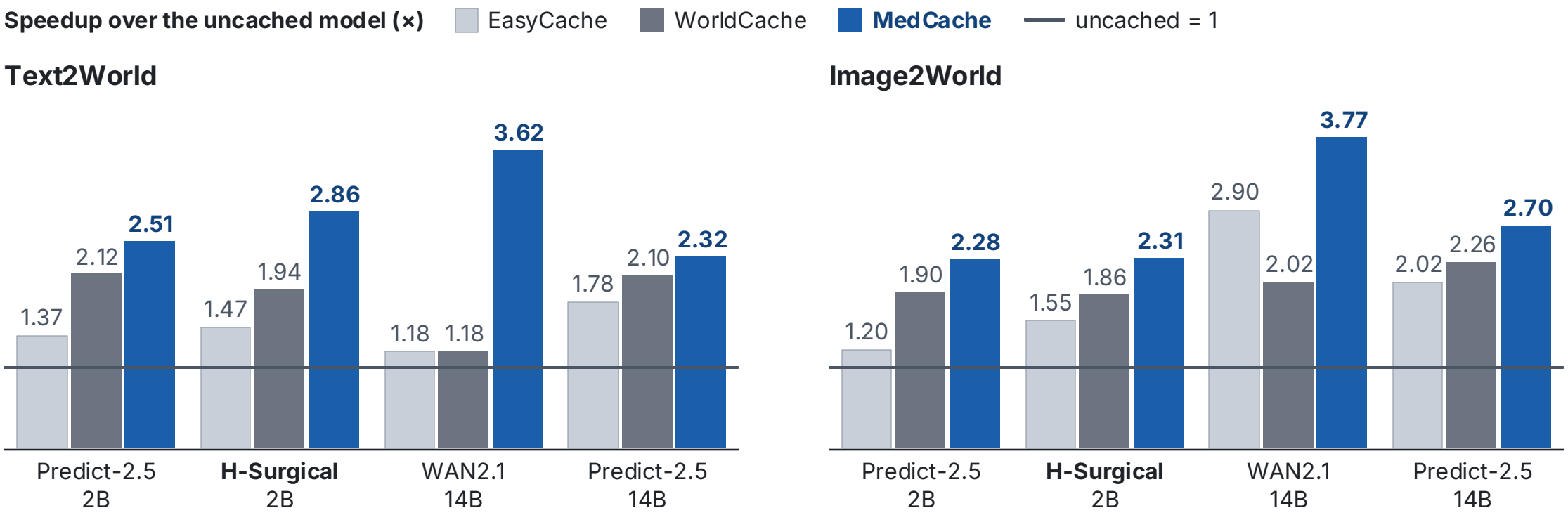
Laplacian energy on the top-12% risk positions; recompute only if it rises past 1.5×

Ê at or above 1.1 τ

Recompute; refresh cache and risk map

- The two guidance branches (conditional and unconditional) keep separate cache states.
- Example Image2World rollout on Cosmos-H-Surgical: **21 of 35** denoising steps skipped, with no guard recompute.

Faster than EasyCache and WorldCache in all 8 settings



- **Surgical model, Text2World: 2.86×.** One clip drops from 272 s to 95 s.
- **WAN2.1 14B: 3.62× and 3.77×.** Image2World drops from 524 s to 139 s.

Quality stays close to the uncached model

Med-VBench Total ↑	Uncached	EasyCache	WorldCache	MedCache	Speedup
Text2World					
Cosmos-Predict-2.5 2B	0.916	0.863	0.891	**0.899**	2.51×
Cosmos-H-Surgical 2B	0.862	0.767	0.814	**0.837**	2.86×
WAN2.1 14B	0.904	**0.898**	**0.898**	0.891	3.62×
Cosmos-Predict-2.5 14B	0.896	0.891	0.888	**0.895**	2.32×
Image2World					
Cosmos-Predict-2.5 2B	0.792	0.788	0.778	**0.795**	2.28×
Cosmos-H-Surgical 2B	0.850	0.847	0.839	**0.850**	2.31×
WAN2.1 14B	0.885	0.883	**0.884**	0.883	3.77×
Cosmos-Predict-2.5 14B	0.865	0.867	0.843	**0.870**	2.70×

Bold: best of the three caches. Total: mean of image quality, subject consistency, background consistency and motion smoothness.

Image2World, Cosmos models

Equal to or above the uncached score on all three, at 2.3–2.7×. On the surgical model: 0.850 vs. 0.850.

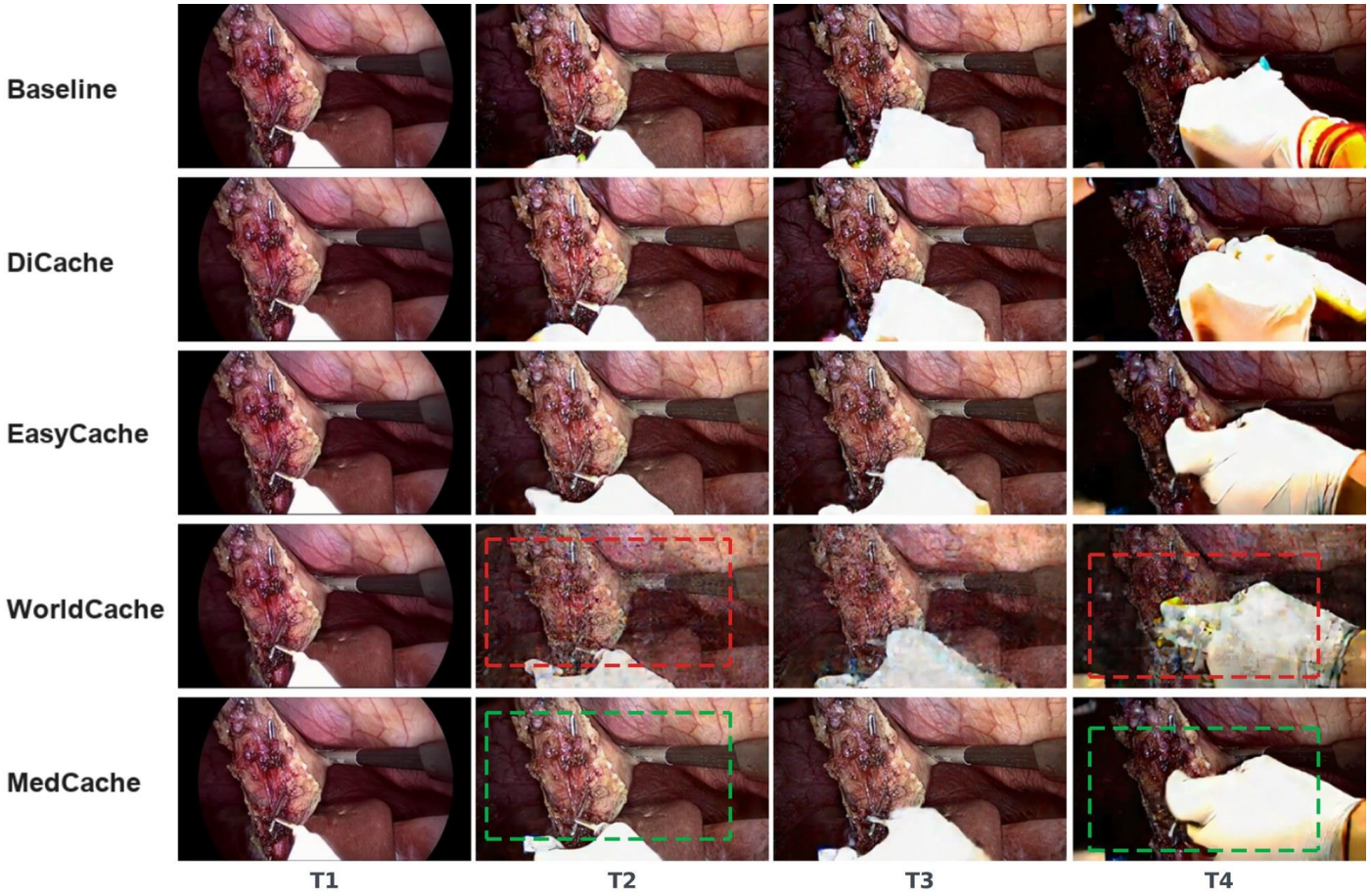
Text2World, surgical model

An out-of-distribution stress test: every cache loses quality. MedCache loses least (−0.025; WorldCache −0.048, EasyCache −0.095).

WAN2.1, general domain

Here we trade quality for speed: −0.013 and −0.002 vs. uncached, for 3.6–3.8×.

Detail is preserved at the tool–tissue interface



Same input image and prompt. **Red:** WorldCache artifacts. **Green:** the same region with MedCache.

Deviation from the uncached video

Mean absolute error (0–255), lower is better

Region	WorldCache	MedCache
Instrument	32.5	**13.7**
Its boundary	31.0	**13.7**

→ 57.9% lower error inside instrument regions, 55.7% on their boundaries

260 Cosmos-H-Surgical Image2World cases, same prompt, image and seed. A cache-fidelity measure, not a claim of clinical correctness.

Where the gain comes from, and whether it transfers

Component ablation · Cosmos-H-Surgical 2B, Image2World

Configuration	Speedup	Skipped	Image qual.	Total
Uncached	1.00×	–	0.610	0.844
EasyCache	1.39×	–	0.594	0.840
+ risk-weighted drift	**2.37×**	61%	0.593	0.840
+ domain prior	2.52×	64%	0.593	0.840
+ lazy guard (full)	2.52×	64%	**0.612**	**0.846**

Transfer · HunyuanVideo, VBench

Method	Total	Speedup
Uncached	0.896	1.00×
EasyCache	0.897	2.03×
WorldCache	0.890	2.93×
MedCache	**0.899**	**3.44×**

- **Risk-weighted drift alone** takes EasyCache from 1.39× to 2.37× at the same quality; the guard then restores image quality.
- **Beyond medicine:** 3.44× on HunyuanVideo with uncached-level quality.

Summary

- Reuse errors in medical video are **spatially local**, and the latent already shows where.
- **MedCache** = latent risk map + risk-weighted drift + lazy guard. Training-free, no network probe.
- **2.3–3.8× faster** than uncached on four checkpoints; quality at or near uncached on the Cosmos models.
- **58% lower error** than WorldCache inside instrument regions.

Limitations: the quality advantage is smaller when motion is spread over the whole frame (e.g. WAN2.1), and the prompt prior is a keyword heuristic.

Thank you

Scan me 👇

MedCache project page
ufaqkhan.github.io/MedCache-Website