READING POINT TRACKS FROM FROZEN VIDEO DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We show that readout design substantially improves point tracking from frozen video diffusion transformers, without training a tracking network. Our readout combines confidence-weighted attention heads, bilinear source sampling, local position decoding and visibility estimation from one transformer block. All settings are selected on 20 labelled synthetic clips and transferred unchanged to real videos. CogVideoX-2B reaches 45.21 average Jaccard (AJ) on TAP-Vid DAVIS and 44.00 on Kinetics. Evaluating released HeFT on the same videos, queries and checkpoints with the official TAP-Vid first-query metric establishes gains on both benchmarks for CogVideoX and Wan, including +18.62 and +12.77 AJ on Kinetics. Controlled ablations on both backbones show why localisation improves: bilinear source sampling and local decoding reinforce each other, and most of the local-decoding gain comes from predicting positions between token centres. On CogVideoX, confidence weighting further improves head aggregation, whereas tested block combinations give no clear gain on the synthetic selection set. Across six checkpoints from three generator families, larger models do not consistently improve tracking under this readout. Our readout improves zero-shot correspondence estimation while keeping the video generator frozen.

1 INTRODUCTION

Point tracking follows physical locations as their position, appearance and visibility change. Dedicated trackers learn this task from synthetic trajectories and real videos, including pseudo-labelled data (Karaev et al., 2024). Video generators offer another source of motion representations: they learn from large image and video collections without trajectory annotations (Yang et al., 2024; Wan-Team, 2025). Reading correspondences from these models reuses that pretraining without fitting a separate tracking network. It also provides a direct test of the motion information learned for video synthesis.

DiffTrack and HeFT demonstrate that internal video-generator features support correspondence (Nam et al., 2025; Yuan et al., 2025). Turning those features into precise tracks requires decoding positions between tokens, combining features from different heads and deciding whether a match is visible. A poor readout can therefore understate the correspondence information in otherwise useful frozen features. We study which readout operations improve tracking, which effects interact, and how tracking accuracy changes with generator size.

Building on DiffTrack's query-key matching, we read one frozen attention block and share extraction and matching across queries at the same frame. Hyperparameters selected on labelled synthetic clips transfer unchanged to real-video benchmarks. Figure 1 shows example tracks from our approach compared to ground truth.

Our contributions are:

- **A simplified tracking readout from frozen generative features.** We combine heads according to matching confidence, sample matching maps at continuous query locations, decode positions locally and estimate visibility from scores and cycle consistency (Figure 2). CogVideoX-2B reaches 45.21 average Jaccard on DAVIS and 44.00 on Kinetics with the same synthetic calibration.

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



Figure 1: **Point tracking with a frozen video generator.** CogVideoX-2B predictions on three DAVIS videos. Colours identify tracks; lines show the preceding 12 frames. Filled/hollow markers denote predicted visibility/occlusion; pale rings mark visible ground truth. Ten frame-zero queries are shown per video; each displayed video's AJ scores its full set of 25 queries over evaluated frames. The examples include accurate motion and transient mismatches.

- **Controlled improvements over frozen-generator trackers.** Using identical videos, queries, checkpoints and the official TAP-Vid first-query metric, our readout improves on released HeFT by +13.10 AJ on CogVideoX and +9.31 on Wan on DAVIS (Table 2). The gains increase to +18.62 and +12.77 on the same 1,144 Kinetics clips. Additional controls preserve the DAVIS advantages under identical input preprocessing and HeFT's own evaluator.
- **An explanation of localisation gains and the limits of scaling.** Factorial ablations on CogVideoX and Wan show that bilinear source sampling and local decoding reinforce each other, with most of the decoding gain coming from positions between token centres. We test whether confidence identifies useful heads and whether combining heads or blocks improves correspondence. Evaluating six checkpoints across three model families shows that larger generators do not consistently improve tracking under this readout.

2 Related Work

Point tracking. TAP-Vid evaluates point trajectories and visibility across videos with motion, deformation and occlusion (Doersch et al., 2022). TAPIR combines per-frame matching with temporal refinement (Doersch et al., 2023). CoTracker3 trains a tracking model with synthetic data and pseudo-labelled real videos (Karaev et al., 2024); TAPNext formulates tracking as sequential token decoding (Zholus et al., 2025). Their reported results provide trained references for our setting. Without tracking annotations, contrastive random walks train cycle-consistent matching (Shrivastava & Owens, 2024), while Opt-CWM learns counterfactual probes (Stojanov et al., 2025). Our readout uses synthetic annotations for selection but trains no tracking parameters.

Correspondence from generative representations. DIFT demonstrates that image-diffusion features support several forms of correspondence without task-specific fine-tuning (Tang et al., 2023).

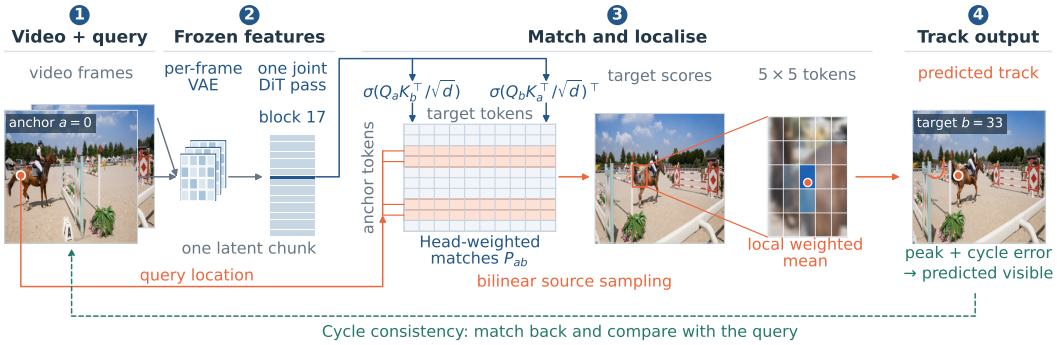


Figure 2: **Tracking from a frozen generator's attention.** Frames are VAE-encoded separately, then processed jointly in one DiT chunk. Per head, σ is row-wise softmax and d is head width; the reverse map is transposed before averaging. Peak weights per anchor row give P_{ab} (schematic). Four neighbouring anchor rows (orange) are bilinearly sampled and reshaped to the target grid; a local mean around the maximum gives the position. Image panels show a real CogVideoX example. The separate return match supplies cycle error for visibility.

Diffusion Hyperfeatures learns to aggregate features across layers and timesteps for semantic correspondence (Luo et al., 2023). Track4Gen jointly trains video generation and point tracking to improve temporal consistency (Jeong et al., 2025). For video diffusion transformers, DiffTrack constructs temporal matches from query-key features (Nam et al., 2025). HeFT uses selected key features, frequency filtering, descriptor refinement and consistency checks (Yuan et al., 2025). We study how head weighting, local decoding and synthetic visibility calibration affect these readouts.

Query-conditioned inference. Point Prompting adds a marker at a query and tracks it through counterfactual video generation (Shrivastava et al., 2025). KL-tracing uses changes in predictive distributions under perturbations to estimate correspondences (Kim et al., 2025a). Both access the model through its response to a query, requiring repeated inference when many points are tracked. Our comparison examines that computation alongside the accuracy of a shared feature readout.

Adaptation of pretrained features. DINO-Tracker adapts DINO features and a tracker to an individual video using self-supervised objectives (Tumanyan et al., 2024). DRIFT combines temporal label propagation from an image diffusion model with test-time optimisation and mask refinement for video object segmentation (Kim et al., 2025b). DiTracker incorporates video DiT features into trained trackers through matching costs, feature fusion and LoRA adaptation (Son et al., 2025). Our main readout fits no tracking parameters; the separate adaptation experiments in Appendix M examine restricted alternatives under their own reference configurations.

3 Tracking with a Frozen Video Generator

Given T frames $I_{0:T-1}$, query q specifies anchor frame a_q and 2D pixel position x_q. We predict its subsequent position and visibility relative to that anchor (Figure 2), without training tracking parameters.

3.1 Feature Extraction

Each frame is encoded separately by the pretrained variational autoencoder (VAE). Following DiffTrack's interleaved chunks (Nam et al., 2025), each transformer pass jointly processes the anchor and up to twelve targets. We remove the anchor from the frame list, sample at stride $\max(1, \lfloor T/12 \rfloor)$, and shift the start one entry per chunk until all targets are covered. For 50 frames, five passes cover 49 targets. Queries at the same anchor share these *anchor passes*.

The frozen transformer receives clean latents at a fixed conditioning timestep t, without added noise or denoising. We capture queries and keys from one block after query/key normalisation and positional encoding where present.

We use CogVideoX, Wan2.1 and Cosmos-Predict2 (Yang et al., 2024; WanTeam, 2025; NVIDIA, 2025). Spatial tokens are 16 model pixels apart, giving a 30×45 grid at 480×720 input. More implementation details are in Appendix A.

3.2 Correspondence estimation

Let $Q_f^h, K_f^h \in \mathbb{R}^{M \times d}$ be head h's query and key matrices for frame f, with M spatial tokens and d channels per head. For anchor $a = a_q$ and target b, we form per-head scores S_{ab}^h and their weighted aggregate P_{ab}:

$$S_{ab}^h = \tfrac{1}{2}\left[\sigma(Q_a^h {K_b^h}^\top/\sqrt{d}) + \sigma(Q_b^h {K_a^h}^\top/\sqrt{d})^\top\right], \qquad P_{ab}(i,j) = \sum_h w_{h,i}^{ab} S_{ab}^h(i,j). \tag{1}$$

In Equation 1, i, j index source and target tokens, and σ is a row-wise softmax over the other frame's tokens. Transposition aligns the reverse scores with the forward scores before averaging. The resulting rows do not necessarily sum to one: S_{ab}^h and P_{ab} are matching scores, not row-normalised probability distributions.

We use each head's peak matching score as a confidence proxy, assigning weights separately for each source token and target frame. The confidence exponent is β; g ranges over all attention heads:

$$w_{h,i}^{ab} = \frac{[\max_j S_{ab}^h(i,j)]^\beta}{\sum_g [\max_j S_{ab}^g(i,j)]^\beta}. \tag{2}$$

Equation 2 averages heads uniformly at $\beta = 0$; $\beta > 0$ favours higher peaks. The scores are not calibrated correctness probabilities. Weights are computed directly, without fitting parameters to a video.

We average repeated anchor-target pairs' head-weighted maps across chunks. Expressing the query relative to token centres, we bilinearly sample the source rows at its continuous location to obtain query q's target score map p_{qb}. With $j^\star = \arg\max_j p_{qb}(j)$, the predicted token coordinate is

$$\widehat{x}_{qb} = \frac{\sum_{j \in \mathcal{W}(j^\star)} p_{qb}(j)\, u_j}{\sum_{j \in \mathcal{W}(j^\star)} p_{qb}(j)}, \tag{3}$$

where $u_j \in \mathbb{R}^2$ is token j's coordinate and $\mathcal{W}(j^\star)$ is the 5×5 window centred on the maximum, clipped at image boundaries. We convert Equation 3's weighted mean to evaluation coordinates: token index zero maps to model pixel eight, versus zero in DiffTrack (more coordinate details in Appendix E). Local weighted decoding is established in tracking, including TAPIR's thresholded soft argmax (Doersch et al., 2023); we test its effect on frozen DiT maps. Predictions do not update the original query descriptor.

3.3 Visibility estimation

We use the peak score $s_{qb} = \max_j p_{qb}(j)$ and cycle-consistency error e_{qb}. We bilinearly sample rows of $P_{ab}^\top$ at the predicted target position and apply the same local decoder to return to the anchor. Head weights are not recomputed. The Euclidean return distance to the original query in token coordinates is e_{qb}. A visible match should have a sufficiently high score and a sufficiently small return error, although an incorrect match can also satisfy these conditions.

Synthetic annotations select cues, thresholds and optional three-frame median smoothing. CogVideoX and Cosmos select both smoothed cues; Wan-1.3B selects unsmoothed cycle error. The later visibility ablation expands the original candidate set and reports its selections separately (more details in Appendix G). At the query frame, we enforce the supplied position and visibility.

3.4 Hyperparameter selection on synthetic data

We select extraction settings and the confidence exponent by position accuracy on 20 Kubric MOVi-E validation clips (Greff et al., 2022), then select visibility by average Jaccard. Changes that alter the score distribution, such as head weighting or input resolution, receive their own synthetic visibility calibration. All settings transfer unchanged to real video. The selected (block, t, β) are $(17, 19, 2)$ for CogVideoX-2B, $(18, 129, 2)$ for Wan-1.3B and $(26, 308, 1)$ for Cosmos-2B. More selection details are in Appendix B.

3.5 SHARING COMPUTATION ACROSS QUERIES

Queries at the same anchor share extraction and matching; each needs only map lookup, decoding and visibility checks. A cached variant reuses features from frame-zero chunks across anchors, saving passes but changing temporal context. Sections 4.5 and 5.1 assess its computational saving and accuracy cost.

4 EXPERIMENTS

We ask how accurately frozen video generators can track points without training a tracker. We compare zero-shot methods, test transfer of synthetic calibration across benchmarks, then control the HeFT comparison to distinguish readout improvements from protocol differences.

4.1 EVALUATION PROTOCOL

With settings fixed on Kubric, we evaluate transfer using TAP-Vid first mode (Doersch et al., 2022): each point is queried at its first visible frame and evaluated strictly afterwards. DAVIS has 30 videos and 650 queries; all defaults use the same 1,144 of 1,147 Kinetics clips recoverable from the available archive (more details in Appendix C). Position accuracy δ averages the fractions of annotated visible positions within 1, 2, 4, 8 and 16 pixels in 256×256 evaluation coordinates. Occlusion accuracy (OA) measures visibility classification; average Jaccard (AJ) evaluates localisation and visibility jointly. Scores range from 0 to 100 and are averaged over videos. In tables, $\uparrow$ means higher is better and $\downarrow$ means lower is better; Δ denotes a difference in score points, with subtraction order stated.

Oracle diagnostics use benchmark annotations retrospectively to select heads or supply visibility. Unless labelled otherwise, $[l, u]$ denotes the lower and upper limits of a 95% confidence interval (CI), computed from 10,000 percentile-bootstrap video resamples at fixed queries and calibration. Paired comparisons resample videos jointly. Intervals are not adjusted for multiple comparisons; those containing zero leave the direction of a difference uncertain.

4.2 ZERO-SHOT BENCHMARK RESULTS

Our primary comparison is with the zero-shot methods in Table 1. CogVideoX-2B reaches 45.21 AJ and 61.27 position accuracy on DAVIS. Its AJ exceeds the published Point Prompting and KL-tracing scores, although KL-tracing retains higher position accuracy. Both CogVideoX and Wan also exceed HeFT's published AJ for the corresponding backbone. A shared attention readout is therefore competitive with both feature-based trackers and query-conditioned generation. Published protocols differ; Section 4.4 tests the HeFT gains on identical examples. Trained trackers and Opt-CWM provide context for methods with additional learning.

4.3 CROSS-BENCHMARK TRANSFER

Our clearest advantage in Table 1 is on Kinetics. With the DAVIS configuration unchanged, CogVideoX-2B reaches 44.00 AJ and 59.53 position accuracy, the highest Kinetics values for both metrics among the zero-shot entries shown. These scores remain close to its DAVIS performance without retuning. Wan and Cosmos reach 38.98 and 39.63 AJ on the same clips, showing that the synthetic calibration transfers across all three generator families.

RGB-Stacking tests the same readout beyond natural video, on 50 synthetic robot-manipulation videos and 1,498 first-mode queries. Without retuning, CogVideoX-2B reaches 54.21 AJ. CoTracker3 scores 74.49 on identical queries, so this establishes transfer of the frozen readout while leaving an accuracy gap to supervised tracking (more details in Appendix C).

4.4 COMPARISON WITH HEFT UNDER A COMMON PROTOCOL

To separate readout quality from protocol, we run released HeFT (Yuan et al., 2025) on first-visible queries with the official TAP-Vid first metric, retaining its settings and preprocessing (Table 2). Its

Table 1: **Benchmark comparison with published methods.** Reference scores use their published protocols and do not establish a matched ranking; Table 2 evaluates HeFT under TAP-Vid first mode. Input is the backbone resolution; metrics use 256×256 coordinates. Training: **S** uses annotated tracks; **V** uses additional learning from video without track labels; **Z**: no additional tracker training. Our **Z** rows use synthetic annotations for hyperparameter selection. $*$ Point Prompting selects one query per DAVIS video. ‡ DiffTrack's DAVIS result uses frame-zero queries. † Point Prompting's Kinetics result uses 30 sampled clips. Published values and source citations are given in Appendix C; HeFT's exact Kinetics clip set is unverified. Our completed Kinetics evaluation uses 1,144 clips. - unreported; ∅ no visibility prediction supplied.

				DAVIS			Kinetics		
Method	Backbone	Training	Input	AJ↑	δ↑	OA↑	AJ↑	δ↑	OA↑
CoTracker3	-	**S**	-	64.45	77.13	90.90	54.35	65.99	89.43
TAPNext	-	**S**	-	64.48	77.29	91.71	54.94	65.23	90.06
DiTracker	CogVideoX-2B	**S**	-	64.80	78.00	89.00	55.20	68.30	86.30
Opt-CWM	-	**V**	-	46.50	62.05	80.07	43.70	53.74	84.45
DIFT	SD2.1	**Z**	-	21.51	39.55	69.71	15.10	25.56	63.17
DiffTrack‡	CogVideoX-2B	**Z**	480×720	∅	46.30	∅	∅	36.35	∅
Point Prompting*	Wan2.1-14B	**Z**	480×832	42.21	57.29	82.90	27.36†	41.51	71.39
KL-tracing	LRAS-7B	**Z**	256×256	44.16	65.20	74.58	-	-	-
HeFT (published)	CogVideoX-2B	**Z**	480×720	39.54	54.14	81.46	29.33	46.78	72.15
HeFT (published)	Wan2.1-1.3B	**Z**	480×832	41.27	55.94	80.09	30.06	47.68	73.63
HeFT (published)	Cosmos-2B	**Z**	*704×1280*	48.61	63.50	82.47	36.64	54.47	78.60
Ours	CogVideoX-2B	**Z**	480×720	45.21	61.27	76.14	44.00	59.53	79.08
Ours	Wan2.1-1.3B	**Z**	480×832	43.64	57.42	79.22	38.98	53.10	79.55
Ours	Cosmos-2B	**Z**	480×720	39.73	54.83	74.64	39.63	54.70	77.58
Ours	Cosmos-2B	**Z**	*704×1280*	39.93	56.37	70.52	-	-	-

Table 2: **DAVIS comparison under TAP-Vid first.** All 30 videos and 650 queries, with identical checkpoints, full video lengths and matched input dimensions within each backbone. Each method retains native preprocessing. Differences are ours minus HeFT, with paired 95% intervals. Adapted DiffTrack supports later query frames and adds Kubric-selected visibility.

Method	AJ↑	δ↑	OA↑	Ours - HeFT AJ↑ [95% CI]
CogVideoX-2B, 480×720				
HeFT (TAP-Vid first)	32.11	47.44	77.99	-
DiffTrack readout + visibility	33.46	45.04	74.14	-
Ours	45.21	61.27	76.14	+13.10 [+9.28, +17.29]
Wan2.1-1.3B, 480×832				
HeFT (TAP-Vid first)	34.33	48.94	79.04	-
Ours	43.64	57.42	79.22	+9.31 [+6.91, +11.70]
Cosmos-2B, 704×1280				
HeFT (TAP-Vid first)	41.47	58.07	79.19	-
Ours	39.93	56.37	70.52	-1.54 [-4.81, +1.82]

own evaluator restricts queries to frame zero, drops incomplete final chunks and counts the query frame.

On DAVIS, our readout improves AJ for CogVideoX and Wan; the Cosmos difference remains uncertain. This pattern also holds under HeFT's evaluator and frame selection. Giving HeFT our exact input frames and physical query locations preserves gains of +13.43 [+9.66, +17.59] AJ for CogVideoX and +10.45 [+8.27, +12.65] for Wan. Cosmos remains uncertain at +0.20 [-2.78, +3.27], with higher OA for HeFT.

The CogVideoX and Wan advantages also hold on all 1,144 Kinetics clips: +18.62 [+18.06, +19.20] and +12.77 [+12.29, +13.23] AJ, respectively. Cosmos gains +7.78 [+6.69, +8.87] on a seeded 200-clip subset, with ours at 480×720 and HeFT at 704×1280. HeFT's published Kinetics

Table 3: **Separating source sampling, coordinates and local decoding.** Shared CogVideoX-2B features: block 17, $t = 19$, clean latents, anchor passes, 480×720, $\beta = 2$; fixed queries and visibility. *DiffTrack* denotes its coordinate mapping, not its full readout: token 0 maps to pixel 0, versus pixel 8 for token centres. A 1×1 window is argmax.

Source sampling	Coordinates	Window (tokens)	Kubric $\delta\uparrow$	DAVIS $\delta\uparrow$
Nearest token	DiffTrack	1×1	43.79	40.84
Nearest token	DiffTrack	5×5	47.60	44.09
Bilinear	DiffTrack	1×1	47.04	43.35
Bilinear	DiffTrack	5×5	59.19	52.42
Nearest token	Token centres	1×1	47.66	45.40
Nearest token	Token centres	5×5	51.81	49.14
Bilinear	Token centres	1×1	50.34	47.71
Bilinear	Token centres	5×5	69.74	61.27

scores remain incompletely reproduced. With its driver unreleased and clip overlap unverified, our comparison concerns released tracking code on identical recovered clips.

Adapted DiffTrack interpolates per-source-token argmax positions from uniformly averaged heads. Our readout gains +11.75 [+9.25, +14.62] AJ on DAVIS and +8.97 on the same Kinetics subset. These complete-system gains motivate the controlled readout ablations in Section 5 (more comparison details in Appendix C.2).

4.5 Computational efficiency

Sharing extraction across queries avoids repeated backbone inference, as in DiffTrack and HeFT. CogVideoX tracking through cycle-consistency checks averages 18.77 seconds per DAVIS video on one H100 NVL. Appendix D details the timer and projects KL-tracing's repeated query-conditioned calls; end-to-end latency across trackers is not measured.

5 Ablations and readout analysis

System comparisons do not identify which operations cause the gains. We isolate readout components and their interactions, then test whether other heads, blocks or larger generators improve correspondence.

5.1 Localisation, visibility and feature reuse

The DiffTrack comparison changes several decoder operations together. To attribute the gain, Table 3 varies them independently at fixed features, queries and visibility, reporting position accuracy.

At fixed bilinear source sampling and token-centre coordinates, replacing argmax with a 5×5 score-weighted mean increases DAVIS δ by +13.56 [+11.31, +15.95]. Rounding to token centres removes 12.42 [10.20, 14.78] points, leaving a gain over argmax of +1.15 [+0.71, +1.62]. Thus the benefit comes mostly from positions between token centres, rather than better token selection.

Bilinear source sampling depends on the target decoder: it adds +2.31 δ with argmax but +12.13 with the local mean. Their interaction is +9.82 [+7.84, +12.06], so the effects are not additive. On Wan, local decoding gains +11.18 [+9.40, +13.00] δ, with a source-sampling interaction of +7.91 [+6.32, +9.62] on DAVIS. Separately, either matching direction alone lowers DAVIS AJ on all three backbones, both with reference visibility masks fixed and after synthetic recalibration (more details in Appendix E.1).

Visibility at fixed positions. To isolate visibility errors, we substitute annotated visibility at fixed positions. This raises CogVideoX AJ by +5.48 [+3.84, +7.31] on DAVIS and +4.60 [+4.31, +4.90] on Kinetics. The selected rule beats declaring all points visible. Smoothed score alone changes DAVIS AJ by +0.07 [-0.10, +0.26] relative to the combined rule, so an added benefit from cycle

Table 4: **Head aggregation with a fixed decoder and visibility.** All rows use Table 3's anchor-pass features. Peak weighting and the fixed head are Kubric-selected; only peak weighting receives an exponent sweep. Oracles use DAVIS annotations. The per-query oracle selects one head per complete trajectory.

Head rule	Kubric δ↑	DAVIS AJ↑	DAVIS δ↑
Uniform reference and Kubric-selected rule			
Uniform mean	66.08	42.96	58.37
Peak weighting (selected)	69.74	45.21	61.27
Other annotation-free candidates			
Entropy weighting	69.14	44.54	60.27
Margin weighting	68.01	44.09	59.86
Cycle weighting	68.60	44.33	59.91
Lowest-entropy four heads	69.26	45.42	61.54
Fixed head selected on Kubric	66.97	42.42	57.05
Retrospective oracles using DAVIS annotations			
Best fixed head (DAVIS oracle)	-	42.83	57.50
Per-video head (oracle)	-	43.54	59.50
Per-query head (oracle)	-	47.43	65.78

error is not established here. HeFT's higher Cosmos OA combines fewer false occlusions with more missed occlusions. Appendix G reports the rules, class errors and sensitivity to the synthetic calibration subset.

Feature reuse. Caching features across anchors saves computation but changes temporal context. To test whether reuse sacrifices accuracy, we compare extraction modes: DAVIS AJ falls by 1.27 [0.66, 1.91] in the original paired runs. Main results therefore retain anchor passes.

5.2 Head aggregation and feature selection

Head weighting requires a confidence signal that predicts accurate matches. Within visible query/target pairs, higher peaks correlate with lower position error (video-averaged Spearman correlation 0.432 [0.394, 0.469]). Cycle consistency has a similar association; entropy and margin are weaker.

On CogVideoX, Kubric-selected peak weighting improves DAVIS δ over uniform averaging by +2.90 [+2.24, +3.59] (Table 4). Calibrating each rule's visibility separately on Kubric gives an AJ gain of +2.28 [+1.51, +3.09]. This tests head weighting with the local decoder; we have not varied both head weighting and decoder on one shared feature extraction.

The four lowest-entropy heads score slightly higher on DAVIS but were not selected on Kubric. Only peak weighting received an exponent sweep; other rules use fixed scales, so this comparison does not rank confidence measures under equal tuning budgets.

Even the DAVIS-selected fixed head underperforms averaging, whereas annotation-based selection per query gains +4.51 [+3.42, +5.58] δ over peak weighting. This does not establish query-specific head specialisation: a simulation where heads differ only in average quality and correlated noise gives comparable gains. More analysis details are in Appendices F and J.

Block combinations. Beyond heads, combining blocks could recover complementary matches. Three CogVideoX block averages score below the selected block on Kubric. A broader sweep with uniform heads and cached features reaches 66.05 δ versus 65.94 for one block (reference minus mixture: [-0.28, 0.05]). Thus fixed averages show no clear selection-set benefit; joint decoding and refinement remain untested (more details in Appendix K).

Frequency bands. To test whether positional frequencies affect correspondence, we apply HeFT's rotary band, which restricts query/key channels to a frequency range. Selecting the band on Kubric raises Cosmos AJ by +3.09 [+2.56, +3.71] on DAVIS and +3.22 [+2.66, +3.81] on the 200-clip

Table 5: **Checkpoint size after synthetic selection.** DAVIS, 30 videos. Each checkpoint uses Kubric-selected settings. Differences are larger minus smaller, with paired 95% video intervals. Input dimensions match within each family; all use anchor passes and peak weighting.

Checkpoint pair	Small AJ↑	Large AJ↑	ΔAJ↑ [95% CI]	Δδ↑ [95% CI]
CogVideoX $2B \to 5B$	45.21	39.06	-6.15 [-8.17, -4.40]	-6.62 [-8.15, -5.22]
Wan2.1 $1.3B \to 14B$	43.64	42.33	-1.31 [-3.33, +0.61]	+0.66 [-1.42, +2.71]
Cosmos $2B \to 14B$	39.73	41.67	+1.94 [+0.08, +3.74]	+1.48 [-0.69, +3.54]

Kinetics subset, with original visibility thresholds. Recalibration raises the DAVIS gain to +3.30 [+2.70, +3.95].

Cosmos then reaches 43.03 DAVIS AJ; its advantage over HeFT remains uncertain and inputs differ. Wan shows no clear AJ gain, and the band does not apply to CogVideoX-2B's absolute positional embeddings. All channels remain the common default; the Cosmos extension is reported separately (more band-selection details in Appendix H).

5.3 Model size and tracking accuracy

Larger generators cost more to run. To test whether they provide better correspondence, Table 5 compares three checkpoint pairs with Kubric-selected settings. The cached-feature search tests every block at the same scheduler levels within each family; deeper models therefore have more candidates. Final evaluation uses anchor passes (more details in Appendix L).

CogVideoX-5B reduces AJ and position accuracy; Wan-14B's differences remain uncertain. Cosmos-14B's AJ interval only narrowly excludes zero, while its position-accuracy interval includes zero. Larger checkpoints provide no uniform benefit under this readout.

Aggregation may limit Wan-14B: its DAVIS-selected fixed head beats uniform averaging by +6.90 [+5.55, +8.40] δ in the cached-feature diagnostic. This motivates testing whether annotation-free aggregation can recover that information.

6 Limitations

Full CogVideoX selection repeats on three synthetic subsets of a 47-clip pool; only visibility calibration was checked on subsets for other backbones. Pretraining may contain benchmark videos, particularly Kinetics, so transfer does not guarantee unseen videos.

The fixed query descriptor can fail after appearance changes, and wrong matches can pass cycle consistency. Trained trackers remain more accurate on these benchmarks.

Decoder controls cover CogVideoX and Wan; head-confidence analysis covers CogVideoX. The Cosmos comparison with HeFT on Kinetics uses 200 clips and different input dimensions. Checkpoints differ in architecture and pretraining as well as size. Bootstrap intervals fix the selected settings. Richer readouts and matched end-to-end timing comparisons remain untested.

7 Conclusion

We improve zero-shot tracking from frozen video transformers, with matched gains over HeFT on CogVideoX and Wan. Controlled ablations explain the benefits of source sampling, local decoding and head aggregation, while scaling results motivate better readouts for larger generators. Our method offers a simple way to reuse video generative pretraining for point tracking, sharing computation across queries without training a separate tracking network.

Reproducibility statement

Section 3 and Appendix A specify the readout, coordinate conventions and selected configurations. Evaluation uses the benchmark metric implementation. The repository records checkpoint revisions, commands, seeds, hardware and result files, and includes scripts that generate the reported numerical macros, tables and figures. Published reference values and measured or projected costs retain separate provenance. The code and result package will be released upon publication.

AI use statement

Generative AI tools assisted implementation, experiment scripting and design, result analysis, literature checking, figure preparation, manuscript organisation, drafting and editing. The qualitative figures are rendered from saved predictions and benchmark frames; charts use recorded results, and the method diagram is a computation schematic. The authors are responsible for the scientific claims, experimental validity and final manuscript, including material developed with AI assistance.

Ethics statement

The study uses existing public benchmark videos and pretrained models, with no new data collection. Its evaluation does not establish reliable tracking outside those benchmarks or remove limitations inherited from the source data and models.

References

João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the Kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. URL https://arxiv.org/abs/1907.06987.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, Jo ao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. In *NeurIPS Datasets and Benchmarks Track*, 2022. URL https://arxiv.org/abs/2211.03726.

Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. *arXiv preprint arXiv:2306.08637*, 2023. URL https://arxiv.org/abs/2306.08637.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2203.03570.

Hyeonho Jeong, Chun-Hao P. Huang, Jong Chul Ye, Niloy J. Mitra, and Duygu Ceylan. Track4Gen: Teaching video diffusion models to track points improves video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7276–7287, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jeong_Track4Gen_Teaching_Video_Diffusion_Models_to_Track_Points_Improves_Video_CVPR_2025_paper.html.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. URL https://arxiv.org/abs/2410.11831.

Seungwoo Kim, Khai Loong Aw, Klemen Kotar, Cristobal Eyzaguirre, Wanhee Lee, Yunong Liu, Jared Watrous, Stefan Stojanov, Juan Carlos Niebles, Jiajun Wu, and Daniel L. K. Yamins. Taming

generative video models for zero-shot optical flow extraction. *arXiv preprint arXiv:2507.09082*, 2025a. URL https://arxiv.org/abs/2507.09082.

Youngseo Kim, Dohyun Kim, Geonhee Han, and Paul Hongsuck Seo. Image diffusion models exhibit emergent temporal propagation in videos. *arXiv preprint arXiv:2511.19936*, 2025b. URL https://arxiv.org/abs/2511.19936.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *arXiv preprint arXiv:2305.14334*, 2023. URL https://arxiv.org/abs/2305.14334.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. *arXiv preprint arXiv:2506.17220*, 2025. URL https://arxiv.org/abs/2506.17220.

NVIDIA. Cosmos-Predict2, 2025. URL https://github.com/nvidia-cosmos/cosmos-predict2.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. URL https://arxiv.org/abs/1704.00675.

Ayush Shrivastava and Andrew Owens. Self-supervised any-point tracking by contrastive random walks. *arXiv preprint arXiv:2409.16288*, 2024. URL https://arxiv.org/abs/2409.16288.

Ayush Shrivastava, Sanyam Mehta, Daniel Geng, and Andrew Owens. Point prompting: Counterfactual tracking with video diffusion models. *arXiv preprint arXiv:2510.11715*, 2025. URL https://arxiv.org/abs/2510.11715.

Soowon Son, Honggyu An, Jisu Nam, Hyunah Ko, Chaehyun Kim, Dahyun Chung, Siyoon Jin, Jung Yi, Junhwa Hur, and Seungryong Kim. Probing and leveraging video diffusion transformer features for robust point tracking. *arXiv preprint arXiv:2512.20606*, 2025. URL https://arxiv.org/abs/2512.20606.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02502.

Stefan Stojanov, David Wendt, Seungwoo Kim, Rahul Venkatesh, Kevin Feigelis, Jiajun Wu, and Daniel LK Yamins. Self-supervised learning of motion concepts by optimizing counterfactuals. *arXiv preprint arXiv:2503.19953*, 2025. URL https://arxiv.org/abs/2503.19953.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. URL https://arxiv.org/abs/2306.03881.

Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. *arXiv preprint arXiv:2403.14548*, 2024. URL https://arxiv.org/abs/2403.14548.

WanTeam. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. URL https://arxiv.org/abs/2408.06072.

Tianyu Yuan, Yuanbo Yang, Lin-Zhuo Chen, Yao Yao, and Zhuzhong Qian. Denoise to track: Harnessing video diffusion priors for robust correspondence. *arXiv preprint arXiv:2512.04619*, 2025. URL https://arxiv.org/abs/2512.04619.

Artem Zholus, Carl Doersch, Yi Yang, Skanda Koppula, Viorica Patraucean, Xu Owen He, Ignacio Rocco, Mehdi S. M. Sajjadi, Sarath Chandar, and Ross Goroshin. Tapnext: Tracking any point (tap) as next token prediction. *arXiv preprint arXiv:2504.05579*, 2025. URL https://arxiv.org/abs/2504.05579.

A IMPLEMENTATION DETAILS

We specify the implementation of Section 3 below; Table A1 lists the selected settings.

A.1 FRAME ENCODING AND VIDEO CHUNKS

We use the public CogVideoX, Wan2.1 and Cosmos-Predict2 checkpoints (Yang et al., 2024; Wan-Team, 2025; NVIDIA, 2025), with revisions recorded in each run. The wrappers preserve native query/key normalisation and positional encoding. Source frames follow the benchmark's 256×256 resize, then are resized bilinearly to the backbone input. Each frame is encoded separately using the VAE's deterministic latent representation and checkpoint-specific scaling. Conditioning uses the empty text prompt. The encoded latent is cached for reuse across anchor groups.

To construct chunks for anchor frame a, we remove a from the ordered frame list and sample the remaining list at stride $\max(1, \lfloor T/12 \rfloor)$. Each chunk contains the anchor followed by up to twelve sampled targets. We shift the starting position by one list entry between chunks until every target has appeared. We construct and head-aggregate maps within each chunk, then average repeated anchor-target pairs. Targets match directly to the anchor; predictions are not chained through intermediate frames.

The cached variant runs only the chunks anchored at frame zero and stores each frame's features from its first occurrence. Other query anchors reuse these features, including pairs extracted in different chunks. This saves passes but changes the temporal context relative to anchor passes.

A.2 FRAME-PAIR SCORE NORMALISATION

Equation 1 computes similarities from the captured query and key matrices. Each softmax compares a token with tokens in one other frame, whereas the generator's original attention normalises over its full token sequence. We therefore recompute the frame-pair scores rather than slice them from a full-sequence attention matrix.

The forward scores have unit row sums. The reverse scores have unit row sums before transposition; afterwards, their row sums can differ from one. Their bidirectional average and the head-weighted map should thus be read as matching scores, rather than row-normalised probabilities. The decoder normalises these scores within its local window; visibility thresholds are calibrated for the resulting score distribution.

A.3 PIXEL AND TOKEN COORDINATES

For a backbone input of height H and width W, the spatial token grid has dimensions $h = H/16$ and $w = W/16$. A query at evaluation coordinates (x, y) maps to token-centre coordinates (u, v) as

$$u = \frac{xW}{256 \cdot 16} - \frac{1}{2}, \qquad v = \frac{yH}{256 \cdot 16} - \frac{1}{2}.$$

We bilinearly interpolate score-map rows at this continuous source position, using aligned grid corners and zero padding outside the grid. Within half a token of an image boundary, zero padding can attenuate a sampled score. We retain this operation across the controlled rows and have not separately measured its effect on visibility near image borders. The decoder takes a score-weighted mean within the 5×5 window around the maximum, clipped to the token grid. A denominator floor of 10^{-12} prevents division by zero. To return to evaluation coordinates, we add $(1/2, 1/2)$, multiply by the token stride 16, and scale each axis by the ratio of evaluation size to backbone input size.

A.4 VISIBILITY AND QUERY-FRAME CONVENTIONS

The cycle-consistency check applies the same local decoder to the transpose of the already aggregated matching map. Its Euclidean return error is measured against the original query in token coordinates. When a rule uses temporal smoothing, we replicate the endpoint values before taking the three-frame median.

At the query frame, we set the predicted position to the supplied position, the matching score to one and the cycle-consistency error to zero, and force predicted visibility to true. Predictions before the query time are excluded from evaluation and qualitative displays.

B HYPERPARAMETER SELECTION

We configure each readout using synthetic annotations and transfer the settings unchanged to real videos. Table A1 lists the configurations used in the benchmark and resolution comparisons.

Table A1: **Selected readout configurations.** All rows use clean latents, anchor passes, chunks of up to 13 frames and a 5×5 decoder window. Blocks are zero-indexed; t is the scheduler timestep and β the confidence exponent. Visibility uses peak score s and cycle-consistency error e, or e alone. The combined rule smooths each cue with a three-frame median; the single-cue rule is unsmoothed. Thresholds list score then error (in token cells), selected on 20 Kubric clips. The first three configurations are used unchanged on both DAVIS and Kinetics.

Checkpoint	Input size	Block	t	β	Visibility cues	Thresholds
CogVideoX-2B	480×720	17	19	2	s and e	0.16, 1
Wan2.1-1.3B	480×832	18	129	2	e	1.5
Cosmos-2B	480×720	26	308	1	s and e	0.08, 1.75
Cosmos-2B (large input)	704×1280	26	308	1	s and e	0.06, 1.75
Wan2.1-14B	480×832	17	129	2	s and e	0.09, 1.25

B.1 SYNTHETIC DATA AND SELECTION PROCEDURE

Kubric videos (Greff et al., 2022) are exported from the MOVi-E validation split using the benchmark's point-sampling settings: 256 trajectories per clip, no random crop and no vertical flip. The configuration and visibility experiments use the first 20 clips of the fixed exported list, giving 5,120 queries. These annotations select hyperparameters; they do not train tracking parameters. The ordered selection IDs are `movi_e_val_00000` through `movi_e_val_00019`. Export uses seed zero, sequential TFDS shard reading and no shuffle buffer. This deterministic prefix is reproducible; we have not tested its independence from generation or export ordering.

Selection proceeds in stages. Initial sweeps compare transformer blocks and scheduler settings, followed by clean versus corrupted latents and alternatives to uniform head aggregation. The confidence sweep includes exponents $0.5, 1, 2, 4, 8, 16$, with zero as the uniform reference. The chosen exponents are then evaluated with anchor passes. We select extraction and aggregation settings by position accuracy.

For each resulting score distribution, visibility calibration maximises synthetic AJ over rules using the score, the cycle-consistency error or both. Single-cue rules are tested with and without three-frame median smoothing. Score-only thresholds range from 0 to 1 in steps of 0.01; cycle-consistency error thresholds range from 0 to 8 token cells in steps of 0.25. The combined rule uses separately smoothed signals and requires both inequalities strictly. Its score thresholds range from 0.03 to 0.20 in steps of 0.01, and its cycle-consistency error thresholds from 0.5 to 4 in steps of 0.25. No threshold is selected on DAVIS or Kinetics.

B.2 BLOCK AND SCHEDULER SEARCH

Figure 3 and Table A2 summarise 1,098 distinct block/scheduler settings evaluated on the same synthetic clips. To make these comparisons interpretable, they retain only runs with clean latents, cached features, uniform heads and 480×720 input. Later decoder, temperature and resolution variants are considered separately.

The horizontal axis identifies the scheduler setting by its requested log-SNR. We convert it to an integer timestep using each model's schedule: $\log[\bar{\alpha}_t/(1-\bar{\alpha}_t)]$ for CogVideoX, $2\log[(1-s)/s]$ for Wan, with $s = 3u/(1+2u)$ and $u = t/1000$, and $-2\log[u/(1-u)]$ for Cosmos. CogVideoX uses the checkpoint's cumulative signal schedule $\bar{\alpha}_t$. The conversion searches the schedule and rounds the resulting timestep; requested and attained levels can differ. Because the input latents remain clean, this axis does not specify noise added to an image. The actual timestep used by each checkpoint is retained in the source manifest. Search coverage differs across checkpoints, including finer sweeps for CogVideoX, so each maximum describes that model's measured selection

set. Positional encoding and pretraining can also differ between checkpoints; these maxima do not isolate an effect of parameter count.

An earlier translation probe tests matching between a frame and a copy with a known spatial shift. This diagnostic checks whether the readout recovers the shift; the headline results instead use the annotation-based selection procedure above.

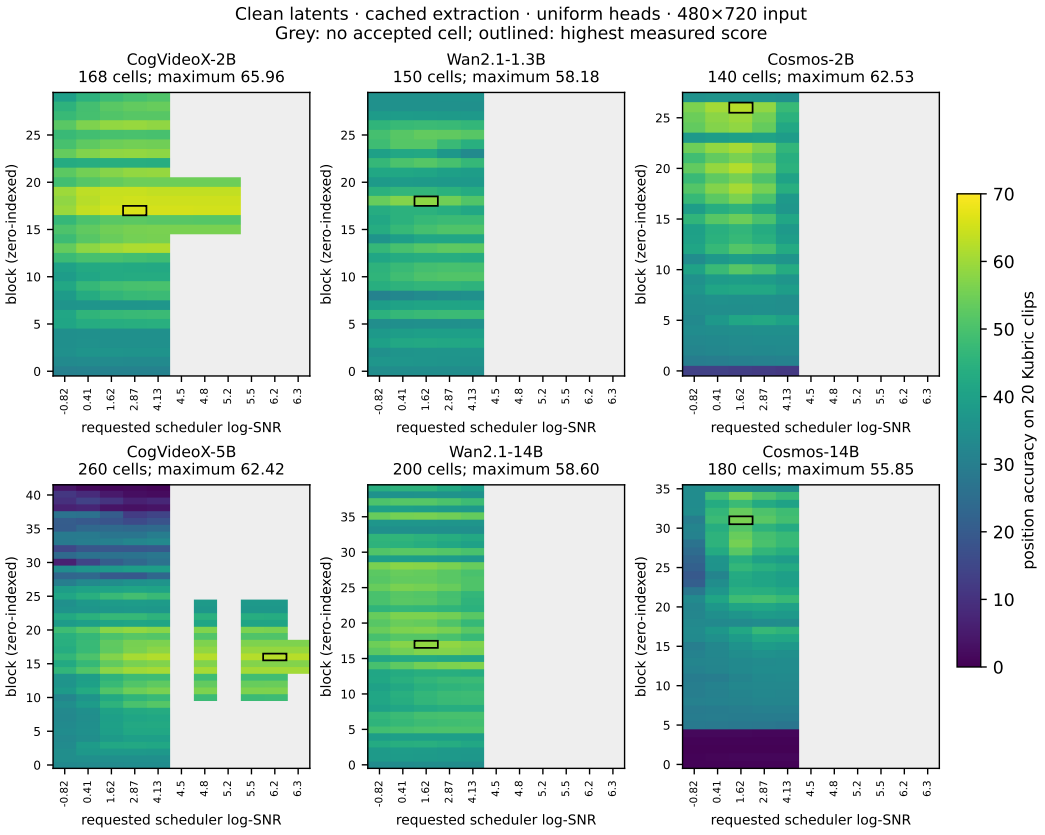


Figure 3: **Block and scheduler search on synthetic videos.** Position accuracy on the same 20 Kubric clips, using clean latents, cached features, uniform heads and 480×720 input. All panels share an absolute colour scale; grey cells are unmeasured in this comparison. Outlines mark each panel's highest measured score. The horizontal axis gives the requested scheduler setting, with no noise added to the input. Search coverage differs between checkpoints.

Table A2: **Original synthetic block and scheduler search.** The same 20 Kubric clips, 480×720 input, clean latents, cached features and uniform heads. Each maximum is over that checkpoint's measured cells in Figure 3. Unequal search coverage makes these selection summaries rather than a controlled scaling comparison.

Checkpoint	Configurations	Best block	Best log-SNR	Best Kubric $\delta\uparrow$
CogVideoX-2B	168	17	2.87	65.96
CogVideoX-5B	260	16	6.20	62.42
Wan2.1-1.3B	150	18	1.62	58.18
Wan2.1-14B	200	17	1.62	58.60
Cosmos-2B	140	26	1.62	62.53
Cosmos-14B	180	31	1.62	55.85

Table A3 extends the search to every block at scheduler levels available to both sizes within each family. This includes nine levels for CogVideoX and five for Wan and Cosmos. CogVideoX-2B

Table A3: **Model-size comparison with matched scheduler coverage.** The same 20 Kubric clips, 480×720 input, clean latents, cached features and uniform heads. Both sizes are evaluated at every block and the same scheduler levels within each family. Deeper checkpoints therefore have more configurations. Best scores are synthetic selection maxima, not held-out test results or estimates of parameter-count effects alone.

Checkpoint	Blocks	Levels	Configurations	Best block	Best log-SNR	Best $\delta\uparrow$
CogVideoX-2B	30	9	270	17	2.87	65.96
CogVideoX-5B	42	9	378	16	5.25	62.28
Wan2.1-1.3B	30	5	150	18	1.62	58.18
Wan2.1-14B	40	5	200	17	1.62	58.60
Cosmos-2B	28	5	140	26	1.62	62.53
Cosmos-14B	36	5	180	31	1.62	55.85

Table A4: **Repeating joint selection on additional synthetic clips.** CogVideoX-2B: a common grid of 30 blocks and nine scheduler levels, followed by seven confidence exponents and visibility selection, all using only the named 20-clip subset. Every subset selects block 17, requested log-SNR 2.87 ($t = 20$), $\beta = 2$, and the median-smoothed combined visibility rule with score/cycle thresholds $(0.16, 1.00)$. Clip numbers index the fixed MOVi-E export. S1 and S2 each exclude S0 but share 13 clips with one another. DAVIS evaluates all 30 videos and 650 queries with unchanged selected settings. Brackets give 95% video intervals. Grid scores use cached features and uniform heads; final selection and DAVIS use anchor passes.

Subset	Clip numbers	Grid $\delta\uparrow$	Anchor $\delta\uparrow$	DAVIS AJ↑ [95% CI]	DAVIS $\delta\uparrow$	OA↑
S0	0 to 19	65.96	69.75	45.33 [39.71, 51.18]	61.35	76.20
S1	20 to 39	69.09	73.14	45.33 [39.71, 51.18]	61.35	76.20
S2	27 to 46	68.39	72.33	45.33 [39.71, 51.18]	61.35	76.20

cannot reach the additional high-log-SNR levels available to 5B; those levels are excluded from this comparison. Deeper checkpoints still have more blocks and hence more evaluated configurations. The resulting maxima compare a common extraction and search procedure, not parameter count alone: positional encoding and pretraining also differ.

B.3 Repeating selection on additional synthetic clips

Table A4 repeats CogVideoX-2B block, timestep, confidence-exponent and visibility selection. S0 uses the original 20 clips, S1 uses the next 20, and S2 uses clips 27 to 46 in the same fixed MOVi-E export. Both additional sets exclude S0; they share 13 clips with one another and therefore are not independent replications of each other. Each subset uses the same 270-cell block/scheduler grid, then selects the confidence exponent from $0, 0.5, 1, 2, 4, 8, 16$ and visibility from its own anchor-pass predictions. All settings are fixed before DAVIS evaluation.

Every subset selects block 17, $t = 20$, $\beta = 2$, and the smoothed combined visibility rule with thresholds $(0.16, 1.00)$. Their DAVIS predictions are identical, reaching 45.33 AJ. The common grid's requested log-SNR 2.87 maps to $t = 20$; the main method retains $t = 19$ from its earlier search. This control tests repeatability within the same 47-clip pool and candidate grid, rather than invariance to arbitrary synthetic data. It does not repeat selection for Wan or Cosmos or change the main benchmark configuration.

C Evaluation protocols and reference results

We compute 95% percentile bootstrap intervals from 10,000 video resamples with seed zero. Comparisons resample the same videos for both methods. These intervals condition on the queries and selected configurations; they do not include synthetic-calibration uncertainty or numerical variation across runs. Exploratory intervals are not adjusted for multiple comparisons.

Reference executions. The main benchmark, decoder and head tables share CogVideoX predictions with 45.21 AJ and 61.27 δ. Earlier feature-reuse, source-resolution, calibration-subset, direction and error-stratification controls, and the qualitative examples, retain an earlier execution: 45.31 AJ and 61.29 δ at the same settings. These are numerical differences between executions, not different deployed configurations. Each comparison uses its own saved reference; we never pair aggregate scores from different runs.

TAP-Vid supplies point annotations and evaluation protocols for videos from the DAVIS 2017 validation set (Pont-Tuset et al., 2017) and Kinetics-700 validation set (Carreira et al., 2019). Both tracking subsets and their metrics are introduced in the TAP-Vid paper (Doersch et al., 2022).

C.1 DAVIS AND THE DIFFTRACK REPRODUCTION

The TAP-Vid first evaluation on DAVIS uses all 650 queries. The frame-zero subset contains 538 queries and is used only for the DiffTrack reproduction and explicitly labelled additional experiments. Table A5 distinguishes the published DiffTrack score, our execution of the released code and our reimplementation. The reimplementation uses posterior-mode VAE latents and the documented decoder coordinate convention; the released pipeline samples its VAE posterior. This reproduction provides a reference for the subsequent changes to localisation and aggregation.

Table 2 uses this reimplementation on all first-visible queries. It retains uniform head averaging, bidirectional query-key matching and DiffTrack's decoder: find the target argmax for each source token, then bilinearly sample that position field at the query using DiffTrack's coordinate convention. Queries after frame zero use our anchor-pass extraction. To obtain AJ and OA, we add score and cycle visibility checks, selecting their thresholds separately on Kubric for this readout. Thus this row compares an adapted DiffTrack readout with our method; it is not the unmodified released tracker.

For orientation, benchmark, decoder and confidence tables use anchor passes and the selected peak rule, except the explicitly varied rows. The block/timestep grid and the older cross-model head diagnostics use cached features and uniform heads. The size study first selects a grid cell with that cached readout, then evaluates anchor passes with checkpoint-specific confidence and visibility. Each paired comparison uses the reference predictions identified in its caption.

Table A5: **DiffTrack reproduction on DAVIS frame-zero queries.** All rows concern the 538 frame-zero queries. The official-code result averages its logged, rounded per-video scores. Our reimplementation uses posterior-mode VAE latents and DiffTrack's released decoder.

Implementation	Position accuracy $\delta\uparrow$
DiffTrack (published)	46.30
Released DiffTrack implementation	46.21
DiffTrack readout reimplementation	46.47

C.2 HEFT REPRODUCTION AND FIRST-MODE ADAPTATION

We use the released HeFT repository at revision `040959e4` (full revision retained in the run manifests). The initial reproduction executes its scripts with unchanged defaults: frame-zero queries, the query frame included in evaluation, and incomplete final 25-frame chunks discarded. Its CogVideoX AJ is 38.45, 1.09 below the published 39.54, outside our predefined tolerance of ± 1.0 AJ. We do not tune the method to close that gap. Subsequent features evaluated with HeFT's own evaluator give 38.56, 39.55 and 47.77 AJ for CogVideoX, Wan and Cosmos; the published values are 39.54, 41.27 and 48.61. The protocol table below uses these subsequent executions. Reproduction differences and protocol differences must therefore be considered separately.

For first-visible queries, we group queries by anchor frame, start the released tracking loop there and stop its backward walk at the anchor. Search, descriptor updates, interpolation and consistency checks retain released settings. We repeat the last input frame to complete the final chunk, then trim tracks to the true length. On every DAVIS video and all three backbones, frame-zero positions and visibility exactly match the original tracker on the same padded features. This verifies the adapter, not equivalence to extracting a truncated video.

Table A6: **Both readouts under HeFT and TAP-Vid evaluation on DAVIS.** Published scores are copied from HeFT. The remaining rows score saved predictions. The release protocol scores frame-zero queries, drops incomplete final chunks and includes the query frame. TAP-Vid first uses all 650 first-visible queries, full video lengths and scores strictly after the query with the official metric. Release scores use HeFT's unchanged evaluator and its 255-coordinate normalisation. This controls evaluation, not feature extraction after truncation. All entries are AJ↑; native preprocessing and matched dimensions are retained. Difference intervals are paired over 30 videos.

Score source / evaluation	CogVideoX-2B	Wan2.1-1.3B	Cosmos-2B
HeFT (published)	39.54	41.27	48.61
HeFT (release protocol)	38.56	39.55	47.77
Ours (release protocol)	50.85	48.08	44.95
Ours - HeFT (release)	+12.28 [+8.19, +16.51]	+8.53 [+5.95, +11.20]	-2.82 [-6.21, +0.75]
HeFT (TAP-Vid first)	32.11	34.33	41.47
Ours (TAP-Vid first)	45.21	43.64	39.93
Ours - HeFT (first)	+13.10 [+9.28, +17.29]	+9.31 [+6.91, +11.70]	-1.54 [-4.81, +1.82]

Both methods are scored with the same vendored TAP-Vid code on identical query annotations and frames. HeFT converts a normalised horizontal query x to $x(W-1)$ model pixels; ours uses xW. This difference can approach one model pixel. Each method's output is converted back through its own convention. Common annotations therefore do not imply identical feature sampling. HeFT uses direct Lanczos4 resizing; ours uses the benchmark resize followed by bilinear resizing. For all three backbone controls in Table A7, HeFT instead receives exactly our float-valued input frames after both resizes, without its own Lanczos operation. We pass queries at $xW-0.5$ in HeFT's pixel-index convention and map output index p back through $(p+0.5)/W$. This gives identical physical query locations. The adapter's frame-zero parity check still passes. Changing HeFT's input changes AJ by -0.32 [-1.07, +0.52] on CogVideoX, -1.14 [-1.95, -0.42] on Wan and -1.75 [-3.02, -0.66] on Cosmos. Our CogVideoX and Wan AJ advantages remain +13.43 [+9.66, +17.59] and +10.45 [+8.27, +12.65], respectively. Matching the inputs therefore does not explain either advantage. The Cosmos AJ difference becomes +0.20 [-2.78, +3.27], still uncertain, while its OA difference is -8.48 [-13.73, -4.04]. HeFT's Cosmos visibility advantage therefore persists with identical inputs.

HeFT's authors selected its head and frequency settings on DAVIS; our method uses Kubric. The comparison does not isolate selection-dataset choice. Published HeFT values are in Table 1; the standard TAP-Vid first comparisons are in Table 2. *Published* denotes a score copied from HeFT's paper; *release protocol* denotes our execution scored with HeFT's evaluator; *TAP-Vid first* denotes our execution scored with the official TAP-Vid first-query metric. Table A6 makes these score sources explicit. We also score our saved predictions using HeFT's unchanged evaluator on the same 538 frame-zero queries and complete 25-frame chunks, including the query frame and using its 255-coordinate normalisation. The CogVideoX and Wan AJ advantages remain +12.28 [+8.19, +16.51] and +8.53 [+5.95, +11.20]; the Cosmos difference remains uncertain. Every HeFT per-video score is reproduced before scoring ours. This isolates the evaluation choice, not the effect of re-extracting features after truncation. The official metric and query samplers are vendored without modification from TAP-Net revision `730cda1c`. The numerical reproduction gap is reported separately from the change in evaluation protocol.

On Kinetics, Table A8 compares both methods on identical clips and queries. Both CogVideoX and Wan cover all 1,144 recovered clips and 32,267 first-mode queries. Under HeFT's own evaluator and released protocol, CogVideoX obtains 25.82 AJ on the recovered clips, below the published 29.33. Their overlap with HeFT's original clips is unverified; we do not attribute this difference solely to the evaluator. Wan likewise obtains 26.35 under HeFT's evaluator versus the published 30.06. These remaining shortfalls of about 3.5 and 3.7 AJ could reflect the source clips, the unreleased Kinetics driver or execution differences; their cause is unresolved. The CoTracker3 check below does not validate HeFT's implementation. Our claim concerns the released code evaluated on our recovered clips, not an exact reproduction of HeFT's published Kinetics result.

Cosmos covers the same seeded 200-clip subset used by the other subset controls, with 5,589 first-mode queries. Our default run uses 480×720 inputs; HeFT uses its released 704×1280 setting. The AJ difference is +7.78 [+6.69, +8.87]. This comparison controls clips and queries, but not input

Table A7: **HeFT with identical model inputs and query locations.** DAVIS, all 30 videos and 650 queries, TAP-Vid first. Each shared-input row gives HeFT our resized frames and maps queries to the same physical positions. The last column is our AJ minus each HeFT row, with paired 95% intervals.

Method / preprocessing	AJ↑	δ↑	OA↑	Ours - HeFT AJ↑ [95% CI]
CogVideoX-2B, 480×720				
HeFT (native)	32.11	47.44	77.99	+13.10 [+9.28, +17.29]
HeFT (shared input)	31.79	47.01	77.88	+13.43 [+9.66, +17.59]
Ours	45.21	61.27	76.14	-
Wan2.1-1.3B, 480×832				
HeFT (native)	34.33	48.94	79.04	+9.31 [+6.91, +11.70]
HeFT (shared input)	33.19	48.05	77.16	+10.45 [+8.27, +12.65]
Ours	43.64	57.42	79.22	-
Cosmos-2B, 704×1280				
HeFT (native)	41.47	58.07	79.19	-1.54 [-4.81, +1.82]
HeFT (shared input)	39.73	56.82	79.00	+0.20 [-2.78, +3.27]
Ours	39.93	56.37	70.52	-

Table A8: **Kinetics comparisons on identical clips and queries under TAP-Vid first.** Within each group, both methods use identical recovered clips and queries, backbone checkpoints and the official metric. Preprocessing is native to each method. CogVideoX and Wan match input dimensions and cover all 1,144 recovered clips and 32,267 queries. Cosmos covers a seeded 200-clip subset (5,589 queries) and uses different input dimensions, shown below. The last column is ours minus HeFT, with paired 95% intervals.

Method	AJ↑	δ↑	OA↑	Ours - HeFT AJ↑ [95% CI]
CogVideoX-2B, 480×720*, all 1,144 clips*				
HeFT (TAP-Vid first)	25.37	42.15	71.64	-
Ours	44.00	59.53	79.08	+18.62 [+18.06, +19.20]
Wan2.1-1.3B, 480×832*, all 1,144 clips*				
HeFT (TAP-Vid first)	26.21	43.23	71.20	-
Ours	38.98	53.10	79.55	+12.77 [+12.29, +13.23]
Cosmos-2B, seeded 200-clip subset				
HeFT (704×1280)	28.05	47.21	68.98	-
Ours (480×720)	35.83	50.90	75.37	+7.78 [+6.69, +8.87]

dimensions; it does not establish a result on all recovered Kinetics clips. HeFT's frame-zero parity check is exact on all 197 subset videos with frame-zero queries; the remaining three have none. Its own-evaluator score on this subset is not directly comparable with its published full-set score.

Table A9 adds the adapted DiffTrack readout on 200 clips drawn without using performance (seed 0) from the recovered Kinetics set. All three readouts use the same 5589 first-mode queries. Our AJ advantage over the adapted DiffTrack readout is +8.97 [+8.03, +9.93]. This subset evaluation does not fill its missing full-set Table 1 cells.

C.3 Transfer to RGB-Stacking

Table A10 evaluates the same CogVideoX-2B configuration on a synthetic manipulation domain without selection on its annotations. The two readout rows change only visibility at fixed positions. CoTracker3 uses its scaled offline checkpoint, with window length 60, 384×512 internal resizing and the released visibility-confidence product threshold of 0.6. Its primary row processes all queries jointly, as in our other CoTracker3 evaluations. The single-point check processes one query at a time with an 8×8 support grid. Both use the same 50 videos, 1,498 queries and metric as our readout. CoTracker3 is ahead on every video; single-point evaluation changes its AJ by only +0.11 [-0.54, +0.77]. These results establish transfer of our fixed configuration, but no generalisation advantage over the trained tracker on this domain.

Table A9: **Three readouts on the same seeded Kinetics subset.** CogVideoX-2B, 480×720, 200 clips and 5589 first-mode queries, seed 0. This subset was drawn without using performance from the 1,144 recovered clips. All rows use the official TAP-Vid metric and identical queries, with native preprocessing. The adapted DiffTrack readout has uniform heads, its argmax position-field decoder and Kubric-selected visibility. These are complete-system comparisons. Brackets give 95% video intervals, paired for differences; subset scores must not be compared with full-set rows.

Method	AJ↑ [95% CI]	δ↑ [95% CI]	OA↑ [95% CI]
HeFT (TAP-Vid first)	22.91 [21.27, 24.56]	39.83 [38.03, 41.63]	69.20 [67.09, 71.34]
DiffTrack readout + visibility	31.83 [29.80, 33.88]	42.00 [39.87, 44.10]	75.76 [73.75, 77.74]
Ours	40.80 [38.32, 43.26]	56.43 [53.95, 58.83]	77.80 [76.12, 79.40]
Ours - DiffTrack readout	+8.97 [+8.03, +9.93]	+14.43 [+13.39, +15.51]	+2.04 [+1.01, +3.11]
Ours - HeFT	+17.90 [+16.54, +19.20]	+16.60 [+15.25, +17.89]	+8.60 [+6.46, +10.80]

Table A10: **RGB-Stacking transfer and a trained reference on identical queries.** All rows use the same 50 synthetic robot-manipulation videos, 1,498 first-mode queries and official TAP-Vid metric. For our CogVideoX-2B readout, the DAVIS/Kinetics configuration is unchanged: block 17, $t = 19$, $\beta = 2$, anchor passes, 480×720, local 5×5 decoding. The selected visibility rule uses Kubric thresholds 0.16 and 1.0. Intervals resample videos. CoTracker3 uses its scaled offline checkpoint and released evaluation settings; joint queries are the primary comparison. Only the two readout rows share positions. Differences have paired intervals.

Method / setting	AJ↑ [95% CI]	δ↑ [95% CI]	OA↑ [95% CI]
Ours, always visible	51.48 [48.52, 54.48]	73.49 [71.41, 75.53]	80.95 [77.92, 83.77]
Ours, Kubric visibility	54.21 [51.33, 57.03]	73.49 [71.41, 75.53]	77.70 [75.65, 79.74]
CoTracker3, joint	74.49 [72.02, 76.84]	85.00 [83.15, 86.75]	91.54 [90.35, 92.67]
CoTracker3, single point	74.60 [72.21, 76.89]	85.34 [83.59, 87.01]	92.16 [91.05, 93.23]
Ours - CoTracker3 joint	-20.28 [-22.39, -18.24]	-11.51 [-12.76, -10.24]	-13.85 [-15.47, -12.24]
CoTracker3 single - joint	+0.11 [-0.54, +0.77]	+0.34 [-0.00, +0.71]	+0.62 [+0.09, +1.14]

C.4 Kinetics reconstruction and validation

We recover 1,144 of the 1,147 labelled clips. The available archive lacks `BBSK3Wv0jXM_000112_000122`, `G4Z1Ug34B5I_000224_000234` and `pk1yi_HMsAI_000053_000063`; no fallback copy was recovered. Source identifiers and temporal alignment are recorded in the reconstruction manifest. For each complete backbone run, we check all 70 expected shard partitions: video names, query counts, saved predictions and consistent extraction and decoding settings. The three backbone runs cover the same recovered clips and use their unchanged synthetic visibility rules.

A CoTracker3 offline check obtains 55.06 AJ, compared with the published 54.70 (Karaev et al., 2024). We process benchmark queries jointly, while the paper evaluates each query separately with support points. The similar scores provide a basic pipeline check, but do not establish identical tracking protocols or source-video reconstruction.

C.5 Published comparisons

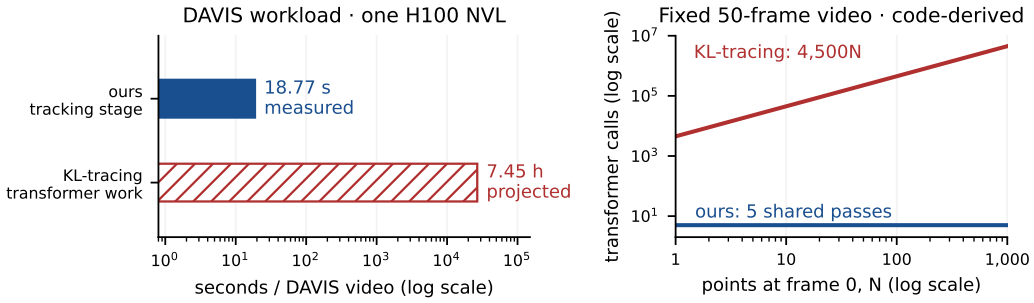
Table A11 identifies the source of each published benchmark entry. Our recovered Kinetics clip set and HeFT's published evaluation are not verified to be identical. Point Prompting reports a 30-clip Kinetics sample; Opt-CWM values are taken from HeFT's reference table, which does not specify their clip count. DiffTrack's DAVIS and Kinetics entries also come from separate reported evaluations.

Table A11: **Sources and scope of published benchmark results.** Provenance of the reference entries in Table 1. Values come from the listed publications; evaluations run by us are reported separately.

Entry	Source	Scope
HeFT	Table 1, v2 (Yuan et al., 2025)	Kinetics clip count unstated
CoTracker3, TAPNext	HeFT Table 1, v2	reference values from that table
DiTracker	Table 3, v2 (Son et al., 2025)	trained CoTracker3 head + DiT
Opt-CWM, DIFT	HeFT Table 1, v2	clip count not specified there
DiffTrack, DAVIS	original paper (Nam et al., 2025)	frame-zero queries
DiffTrack, Kinetics	HeFT Table 1, v2	separate reported evaluation
Point Prompting	Table 1, v2 (Shrivastava et al., 2025)	one query/video; 30 Kinetics clips
KL-tracing	Table 3 (Kim et al., 2025a)	DAVIS first mode

D COMPUTATIONAL COST

The cost comparison combines measured tracking-stage time, measured transformer-call time and workload counts. This section specifies each quantity and derives the projections in Figure 4.



Forward medians on H100 NVL: CogVideoX 0.419 s; LRAS 0.230 s. Decoding is additional point-dependent work.

Figure 4: **Shared inference and repeated query-conditioned calls.** Left: our measured tracking stage and projected KL-tracing transformer work for the DAVIS workload on the same H100 NVL. LRAS uses its released architecture with random weights; this is not an end-to-end comparison. Right: counted transformer calls for a 50-frame video with all queries at frame zero. Decoding grows with query count, and new anchors can add shared passes.

D.1 TIMING MEASUREMENTS

The headline timer starts after input resizing and ends after CUDA synchronisation, before final predictions are copied to the host. It includes VAE encoding, transformer passes, score-map construction, head weighting, position decoding and cycle-consistency checks. It excludes dataset/model loading, resizing, final host copies, separate visibility smoothing and thresholding, metric evaluation and output serialisation.

Table A12 reports this tracking stage separately from individual transformer calls. Call measurements use bfloat16 on one H100 NVL, with 15 repetitions after warm-up. We report their median and observed range. For KL-tracing, we instantiate its LRAS backbone with random weights at the released call shape: 6,552 tokens, including 922 target tokens. This measures the architecture's call cost without reproducing its tracking predictions. The forward-only median is 0.230 seconds; including mask construction gives 0.262 seconds.

Table A12: **Measured timings on one H100 NVL.** The tracking stage covers encoding through position decoding and cycle-consistency checks; timer boundaries are specified in the text. Transformer calls use bfloat16 and 15 timed repetitions after warm-up. LRAS uses random weights in the released architecture. The KL-tracing rows are individual call measurements, not full-tracker runtimes.

Workload	Statistic	Seconds↓	Call range (s)↓
ours, DAVIS video	mean over 30 videos	18.77	-
CogVideoX, 13-frame call	median	0.419	0.418-0.421
LRAS, cached-mask call	median	0.230	0.214-0.241
LRAS, mask creation + call	median	0.262	0.260-0.270

D.2 KL-TRACING WORKLOAD PROJECTION

The DAVIS workload contains 30 videos, 1,999 frames, 650 queries and 38,881 point-frame entries. Entries include the query frames, following KL-tracing's preprocessing. Its released inference loop uses nine zoom evaluations, five mask rollouts and two predictive calls per entry, giving 90 transformer calls per entry and 3,499,290 calls in total.

Multiplying this count by the measured forward-only median and dividing by the number of videos gives 7.45 hours of transformer work per average video. Including mask construction gives 8.49 hours. These projections omit quantisation, perturbation preparation, KL-map computation, remaining orchestration and I/O. They quantify transformer work on the measured hardware, rather than complete end-to-end tracking latency.

D.3 DEPENDENCE ON QUERY COUNT

The right panel of Figure 4 fixes a separate workload: a 50-frame video with every query at frame zero. We run the chunk construction function to count the shared transformer passes and plot that count as the number of points N increases. For a video of T frames, KL-tracing uses $90NT$ calls in this setting, including the query frame. The panel compares transformer-call counts, not total runtime.

For our readout, fixing the video, resolution and anchors fixes the shared encoding, transformer execution and score-map construction. Each added point requires row interpolation, local decoding, a cycle-consistency check and visibility processing. This point-dependent work grows with N, while the per-point share of extraction falls as more points reuse it. Adding new anchor frames can increase the shared work as well. The computational advantage therefore comes from reusing inference, without requiring a perfectly flat total-runtime curve.

D.4 OPERATION COUNTS AND FEATURE CACHING

Table A13 reports call counts and measured per-call FLOPs. The headline readout dispatches 1,040 transformer passes: 1,033 contain 13 frames and seven contain 12. Applying the measured 13-frame cost to every call approximates the transformer total. VAE encoding is counted per distinct video frame and shared across anchors.

The cached variant uses 277 transformer passes on the same videos. Its saved runtime was measured on a different GPU, so the comparison establishes a reduction in calls rather than a matched latency improvement. We likewise do not assign these per-call costs to HeFT or Point Prompting, whose tokenisation, input shapes and inference procedures would need separate measurements.

Table A13: **Operation accounting for the DAVIS workload.** Readout call counts are recorded in the runs; the KL count follows the released preprocessing and inference loops. Per-call FLOPs count dispatched operations, not every operation in a tracking pipeline. The CogVideoX call cost is for 13 frames; seven anchor-pass calls have only 12. VAE encoding is separate and shared across anchors.

Operation	Calls	TFLOP / call↓
CogVideoX, anchor passes	1,040	120.00
CogVideoX, cached extraction	277	120.00
CogVideoX VAE, one frame	1,999	4.78
LRAS, 6,552 tokens	3,499,290	107.44

E CONTROLLED LOCALISATION EXPERIMENTS

Tables A14 and A15 report the full set of component combinations underlying Table 3. All variants share one feature extraction per video, descriptors, head weights, block, timestep, queries and visibility predictions. For each coordinate convention we test nearest and bilinear source-row sampling and windows of 1×1, 3×3 and 5×5 tokens. The one-token case returns argmax; larger windows average existing scores without another softmax or a fitted temperature. Bidirectional score averaging is held fixed in the decoder controls; Appendix E.1 tests it separately.

The DiffTrack rows use the coordinate mapping in its released decoder. Source queries at model-pixel coordinates (x_m, y_m) map to normalised sampling coordinates

$$g_x = 2\frac{x_m/16}{w-m} - 1, \qquad g_y = 2\frac{y_m/16}{h-m} - 1, \qquad m = \frac{W}{64 \cdot 16}.$$

Here W is the model-input width and h, w are the token-grid dimensions. DiffTrack converts decoded target token coordinates (u, v) to model pixels as $(16u, 16v)$. The token-centre convention instead samples source rows at $(x_m/16 - 1/2, y_m/16 - 1/2)$, normalised by $(w-1, h-1)$, and returns target pixels $(16(u+1/2), 16(v+1/2))$. For example, target token index zero maps to pixel zero under DiffTrack's convention and pixel eight under the token-centre convention, in each axis. Table 3 switches both source and target mappings together; these are coordinate changes within the same readout. Two additional controls change only one side's convention. Both are worse here, showing why a consistent source/target convention matters.

The position-field controls first decode a target argmax at each source token, then interpolate the resulting positions at the query. They separate the stage of interpolation from its use. The rounding control starts from the current local prediction and rounds it to the nearest corrected token centre. The nearest-ground-truth-token row uses annotated positions; it gives the best possible position accuracy when outputs are restricted to token centres.

Tables 1 to 4 use the same CogVideoX-2B DAVIS reference predictions, including their visibility masks. This shared extraction gives 61.27 δ; an earlier execution of the same configuration gave 61.29. We recompute the HeFT and DiffTrack differences and visibility controls from the shared reference. Earlier feature-reuse, source-resolution, calibration-subset and error-stratification studies retain their original predictions and references. Kubric scores are selection-set diagnostics. DAVIS measures transfer of the fixed interventions.

Replication on Wan. Tables A16 and A17 repeat the decoder controls on Wan2.1-1.3B at its original selected configuration. Each variant uses the same extracted features and the reference's actual visibility mask, not just the same visibility thresholds. The local mean improves DAVIS δ by +11.18 [+9.40, +13.00] over argmax with bilinear source sampling and token-centre coordinates fixed. Rounding removes 10.06 [8.53, 11.64] points; the source-sampling interaction is +7.91 [+6.32, +9.62]. Thus the sub-token explanation and the interaction both extend to a second backbone. This new paired extraction gives 57.40 DAVIS δ, compared with 57.42 in the earlier benchmark run. All decoder differences use the new shared extraction; the benchmark tables retain their original reference. No hyperparameters are reselected.

Table A14: **Complete decoder controls.** Same features and visibility throughout. DiffTrack denotes its coordinate convention, not its complete readout. Field rows decode each source token before interpolating positions; all other rows interpolate scores before decoding. The final row uses annotations and only bounds token-centre outputs.

Decoder variant	Kubric $\delta\uparrow$	DAVIS AJ$\uparrow$	DAVIS $\delta\uparrow$
Nearest, DiffTrack, 1×1	43.79	28.23	40.84
Nearest, DiffTrack, 3×3	47.54	31.02	44.18
Nearest, DiffTrack, 5×5	47.60	30.86	44.09
Bilinear, DiffTrack, 1×1	47.04	30.45	43.35
Bilinear, DiffTrack, 3×3	58.21	37.07	51.57
Bilinear, DiffTrack, 5×5	59.19	37.70	52.42
Nearest, token centres, 1×1	47.66	32.52	45.40
Nearest, token centres, 3×3	51.99	36.32	49.79
Nearest, token centres, 5×5	51.81	35.83	49.14
Bilinear, token centres, 1×1	50.34	34.15	47.71
Bilinear, token centres, 3×3	68.17	44.93	60.96
Bilinear, token centres, 5×5	69.74	45.21	61.27
Interpolate argmax field, DiffTrack coordinates	49.82	31.66	45.09
Interpolate argmax field, centres	53.35	36.93	51.12
DiffTrack source, centre target (argmax)	42.73	28.19	40.31
Centre source, DiffTrack target (argmax)	37.36	23.95	34.95
Round current prediction to token centre	51.86	35.10	48.85
GT nearest target token (diagnostic)	63.40	41.42	63.35

Table A15: **Controlled changes in decoder position accuracy.** Source interpolation compares bilinear sampling with the nearest source token. Intervals use paired video bootstrap resampling. The interaction subtracts the nearest-source local-decoding gain from the bilinear-source gain. These contrasts do not form an additive attribution.

Change	Kubric $\Delta\delta\uparrow$ [95% CI]	DAVIS $\Delta\delta\uparrow$ [95% CI]
Source interpolation (argmax, token centres)	+2.68 [+2.33, +2.99]	+2.31 [+1.78, +2.83]
Source interpolation (local mean, token centres)	+17.93 [+16.20, +19.63]	+12.13 [+10.11, +14.30]
Local mean vs argmax (bilinear source, token centres)	+19.41 [+17.51, +21.26]	+13.56 [+11.31, +15.95]
Interaction: window size and source sampling	+15.26 [+13.54, +16.96]	+9.82 [+7.84, +12.06]
Local mean vs its rounded prediction	+17.89 [+16.19, +19.54]	+12.42 [+10.20, +14.78]
Rounded local mean vs argmax	+1.52 [+1.15, +1.87]	+1.15 [+0.71, +1.62]
Full decoder vs interpolated argmax (DiffTrack coordinates)	+19.92 [+18.10, +21.63]	+16.18 [+13.86, +18.72]
Interpolate scores vs positions (token centres)	-3.02 [-5.43, -0.72]	-3.42 [-6.16, -0.92]
Interpolate scores vs positions (DiffTrack coordinates)	-2.78 [-4.47, -1.08]	-1.74 [-3.28, -0.15]

E.1 Matching direction

Equation 1 averages forward query-key scores and transposed backward query-key scores. Tables A18 and A19 replace this average with either direction alone on all three backbones. The selected block, timestep, input size, head-weighting rule and local decoder remain fixed. Confidence weights are computed from each direction's maps, so the intervention measures the effect of matching direction within the current aggregation rule.

We distinguish three visibility controls. Copying the bidirectional reference's actual masks to each variant isolates position changes in AJ. Applying the original thresholds to a variant's new scores and cycles also changes visibility; this is a separate control. Finally, reselecting visibility on that variant's Kubric predictions evaluates the recalibrated system. All six single-direction variants have lower DAVIS AJ in every comparison, with paired intervals below zero. Five also have lower position accuracy; backward-only Wan has a change of +1.06 [-0.04, +2.20]. Its AJ still falls with fixed masks and after recalibration.

The reference predictions come from earlier executions at the same settings, rather than a shared extraction with the variants. Their CogVideoX score is 45.31 AJ; the main comparison retains 45.21 from the later common reference. The intervals describe variation across videos, not numerical variation across devices.

Table A16: **Decoder controls repeated on Wan2.1-1.3B.** Block 18, $t = 129$, $\beta = 2$, clean latents, anchor passes and 480×832 input. All variants share features, queries and the reference's predicted visibility mask (cycle threshold 1.5); no settings are reselected. The 20 Kubric clips and 30 DAVIS videos are the same as in the CogVideoX control. DiffTrack denotes the coordinate formula in Appendix E, not a released Wan tracker. The final row uses annotations.

Decoder variant	Kubric $\delta\uparrow$	DAVIS AJ↑	DAVIS $\delta\uparrow$
Nearest, DiffTrack, 1×1	40.84	27.87	40.12
Nearest, DiffTrack, 3×3	43.90	30.54	43.39
Nearest, DiffTrack, 5×5	43.77	30.39	43.16
Bilinear, DiffTrack, 1×1	43.73	29.65	41.82
Bilinear, DiffTrack, 3×3	53.71	36.17	50.22
Bilinear, DiffTrack, 5×5	54.12	37.09	51.38
Nearest, token centres, 1×1	44.87	32.37	44.76
Nearest, token centres, 3×3	48.68	35.85	48.93
Nearest, token centres, 5×5	48.48	35.20	48.03
Bilinear, token centres, 1×1	46.77	34.56	46.21
Bilinear, token centres, 3×3	59.89	43.59	57.41
Bilinear, token centres, 5×5	60.33	43.61	57.40
Interpolate argmax field, DiffTrack coordinates	45.59	29.30	42.48
Interpolate argmax field, centres	47.98	33.82	47.31
DiffTrack source, centre target (argmax)	38.57	26.86	38.31
Centre source, DiffTrack target (argmax)	35.29	25.35	34.70
Round current prediction to token centre	48.46	35.56	47.34
GT nearest target token (diagnostic)	64.91	44.93	64.74

Table A17: **Controlled decoder differences on Wan2.1-1.3B.** Definitions match Table A15; 95% intervals use paired video resampling. Every contrast uses the shared extraction in Table A16. Bilinear source sampling and local decoding again interact.

Change	Kubric $\Delta\delta\uparrow$ [95% CI]	DAVIS $\Delta\delta\uparrow$ [95% CI]
Source interpolation (argmax, token centres)	+1.90 [+1.54, +2.24]	+1.45 [+0.97, +1.95]
Source interpolation (local mean, token centres)	+11.85 [+10.55, +13.18]	+9.37 [+7.57, +11.30]
Local mean vs argmax (bilinear source, token centres)	+13.56 [+12.11, +14.98]	+11.18 [+9.40, +13.00]
Interaction: window size and source sampling	+9.95 [+8.84, +11.11]	+7.91 [+6.32, +9.62]
Local mean vs its rounded prediction	+11.87 [+10.60, +13.13]	+10.06 [+8.53, +11.64]
Rounded local mean vs argmax	+1.69 [+1.36, +2.01]	+1.12 [+0.53, +1.74]
Full decoder vs interpolated argmax (DiffTrack coordinates)	+14.74 [+12.67, +16.65]	+14.91 [+12.77, +17.01]
Interpolate scores vs positions (token centres)	-1.20 [-3.03, +0.62]	-1.10 [-3.34, +1.11]
Interpolate scores vs positions (DiffTrack coordinates)	-1.87 [-3.48, -0.28]	-0.67 [-2.61, +1.24]

Table A18: **Forward, backward and bidirectional matching.** Original blocks, timesteps, input sizes, clean latents, anchor passes, head-weighting rules and local decoder are fixed. Head weights are computed from each variant's maps. Kubric uses 20 selection clips; DAVIS uses 30 videos and 650 first-mode queries. DAVIS AJ and OA use each variant's Kubric-selected visibility rule. Bidirectional references retain the earlier executions used by this study; the CogVideoX reference is 45.31 AJ, rather than the later shared extraction in Table 1.

Matching direction	Kubric $\delta\uparrow$	DAVIS AJ↑	DAVIS $\delta\uparrow$	DAVIS OA↑
CogVideoX-2B, 480×720				
Bidirectional	69.72	45.31	61.29	76.19
Forward only	68.74	44.13	60.07	76.07
Backward only	66.90	43.13	58.96	78.18
Wan2.1-1.3B, 480×832				
Bidirectional	60.34	43.64	57.42	79.22
Forward only	52.52	39.74	50.19	76.49
Backward only	60.46	41.23	58.48	79.94
Cosmos-2B, 480×720				
Bidirectional	63.84	39.73	54.83	74.64
Forward only	53.35	35.36	48.77	75.56
Backward only	54.35	34.11	48.19	75.38

Table A19: **Separating position and visibility effects of matching direction.** Each entry is the single direction minus its bidirectional reference, with a paired 95% video interval. Fixed masks copy the reference's predicted visibility to the variant positions, isolating position changes in AJ. Retained thresholds apply the original rule to the variant's own scores and cycles; recalibrated rules are selected on its Kubric predictions. Position accuracy is independent of these masks. All settings other than matching direction and the stated visibility intervention are fixed. References and variants were extracted in separate runs.

Backbone	Direction	$\Delta\delta\uparrow$	ΔAJ↑: fixed masks	ΔAJ↑: retained thresholds	ΔAJ↑: recalibrated
CogVideoX-2B	Forward	-1.22 [-1.87, -0.55]	-0.81 [-1.33, -0.31]	-1.05 [-1.93, -0.20]	-1.18 [-1.98, -0.42]
CogVideoX-2B	Backward	-2.34 [-2.96, -1.71]	-1.79 [-2.36, -1.23]	-2.52 [-3.82, -1.22]	-2.18 [-3.30, -1.08]
Wan2.1-1.3B	Forward	-7.23 [-8.22, -6.25]	-7.30 [-8.43, -6.23]	-3.88 [-4.93, -2.75]	-3.90 [-4.99, -2.76]
Wan2.1-1.3B	Backward	+1.06 [-0.04, +2.20]	-1.59 [-2.14, -1.07]	-4.51 [-5.40, -3.68]	-2.41 [-3.62, -1.19]
Cosmos-2B	Forward	-6.06 [-7.34, -4.83]	-4.63 [-5.74, -3.58]	-4.33 [-5.51, -3.21]	-4.38 [-5.52, -3.31]
Cosmos-2B	Backward	-6.64 [-8.09, -5.17]	-5.14 [-6.42, -3.90]	-5.78 [-7.18, -4.40]	-5.62 [-6.94, -4.34]

F CONFIDENCE STATISTICS AND DEPLOYABLE HEAD RULES

Figure 5 relates head confidence to correctness; Table A20 reports the annotation-free rules and oracle diagnostics. For a score row S_h, let $z_h = \sum_j S_h(j)$ and $p_h = S_h/z_h$. Peak confidence is $\max_j S_h(j)$. Entropy is $H_h = -\sum_j p_h(j) \log p_h(j) / \log M$. We save both the raw top-two difference and the difference divided by z_h; the deployable margin rule uses the latter. Cycle error applies the same local forward and reverse decoder to each head and measures return distance in token coordinates.

Weights are proportional to squared peak score, $\exp(-H_h/0.3)$, normalised margin, or $\exp(-e_h)$, respectively, then normalised across heads. Peak exponents were swept on Kubric; entropy, margin and cycle weighting use single fixed scales. Their comparison is not a ranking after equal hyperparameter-search budgets. None is fitted on DAVIS. Top-k rules equally average the best $k \in \{1, 2, 4\}$ heads according to each statistic. All deployable rules compute weights at source tokens before continuous query interpolation and before repeated-pair averaging, as in the current readout.

The confidence analysis measures statistics at continuous query locations after repeated-pair averaging. It correlates confidence ranks with negative Euclidean position error within each visible query/target pair, averages within videos and bootstraps videos. Lower entropy and cycle error count as higher confidence. The correlation describes the relation between confidence and correctness; the aggregation experiment tests whether using that confidence improves tracks.

Peak weighting and the deployable fixed head are selected by Kubric position accuracy. Retrospective choices use DAVIS annotations to select a fixed head, a head for each video, or one head for each complete query trajectory. Results use the official metric's pooling of visible positions within videos. They differ from the earlier equal-query, cached-feature diagnostic in Appendix J. All controlled rows use the peak rule's predicted visibility mask. A separate full-system check selects visibility independently on Kubric. Peak weighting gives 45.21 AJ, 61.27 δ and 76.14 OA, versus 42.94, 58.37 and 76.41 for uniform averaging. Paired peak-minus-uniform differences are +2.28 [+1.51, +3.09] AJ, +2.90 [+2.24, +3.59] δ and -0.27 [-1.15, +0.62] OA. The localisation and AJ gains persist; the OA difference remains uncertain. No benchmark annotation enters a deployable selection rule. The head gain in Table 4 is conditional on the local decoder. The adapted DiffTrack row uses uniform heads from a separate extraction; comparing it with the peak-weighted position-field control does not form a shared-feature head-by-decoder factorial experiment.

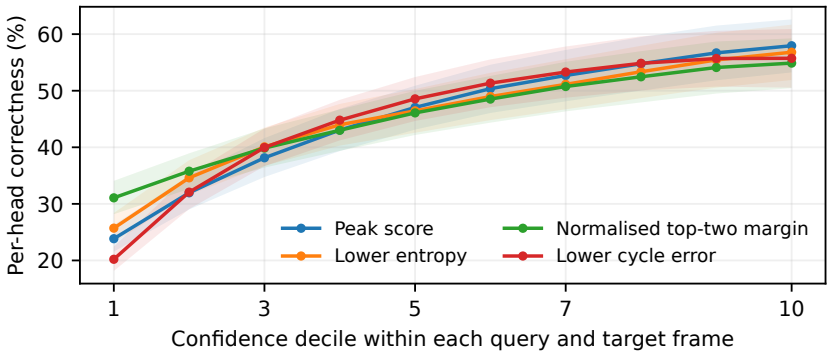


Figure 5: **Confidence and per-head correctness.** CogVideoX-2B on DAVIS. Heads are ranked within each visible query/target pair and grouped into confidence deciles, controlling query difficulty. The vertical axis averages correctness over the five TAP-Vid distance thresholds, then over videos. Shading gives 95% video bootstrap intervals. Lower entropy and cycle error are treated as higher confidence. These are retrospective measurements of annotation-free statistics; annotations are not supplied to the aggregation rules.

Table A20: **All annotation-free head rules and oracle diagnostics.** DAVIS video means and 95% video bootstrap intervals. All rows use the same visibility predictions.

Rule	AJ↑	δ↑ [95% CI]	OA↑
Cycle weighting	44.33	59.91 [54.90, 64.89]	76.14
Cycle top 1	42.06	56.99 [51.78, 62.18]	76.14
Cycle top 2	44.03	59.29 [54.02, 64.50]	76.14
Cycle top 4	45.12	60.76 [55.47, 65.97]	76.14
Entropy weighting	44.54	60.27 [55.28, 65.23]	76.14
Entropy top 1	43.35	59.03 [54.33, 63.63]	76.14
Entropy top 2	44.54	60.55 [55.79, 65.24]	76.14
Lowest-entropy four heads	45.42	61.54 [56.67, 66.35]	76.14
Margin weighting	44.09	59.86 [54.98, 64.69]	76.14
Margin top 1	41.18	56.20 [51.71, 60.63]	76.14
Margin top 2	42.30	57.80 [53.29, 62.23]	76.14
Margin top 4	43.30	58.92 [54.28, 63.50]	76.14
Peak weighting (selected)	45.21	61.27 [56.26, 66.23]	76.14
Peak top 1	42.78	58.48 [53.95, 62.92]	76.14
Peak top 2	43.73	59.64 [55.00, 64.18]	76.14
Peak top 4	44.65	60.67 [55.87, 65.40]	76.14
Uniform mean	42.96	58.37 [53.65, 62.98]	76.14
Fixed head selected on Kubric	42.42	57.05 [51.88, 62.14]	76.14
Retrospective oracles using DAVIS annotations			
Best fixed head (DAVIS oracle)	42.83	57.50 [52.34, 62.61]	76.14
Per-video head (oracle)	43.54	59.50 [54.91, 64.17]	76.14
Per-query head (oracle)	47.43	65.78 [61.41, 70.15]	76.14

G VISIBILITY CONTROLS AND CALIBRATION SENSITIVITY

Tables A22, A23 and A24 hold positions fixed and vary visibility. Rows with retained thresholds use the current rule's values for available cues; for a cue absent from the current rule, we use its Kubric-selected single-cue threshold. Rows with reselected thresholds maximise each rule's synthetic AJ. This separates a change to the rule from a change to its calibration. Smoothing uses a three-frame median with replicated endpoints. Delta is identical for all rules within a model and benchmark.

Within each video, the visible-to-occluded error rate is the fraction of annotated visible positions predicted as occluded. The occluded-to-visible rate is the fraction of annotated occluded positions

Table A21: **Visibility at fixed CogVideoX positions.** Each deployable row selects thresholds on Kubric. DAVIS δ stays at 61.27; Kinetics δ stays at 59.53. False occlusion is the fraction of visible points predicted occluded; missed occlusion is the fraction of occluded points predicted visible. Rates (%) average videos containing that class. Median uses three frames.

	DAVIS				Kinetics
Visibility rule	AJ↑	OA↑	False occlusion ↓	Missed occlusion ↓	AJ↑
Always visible	40.13	75.13	0.00	100.00	36.32
Peak score	44.82	75.89	23.03	30.51	43.53
Peak score + median	45.29	76.49	22.53	30.00	43.86
Cycle consistency	42.39	76.47	14.03	52.19	40.44
Cycle + median	42.51	76.81	13.37	52.86	40.59
Score + cycle	45.06	76.41	22.17	31.06	43.71
Score + cycle + median	45.21	76.14	23.69	27.57	44.00
Selected method	45.21	76.14	23.69	27.57	44.00
Oracle visibility	50.70	100.00	0.00	0.00	48.60

predicted as visible. We average each rate over videos containing positions of that class. OA instead measures the fraction of all evaluated positions classified correctly within each video, then averages over videos. Because the proportion of visible positions varies between videos, the two averaged error rates alone do not determine OA. We save the underlying counts and the number of videos contributing to each rate.

Table A25 applies current, always-visible and annotated visibility to saved HeFT and readout positions. The entries using predicted visibility reproduce the common scorer's per-video metrics to numerical tolerance. This differs from removing a visibility check inside HeFT: that change can also alter search and descriptor updates, and hence positions.

Table A26 repeats only visibility selection. A seed-zero permutation partitions the original 20 Kubric clips into disjoint groups at each size. Groups at different sizes overlap; size 20 has only one group. Each selected rule transfers unchanged to DAVIS and Kinetics. The expanded candidate set can select a different full-set rule from the original deployment rule, as for Wan. Specifically, the original combined rule always smoothed both cues; this ablation also allows their unsmoothed conjunction. We report that conditional study separately rather than replacing the benchmark setting. Appendix B.3 separately repeats joint block, timestep, confidence and visibility selection on additional CogVideoX calibration clips.

G.1 Results at fixed positions

Table A21 changes visibility while keeping positions fixed. Each deployable rule selects its thresholds on Kubric. For CogVideoX, the selected combined rule improves AJ over declaring every point visible. Score alone with median smoothing gives a similar DAVIS AJ point estimate; the data do not establish that adding cycle consistency is necessary for this configuration.

Replacing predicted visibility with annotated visibility raises CogVideoX AJ to 50.70 on DAVIS and 48.60 on Kinetics: gains of +5.48 [+3.84, +7.31] and +4.60 [+4.31, +4.90], respectively. This diagnostic isolates the effect of correcting visibility at fixed positions. It is not a general upper bound on AJ, because suppressing an incorrect visible prediction can sometimes increase that metric.

For Cosmos at matched input dimensions, HeFT labels fewer visible points as occluded (12.87% versus 31.64%), but misses more actual occlusions (54.23% versus 26.64%). Its higher OA therefore combines opposing class-error differences. Giving both methods annotated visibility leaves an AJ difference of -2.38 [-5.71, +1.15]. This result does not establish whether visibility accounts for the Cosmos AJ difference. The following tables give all backbones and both benchmarks.

Calibration sensitivity. Holding features, positions and confidence weighting fixed, we repeat visibility selection on disjoint subsets of the original synthetic clips. With one clip, CogVideoX DAVIS AJ ranges from 41.31 to 45.47; with five clips, it ranges from 44.77 to 45.47. Table A26

Table A22: **Visibility rules and threshold selection: CogVideoX-2B.** Retained thresholds come from the selected method, or from the Kubric single-cue selection if that cue is absent from the method. Reselected thresholds maximise Kubric AJ for that rule. Positions remain fixed. Median uses three frames. False and missed occlusion are class-conditional error rates (%), defined in Table A21. Oracle visibility uses benchmark annotations. δ is 61.27 on DAVIS and 59.53 on Kinetics.

		DAVIS				Kinetics			
Visibility rule	Thresholds	AJ↑	OA↑	False occl. ↓	Missed occl. ↓	AJ↑	OA↑	False occl. ↓	Missed occl. ↓
Always visible	-	40.13	75.13	0.00	100.00	36.32	68.14	0.00	100.00
Selected method	Original	45.21	76.14	23.69	27.57	44.00	79.08	23.24	27.46
Peak score	Retained	44.82	75.89	23.03	30.51	43.53	78.65	22.54	30.39
Peak score	Reselected	44.82	75.89	23.03	30.51	43.53	78.65	22.54	30.39
Peak score + median	Retained	45.29	76.49	22.53	30.00	43.86	79.07	22.40	29.59
Peak score + median	Reselected	45.29	76.49	22.53	30.00	43.86	79.07	22.40	29.59
Cycle consistency	Retained	42.35	77.66	8.60	66.41	39.85	76.30	8.15	61.89
Cycle consistency	Reselected	42.39	76.47	14.03	52.19	40.44	76.56	12.49	52.57
Cycle consistency + median	Retained	42.30	77.74	7.78	68.61	39.89	76.61	7.55	62.13
Cycle consistency + median	Reselected	42.51	76.81	13.37	52.86	40.59	77.01	12.08	52.15
Score + cycle	Retained	44.77	75.53	24.40	27.82	43.69	78.66	23.51	28.01
Score + cycle	Reselected	45.06	76.41	22.17	31.06	43.71	79.17	21.47	30.58
Score + cycle + median	Retained	45.21	76.14	23.69	27.57	44.00	79.08	23.24	27.46
Score + cycle + median	Reselected	45.21	76.14	23.69	27.57	44.00	79.08	23.24	27.46
Oracle visibility	-	50.70	100.00	0.00	0.00	48.60	100.00	0.00	0.00

Table A23: **Visibility rules and threshold selection: Wan2.1-1.3B.** Retained thresholds come from the selected method, or from the Kubric single-cue selection if that cue is absent from the method. Reselected thresholds maximise Kubric AJ for that rule. Positions remain fixed. Median uses three frames. False and missed occlusion are class-conditional error rates (%), defined in Table A21. Oracle visibility uses benchmark annotations. δ is 57.42 on DAVIS and 53.10 on Kinetics.

		DAVIS				Kinetics			
Visibility rule	Thresholds	AJ↑	OA↑	False occl. ↓	Missed occl. ↓	AJ↑	OA↑	False occl. ↓	Missed occl. ↓
Always visible	-	36.87	75.13	0.00	100.00	31.39	68.14	0.00	100.00
Selected method	Original	43.64	79.22	16.60	37.29	38.98	79.55	17.91	32.74
Peak score	Retained	38.93	75.89	12.61	62.46	34.83	72.04	21.94	46.21
Peak score	Reselected	38.93	75.89	12.61	62.46	34.83	72.04	21.94	46.21
Peak score + median	Retained	38.93	75.85	11.55	65.51	35.07	72.31	21.79	45.79
Peak score + median	Reselected	38.93	75.85	11.55	65.51	35.07	72.31	21.79	45.79
Cycle consistency	Retained	43.64	79.22	16.60	37.29	38.98	79.55	17.91	32.74
Cycle consistency	Reselected	43.64	79.22	16.60	37.29	38.98	79.55	17.91	32.74
Cycle consistency + median	Retained	43.50	80.76	15.22	36.20	39.10	80.49	17.40	31.21
Cycle consistency + median	Reselected	43.49	80.40	16.34	34.47	39.20	80.18	18.78	29.02
Score + cycle	Retained	43.62	77.75	22.20	27.99	38.77	76.56	28.70	20.10
Score + cycle	Reselected	43.97	79.44	17.40	34.34	39.48	79.34	21.01	27.61
Score + cycle + median	Retained	43.50	79.13	20.18	29.07	38.96	77.18	28.06	19.67
Score + cycle + median	Reselected	43.69	80.80	15.76	34.41	39.53	80.12	20.18	27.15
Oracle visibility	-	46.19	100.00	0.00	0.00	41.95	100.00	0.00	0.00

gives all subset sizes and backbones. This tests visibility calibration conditional on the selected features, not the robustness of block, timestep or confidence-exponent selection.

Table A24: **Visibility rules and threshold selection: Cosmos-2B.** Retained thresholds come from the selected method, or from the Kubric single-cue selection if that cue is absent from the method. Reselected thresholds maximise Kubric AJ for that rule. Positions remain fixed. Median uses three frames. False and missed occlusion are class-conditional error rates (%), defined in Table A21. Oracle visibility uses benchmark annotations. δ is 54.83 on DAVIS and 54.70 on Kinetics.

		DAVIS				Kinetics			
Visibility rule	Thresholds	AJ↑	OA↑	False occl. ↓	Missed occl. ↓	AJ↑	OA↑	False occl. ↓	Missed occl. ↓
Always visible	-	35.18	75.13	0.00	100.00	33.16	68.14	0.00	100.00
Selected method	Original	39.73	74.64	26.46	25.13	39.63	77.58	23.04	31.73
Peak score	Retained	39.32	74.44	25.83	27.10	39.28	77.25	22.12	34.71
Peak score	Reselected	39.32	74.44	25.83	27.10	39.28	77.25	22.12	34.71
Peak score + median	Retained	39.88	75.04	25.49	25.82	39.51	77.64	22.02	33.96
Peak score + median	Reselected	39.88	75.04	25.49	25.82	39.51	77.64	22.02	33.96
Cycle consistency	Retained	36.73	75.96	7.89	71.42	35.24	72.89	7.39	72.40
Cycle consistency	Reselected	36.79	74.83	12.72	61.49	35.91	73.49	11.43	62.14
Cycle consistency + median	Retained	36.61	76.25	6.32	75.32	35.20	73.02	6.49	73.54
Cycle consistency + median	Reselected	36.76	75.27	11.20	64.00	35.96	73.82	10.61	62.63
Score + cycle	Retained	39.16	73.89	27.21	25.23	39.41	77.13	23.39	32.21
Score + cycle	Reselected	39.49	75.53	22.87	30.92	39.02	77.55	19.75	38.06
Score + cycle + median	Retained	39.73	74.64	26.46	25.13	39.63	77.58	23.04	31.73
Score + cycle + median	Reselected	39.73	74.64	26.46	25.13	39.63	77.58	23.04	31.73
Oracle visibility	-	44.13	100.00	0.00	0.00	43.66	100.00	0.00	0.00

Table A25: **Visibility diagnostics for HeFT and our readout on DAVIS.** Predicted positions are held fixed within each method. Inputs have matched dimensions (Cosmos: 704×1280) but retain native preprocessing. Both methods use TAP-Vid first. Oracle visibility uses annotations. Error rates (%) use each method's predicted visibility and the definitions in Table A21.

		AJ↑ with indicated visibility			Error rate (%)↓	
Backbone	Method	Predicted	All visible	Oracle	False occlusion ↓	Missed occlusion ↓
CogVideoX-2B	HeFT	32.11	29.51	36.97	13.17	55.92
CogVideoX-2B	Ours	45.21	40.13	50.70	23.69	27.57
Wan2.1-1.3B	HeFT	34.33	30.97	38.56	15.35	46.14
Wan2.1-1.3B	Ours	43.64	36.87	46.19	16.60	37.29
Cosmos-2B	HeFT	41.47	37.27	46.67	12.87	54.23
Cosmos-2B	Ours	39.93	35.81	44.29	31.64	26.64

Table A26: **Sensitivity of visibility selection to the synthetic subset.** Mean and range of real-video AJ after selection, not confidence intervals. Groups are disjoint within each size and drawn from the original 20 clips; different sizes overlap. Block, timestep, confidence exponent and positions remain fixed. CogVideoX uses the earlier reference execution described in Appendix E.

Backbone	Clips per subset	Subsets	DAVIS AJ↑ [min, max]	Kinetics AJ↑ [min, max]
CogVideoX-2B	1	20	44.47 [41.31, 45.47]	43.36 [41.90, 44.00]
CogVideoX-2B	5	4	45.24 [44.77, 45.47]	43.96 [43.86, 44.00]
CogVideoX-2B	10	2	45.36 [45.31, 45.40]	44.00 [43.99, 44.00]
CogVideoX-2B	20	1	45.31 [45.31, 45.31]	44.00 [44.00, 44.00]
Wan2.1-1.3B	1	20	43.50 [38.54, 43.97]	39.01 [34.80, 39.48]
Wan2.1-1.3B	5	4	43.87 [43.75, 43.97]	39.39 [39.35, 39.48]
Wan2.1-1.3B	10	2	43.96 [43.95, 43.97]	39.46 [39.43, 39.48]
Wan2.1-1.3B	20	1	43.97 [43.97, 43.97]	39.48 [39.48, 39.48]
Cosmos-2B	1	20	38.78 [32.07, 39.94]	38.81 [36.07, 39.71]
Cosmos-2B	5	4	39.71 [39.65, 39.76]	39.55 [39.26, 39.66]
Cosmos-2B	10	2	39.76 [39.65, 39.87]	39.44 [39.21, 39.66]
Cosmos-2B	20	1	39.73 [39.73, 39.73]	39.63 [39.63, 39.63]

H FREQUENCY BANDS AND HEFT COMPONENTS

H.1 TRANSFERRING HEFT'S FEATURE CHOICES

Table A27 tests HeFT components in our Cosmos readout at 704×1280. The key-key row replaces query-key descriptors; the block/timestep row uses block 18 and $t = 2$ together. The frequency band retains HeFT's $(0.1, 1.0)$ channel range. HeFT selected this range on DAVIS, so it is a benchmark-informed diagnostic. Each row recalibrates visibility on Kubric because the matching-score distribution changes.

The band improves AJ by +5.11 [+3.99, +6.37]. Key-key descriptors instead change AJ by -4.33, and HeFT's block and timestep change it by -30.90. Combining the descriptors, band, block and timestep gives 20.50 AJ, below the original 39.93. These effects depend on the readout in which the features are used; a favourable result for one component does not justify transferring all of them.

Table A28 separates block and timestep on Kubric using cached features and uniform heads. Switching from $t = 308$ to $t = 2$ reduces accuracy at either block, while changing block at a fixed timestep has a smaller effect. Input latents remain clean: this tests timestep conditioning, not the effect of adding noise. Its extraction settings differ from the preceding anchor-pass experiment, so the changes are not additive.

H.2 BAND SELECTION USING ONLY SYNTHETIC ANNOTATIONS

For Wan and Cosmos, we select the lower band edge from $\{0, 0.05, 0.1, 0.15, 0.2, 0.3\}$ on the same 20 Kubric clips, with upper edge fixed at 1 (Table A29). These fractions index the rotary channel pairs using HeFT's convention, separately for the temporal and two spatial axes. Zero retains all channels. Matching uses the retained query/key channels and their dimension for score normalisation. Block, timestep, confidence exponent, extraction, input size and decoder remain fixed. CogVideoX-2B has no rotary positional embedding, so this intervention does not apply to it.

The selected lower edges are 0.05 for Cosmos and 0.15 for Wan. Their Kubric gains are +0.87 and +0.41 δ, respectively. For each chosen band, we select visibility thresholds on its own Kubric predictions and transfer them unchanged to DAVIS. Table A30 reports both the original visibility thresholds and those reselected on Kubric for the band. With thresholds retained, Cosmos gains +3.09 [+2.56, +3.71] AJ. Recalibration adds +0.21 [-0.12, +0.51] at unchanged positions. The AJ gain therefore persists with the original thresholds. After recalibration, Cosmos gains +3.30 [+2.70, +3.95] AJ, with an increase on every video. Wan has no clear AJ change (-0.04 [-0.71, +0.57]) and its position accuracy changes by -1.15 [-1.93, -0.47]. Synthetic gains therefore do not predict the direction of real-video localisation changes for both backbones.

The same Cosmos band transfers to the seeded 200-clip Kinetics subset (Table A31). Retaining the original visibility thresholds gives +3.22 [+2.66, +3.81] AJ; transferring the thresholds reselected on Kubric gives +3.24 [+2.69, +3.82]. Recalibration contributes +0.02 [-0.07, +0.12] at fixed positions. The position gain is +1.91 [+1.49, +2.33]. This supports transfer of the band effect beyond DAVIS without attributing it to visibility recalibration.

Table A32 compares the band variants with released HeFT under TAP-Vid first. Cosmos uses 480×720 input while HeFT uses 704×1280. Its AJ difference changes from -1.74 without the band to +1.56 [-1.45, +4.69] with the band. The interval includes zero, so the higher point estimate does not establish an advantage over HeFT. Wan retains an AJ lead of +9.27 [+6.81, +11.71], and CogVideoX retains its unmodified comparison in Table 2. The band remains an ablation, separate from the proposed method.

H.3 REMOVING COMPONENTS FROM HEFT

Table A33 changes one component at a time inside HeFT while retaining its other released settings. Query-key descriptors reduce AJ on all three backbones; descriptor averaging improves Wan and Cosmos. Removing the band changes Cosmos AJ by -1.19 [-1.85, -0.56] but has no clear effect on Wan. Removing the backward-consistency visibility check lowers AJ on all three backbones. This check also controls search and descriptor updates, so the intervention changes positions as well

Table A27: **Transferring HeFT components into our Cosmos readout.** DAVIS, 704×1280 input; each row recalibrates visibility on Kubric. The frequency band is HeFT's DAVIS-selected value, so rows containing it are diagnostics. The block/timestep row changes both settings together. AJ differences are relative to the unchanged readout, with paired video intervals.

Change to our readout	AJ↑	δ↑	OA↑	AJ↑ change [95% CI]
Current readout	39.93	56.37	70.52	-
Key-key descriptors	35.60	52.09	73.27	-4.33 [-6.43, -2.45]
HeFT frequency band (DAVIS-selected)	45.04	60.15	76.11	+5.11 [+3.99, +6.37]
HeFT block and timestep	9.03	17.65	75.13	-30.90 [-36.47, -25.64]
Key-key + band + block/timestep	20.50	34.48	59.94	-19.43 [-23.86, -15.22]

Table A28: **Separating block choice from timestep conditioning on Cosmos.** Kubric position accuracy δ, with clean latents, cached extraction, uniform heads and 480×720 input. No noise is added to the input; t changes the transformer's conditioning. Block 26 and $t = 308$ are our choices; block 18 and $t = 2$ are HeFT's. This extraction differs from the anchor-pass component comparison in Table A27.

Block	$t = 308$: δ↑	$t = 2$: δ↑
26	62.53	42.76
18	60.68	42.44

as visibility; Table A25 separately holds positions fixed. Removing the local search restriction has paired AJ intervals containing zero on all three backbones.

For Cosmos, unrestricted search uses HeFT's existing global-search path because a frame-covering radius exceeds the local kernel's grid limit. A Wan control gives a global-minus-local difference of +0.00 [+0.00, +0.01] AJ. No component setting is selected on DAVIS here.

Table A29: **Selecting the rotary frequency band on Kubric.** Position accuracy $\delta\uparrow$ on the same 20 synthetic clips, with the original block, timestep, confidence exponent, anchor passes and decoder fixed. The upper band edge is 1; columns give the lower edge in HeFT's channel-index convention. Zero retains all channels. Bold entries are selected by synthetic position accuracy. Cosmos uses 480×720 input and Wan 480×832.

Backbone	0	0.05	0.1	0.15	0.2	0.3
Cosmos-2B	63.84	**64.71**	63.97	63.37	63.27	62.21
Wan2.1-1.3B	60.34	60.35	60.70	**60.75**	59.17	47.66

Table A30: **Transfer of Kubric-selected frequency bands to DAVIS.** All 30 videos, TAP-Vid first. Bands are selected on Kubric. Retained visibility thresholds come from the original method; reselected thresholds use the band's Kubric predictions. Both rules transfer unchanged to DAVIS. Positions are identical between these two band rows. AJ differences are relative to all channels, with paired video intervals. These are controlled band ablations; the default method in Table 1 retains all channels.

Channels	Visibility thresholds	AJ↑	$\delta\uparrow$	OA↑	AJ↑ change [95% CI]
Cosmos-2B, 480×720					
All channels	Original	39.73	54.83	74.64	-
Selected band	Retained	42.83	57.44	77.14	+3.09 [+2.56, +3.71]
Selected band	Reselected	43.03	57.44	78.47	+3.30 [+2.70, +3.95]
Wan2.1-1.3B, 480×832					
All channels	Original	43.64	57.42	79.22	-
Selected band	Retained	43.60	56.27	78.86	-0.04 [-0.71, +0.57]
Selected band	Reselected	43.60	56.27	78.86	-0.04 [-0.71, +0.57]

Table A31: **Transfer of the selected Cosmos band to Kinetics.** Cosmos-2B, 480×720, the same seeded 200 clips and 5589 first-mode queries as Table A9. The lower band edge 0.05 and both visibility settings were selected on Kubric. The retained row uses the original thresholds (0.08, 1.75); the reselected row uses (0.07, 1.25). All other settings are fixed. The band rows have identical positions. Differences are relative to all channels, with paired 95% video intervals; this is an ablation, not a change to the default method.

Channels	Visibility	AJ↑	$\delta\uparrow$	OA↑	ΔAJ↑ [95% CI]
All channels	Original	35.83	50.90	75.37	-
Selected band	Retained	39.05	52.81	78.35	+3.22 [+2.66, +3.81]
Selected band	Reselected	39.07	52.81	79.21	+3.24 [+2.69, +3.82]

Table A32: **Band ablations compared with HeFT under TAP-Vid first.** DAVIS: identical videos, queries, checkpoints and metric implementation; native preprocessing. Cosmos input dimensions differ, as shown. Differences are band minus HeFT with paired 95% intervals. CogVideoX-2B has no rotary band and retains the result in Table 2.

Backbone	Band input	HeFT input	Band AJ↑	HeFT AJ↑	ΔAJ↑ [95% CI]
Cosmos-2B	480×720	704×1280	43.03	41.47	+1.56 [-1.45, +4.69]
Wan2.1-1.3B	480×832	480×832	43.60	34.33	+9.27 [+6.81, +11.71]

Table A33: **Removing components from HeFT.** Change in DAVIS AJ relative to the complete released tracker under TAP-Vid first; paired 95% intervals over 30 videos. Native inputs are 480×720 (CogVideoX), 480×832 (Wan) and 704×1280 (Cosmos). Removing the backward check also changes search and descriptor updates, so positions are not fixed. Cosmos uses HeFT's global search path for the final row. CogVideoX has no rotary band.

Change to HeFT (ΔAJ↑)	CogVideoX-2B	Wan2.1-1.3B	Cosmos-2B
Key-key to query-key	-11.32 [-13.68, -9.07]	-9.40 [-11.23, -7.63]	-2.37 [-3.45, -1.29]
Remove frequency band	-	+0.44 [-0.64, +1.52]	-1.19 [-1.85, -0.56]
Remove descriptor averaging	-0.54 [-1.46, +0.44]	-2.62 [-3.93, -1.43]	-3.25 [-4.61, -1.88]
Remove backward visibility check	-4.10 [-5.82, -2.62]	-5.27 [-7.96, -3.04]	-6.34 [-9.30, -3.87]
Remove local search restriction	+0.12 [-0.60, +0.78]	-0.34 [-1.18, +0.34]	-0.47 [-1.51, +0.52]

I WHERE CORRESPONDENCE ERRORS OCCUR

Tables A34 and A35 stratify saved predictions by annotated displacement from the query, time since the query, and reappearance after occlusion. Bins are specified before analysis and do not select deployment settings. A reappearance begins at an annotated occluded-to-visible transition after the query and ends at the next occlusion. We report the first three visible frames and compare them with visible frames without prior occlusion.

Accuracy is lower at large displacements and longer intervals from the query, and on the frames immediately after reappearance. These are descriptive comparisons between different populations, not causal estimates of occlusion effects. The uniform-head DAVIS comparison uses the current visibility mask but separately extracted historical features; the controlled shared-feature head comparison remains Table 4. These strata compare our readouts only.

Table A34: **Position accuracy by motion and visibility: DAVIS.** All positions use evaluation coordinates. Scores average videos with a defined denominator; different rows therefore describe different populations. Cog, Wan and Cosmos denote CogVideoX-2B, Wan2.1-1.3B and Cosmos-2B. Cosmos uses 480×720 input. The uniform-head comparison uses separately extracted features.

Evaluation subset	Videos	Visible positions	Cog $\delta\uparrow$	Wan $\delta\uparrow$	Cosmos $\delta\uparrow$	Cog uniform $\delta\uparrow$
Displacement [0,4) px	29	3,071	84.12	74.24	79.86	79.38
Displacement [4,16) px	30	7,130	69.27	63.89	61.73	66.21
Displacement [16,64) px	30	13,019	53.32	51.91	46.30	51.25
Displacement $\geq$ 64 px	29	4,954	43.93	41.23	36.93	41.32
1 to 5 frames after query	30	3,147	80.08	74.11	73.75	75.90
6 to 15 frames after query	30	5,419	67.07	63.98	60.38	63.93
16 to 30 frames after query	30	6,873	57.12	54.37	51.81	54.58
$\geq$ 31 frames after query	30	12,735	46.62	42.95	41.28	44.59
First 3 visible frames after occlusion	25	1,329	49.78	44.01	43.83	49.20
Visible, no previous occlusion	30	21,375	62.72	58.73	55.82	59.53

Table A35: **Position accuracy by motion and visibility: Kinetics.** All positions use evaluation coordinates. Scores average videos with a defined denominator; different rows therefore describe different populations. Cog, Wan and Cosmos denote CogVideoX-2B, Wan2.1-1.3B and Cosmos-2B. Cosmos uses 480×720 input.

Evaluation subset	Videos	Visible positions	Cog $\delta\uparrow$	Wan $\delta\uparrow$	Cosmos $\delta\uparrow$
Displacement [0,4) px	1143	1,520,840	81.81	68.43	80.76
Displacement [4,16) px	1140	1,035,715	62.27	54.90	57.60
Displacement [16,64) px	1116	1,455,010	42.76	43.46	37.51
Displacement $\geq$ 64 px	1119	687,087	26.07	28.30	19.55
1 to 5 frames after query	1144	158,324	80.82	70.74	78.41
6 to 15 frames after query	1144	299,323	71.33	63.05	66.20
16 to 30 frames after query	1144	416,451	65.76	59.12	59.90
$\geq$ 31 frames after query	1144	3,824,554	56.21	50.23	51.41
First 3 visible frames after occlusion	1061	76,757	41.55	37.51	36.79
Visible, no previous occlusion	1144	3,682,155	63.15	56.01	58.28

J HEAD-SELECTION DIAGNOSTICS

The head-selection experiment asks whether different queries benefit from different individual heads. Table A36 compares uniform averaging with three retrospective choices: the best head over the dataset, the best head for each video and the best head for each query trajectory.

For each valid query, we compute position accuracy using every individual head and uniform averaging. We average these query scores within each video, then average videos equally. This gives each query equal weight within its video; the headline TAP-Vid metric instead pools evaluated visible positions before averaging videos. All columns use the same 649 valid queries. One of the 650 queries has no evaluated visible frame and is omitted throughout.

Table A36: **Retrospective head selection on DAVIS.** Equal-video means of query-level position accuracy, using the same 649 valid queries throughout. All runs use cached features; block/timestep pairs are 17/19 for CogVideoX, 18/129 for Wan and 26/308 for Cosmos. Uniform averaging uses no annotations; the three head choices use ground truth retrospectively. These scores use a different aggregation from the headline benchmark results.

Backbone	Uniform $\delta\uparrow$	Fixed-head $\delta\uparrow$	Per-video $\delta\uparrow$	Per-query $\delta\uparrow$
CogVideoX-2B	57.70	56.56	58.77	66.04
Wan2.1-1.3B	56.51	56.91	58.38	63.37
Cosmos-2B	54.72	50.37	52.30	57.81

The fixed choice does not consistently improve on averaging, while per-query choices improve the diagnostic score on every backbone. This measures a retrospective improvement, whose recoverability must be tested using annotation-free signals. The experiment itself does not provide such a selector: all three choices use annotations. Their performance on unseen videos would require a separate evaluation, and selecting one head does not bound weighted combinations of heads.

Table A37: **Model size and retrospective head selection on DAVIS.** Cached features at each checkpoint's Kubric-selected block and timestep, with the equal-query-within-video statistic of Table A36. Both head choices use DAVIS annotations. Gains are in position accuracy ($\Delta\delta\uparrow$) relative to uniform averaging, with paired video intervals. More heads provide more opportunities for oracle selection.

Checkpoint	Heads	Uniform $\delta\uparrow$	Best fixed gain [95% CI]	Per-query gain [95% CI]
CogVideoX-2B	30	57.70	-1.15 [-2.31, -0.03]	+8.34 [+7.17, +9.45]
CogVideoX-5B	48	53.62	+2.94 [+1.45, +4.46]	+10.63 [+9.40, +11.94]
Wan2.1-1.3B	12	56.51	+0.41 [-0.83, +1.68]	+6.86 [+5.66, +8.10]
Wan2.1-14B	40	56.09	+6.90 [+5.55, +8.40]	+15.10 [+14.01, +16.24]
Cosmos-2B	16	54.72	-4.36 [-5.92, -2.91]	+3.09 [+1.62, +4.53]
Cosmos-14B	40	53.14	-3.01 [-4.96, -1.29]	+7.87 [+6.29, +9.59]

Table A38: **Observed oracle gains and a reference model of head variation.** Gain of per-query selection over the best fixed head, pooling all 649 valid DAVIS queries equally. The reference simulation preserves mean head quality, query difficulty and residual covariance across heads, using Gaussian residuals (400 draws, seed zero). This uses a different averaging rule from Table A37. The final column is observed minus simulated gain; this diagnostic is not a test of a deployable selector.

Checkpoint	Observed $\Delta\delta\uparrow$	Simulated $\Delta\delta\uparrow$	Difference
CogVideoX-2B	+9.89	+11.29	-1.40
CogVideoX-5B	+7.89	+9.32	-1.43
Wan2.1-1.3B	+6.69	+8.45	-1.76
Wan2.1-14B	+8.17	+10.51	-2.33
Cosmos-2B	+7.53	+9.37	-1.84
Cosmos-14B	+11.31	+15.04	-3.73

J.1 Head selection across model sizes

Table A37 extends the same cached-feature analysis to CogVideoX-5B, Wan-14B and Cosmos-14B. Wan-14B has a strong fixed head: selected retrospectively on DAVIS, it improves on uniform averaging by +6.90 [+5.55, +8.40], versus +0.41 [-0.83, +1.68] for Wan-1.3B. CogVideoX-5B also has a fixed-head advantage of +2.94 [+1.45, +4.46]; Cosmos-14B has none. These comparisons reveal information missed by uniform averaging but do not supply a deployable head selector.

The per-query oracle also has more heads to choose from in the larger models. To assess that effect, Table A38 compares its gain over the best fixed head with a simulated reference. We fit additive mean effects for head quality and query difficulty, then simulate Gaussian residuals with the observed covariance across heads. Scores are pooled equally over all valid queries in both observed and simulated data. The observed gains do not exceed this reference for either size in any of the three families. Oracle headroom alone therefore does not establish that heads specialise for different queries; the confidence analysis in Appendix F tests whether annotation-free signals identify useful heads. This simulation is a diagnostic with a Gaussian assumption, not a proof that query-dependent head selection cannot help.

K COMBINING SCORE MAPS ACROSS BLOCKS

We test two ways to combine block score maps before localisation. Both use fixed weights across queries and compare with a single-block reference under the same extraction and head-aggregation settings.

K.1 UNIFORM BLOCK AVERAGES

Table A39 uses the default CogVideoX configuration: anchor passes, confidence exponent two and the same 20 synthetic clips. All three tested combinations score below the block-17 reference, supporting the simpler single-block choice for this readout.

Table A39: **Uniform averages of block score maps.** CogVideoX-2B on 20 Kubric clips, with anchor passes, $\beta = 2$, $t = 19$ and 480×720 input. Localisation follows the average across the listed blocks.

Blocks	Position accuracy $\delta\uparrow$
17	69.72
15,17	68.54
17,20	66.69
15,17,20	67.01

K.2 WEIGHTED BLOCK MIXTURES

Table A40 reports a broader sweep with cached features, uniform heads, timestep 19 and 480×720 input. We form

$$P = \alpha P_{17} + (1 - \alpha) P_b, \qquad b \in \{14, 15, 16, 18, 20, 24\}, \quad \alpha \in \{0.5, 0.7, 0.8, 0.9, 0.95\},$$

then localise using the mixed map. The complete 30-combination sweep is shown. Its highest score is 66.05, compared with 65.94 for a separately executed block-17 reference with the same configuration and sample list. This small difference provides little evidence of a practical benefit from the tested mixtures.

Table A40: **Fixed convex mixtures of block score maps.** Position accuracy $\delta\uparrow$ on 20 Kubric clips. Columns give block 17's coefficient α; the partner receives $1 - \alpha$. Cached features, uniform heads, $t = 19$, input 480×720. The separately executed block-17 reference scores 65.94.

Partner block	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 0.95$
14	62.53	64.17	64.84	65.43	65.71
15	62.72	64.27	64.93	65.45	65.71
16	61.53	63.78	64.68	65.41	65.70
18	65.99	66.05	66.04	65.99	65.97
20	62.92	64.45	65.04	65.53	65.74
24	64.02	65.07	65.42	65.71	65.81

Both experiments combine completed matching maps. A decoder operating jointly on features, query-dependent or signed combinations, and refinement of one block's prediction by another remain distinct possibilities. The present results support the selected readout without establishing a limit on tracking information available across depth.

Table A41: **Settings and absolute DAVIS scores for the completed size study.** Both checkpoints use the same resolution within each family: CogVideoX and Cosmos 480×720, Wan 480×832. All use clean latents, anchor passes, a local 5×5 decoder and peak weighting. Blocks and timesteps come from the common-level Kubric grid; confidence and visibility are subsequently selected on Kubric. The CogVideoX-2B scores use the canonical reference in Tables 1 to 4. Cosmos-14B offloads 16 blocks to host memory. The earlier fixed-exponent Wan study in Table A42 is a separate setting.

Checkpoint	Block	t	β	Visibility	AJ↑	δ↑	OA↑
CogVideoX-2B	17	19	2.0	combo 0.16, 1	45.21	61.27	76.14
CogVideoX-5B	16	4	1.0	combo 0.1, 1.25	39.06	54.65	73.50
Wan2.1-1.3B	18	129	2.0	cycle 1.5	43.64	57.42	79.22
Wan2.1-14B	17	129	4.0	combo 0.12, 1.25	42.33	58.07	77.33
Cosmos-2B	26	308	1.0	combo 0.08, 1.75	39.73	54.83	74.64
Cosmos-14B	31	308	2.0	combo 0.07, 1.25	41.67	56.31	78.99

L Model size and input resolution

Model size, source-image preprocessing and backbone input dimensions can all affect accuracy. The comparisons below vary these choices separately.

L.1 Increasing model size

Table A41 gives the completed three-family DAVIS study summarised in Table 5. Each larger checkpoint uses the best cell from the matched scheduler grid, followed by its own Kubric confidence and visibility selection. The paired differences use the canonical CogVideoX-2B reference. Cosmos-14B improves OA by +4.35 [+1.74, +7.02]; neither its nor Wan-14B's position-accuracy interval excludes zero. These intervals allow positive gains and do not establish equality. The checkpoints also differ in pretraining and architecture, so these are empirical comparisons of available models rather than a controlled intervention on parameter count.

Table A42 separates our Wan size comparison from Point Prompting's published ablation (Shrivastava et al., 2025). Each row supports its own within-method size trend. Comparing the gains between rows would also involve their different query and source-image protocols, as discussed in Section 5.3.

Table A42: **Tracking accuracy with two Wan model sizes.** Our DAVIS pair uses all 650 first-mode queries, benchmark-resolution source frames, 480×832 backbone input, anchor passes and $\beta = 2$; selected blocks are 18 and 17. Point Prompting's published ablation uses one query per video and original-resolution source frames (Shrivastava et al., 2025). Compare model sizes within each row.

Method and setting	1.3B δ↑	14B δ↑	Change in δ↑
Ours (TAP-Vid first)	57.42	57.94	+0.52
Point Prompting, published	58.77	63.47	+4.70

L.2 Source-frame preprocessing

Table A43 changes the source-frame resize while holding backbone input dimensions and readout settings fixed. This study retains its original reference execution (more details in Appendix E). The direct path skips the intermediate 256×256 resize. It increases the CogVideoX point estimate and leaves Wan almost unchanged. The Cosmos increase has a paired interval including zero. The standard path is retained for the main benchmark comparisons.

L.3 Backbone input resolution

Increasing the Cosmos input resolution raises the position-accuracy point estimate, with paired interval [-0.29, 3.37], and lowers OA, with interval [-6.37, -2.16] (Table A44). We keep the stan-

Table A43: **Source-frame resizing.** DAVIS position accuracy with fixed backbone input size and readout settings. The standard path resizes to 256×256 before the backbone input; the direct path skips that intermediate resize. All 650 first-mode queries are scored.

Checkpoint	Backbone input	Via 256^2: $\delta\uparrow$	Direct source: $\delta\uparrow$
CogVideoX-2B	480×720	61.29	62.07
Wan2.1-1.3B	480×832	57.42	57.38
Cosmos-2B	480×720	54.83	55.57

dard source-frame path and recalibrate visibility on synthetic data for each size. These full-system changes do not isolate visibility errors; Appendix G examines visibility at fixed positions. The direct HeFT comparison in Table 2 shares input dimensions and has an uncertain AJ difference.

Table A44: **Backbone input resolution on Cosmos-2B.** Our runs share the standard source-frame path and decoder, with visibility calibrated on synthetic data separately for each size. HeFT retains its native preprocessing and tracking settings. Its TAP-Vid first row uses the same queries and frames as ours; its published row uses a different protocol (more details in Appendix C.2).

Method	Input	AJ↑	$\delta\uparrow$	OA↑
Ours	480×720	39.73	54.83	74.64
Ours	704×1280	39.93	56.37	70.52
HeFT (TAP-Vid first)	704×1280	41.47	58.07	79.19
HeFT (published)	704×1280	48.61	63.50	82.47

M ADAPTATION AND TRAJECTORY DECODING

We investigate three additional ways to use the frozen generator: changing its input latents, fitting head weights for each video and decoding a trajectory jointly across frames. Table A45 compares each with a reference that uses the same query set and extraction mode. These experiments use CogVideoX-2B and uniform-head references, separately from the final confidence-weighted configuration.

M.1 LATENT INVERSION

DDIM inversion (Song et al., 2021) offers an alternative to reading features from clean latents at a nonzero timestep. We deterministically invert the per-frame latents along CogVideoX's velocity-prediction schedule, using an empty prompt and no guidance. The inversion grid follows the 50-step trailing schedule, starting with cumulative signal coefficient one and retaining latents at the requested timestep. The comparison reads block 17 at timestep 19 using identical frame-zero anchor chunks. Inverted chunk latents are cached before readout. The stored timer excludes cache construction, so it does not measure total adaptation cost.

M.2 PER-VIDEO HEAD WEIGHTS

To test adaptation of the aggregation rule, we fit one vector of normalised head weights per video and discard it afterwards. The objective combines VAE-latent appearance agreement, cycle consistency, temporal smoothness and regularisation towards uniform weights. Their coefficients are 10, one, 0.1 and 0.01, respectively. Optimisation uses up to 150 steps, with held-out within-video triplets for early stopping.

Cycle consistency alone can reward a head that always matches a location to itself. A motion-based guard therefore excludes such heads, and the objectives give greater weight to moving regions. This experiment asks whether the remaining heads can be combined more effectively using these within-video objectives.

M.3 JOINT TRAJECTORY DECODING

With cached features, we construct both anchor-to-target appearance scores and consecutive-frame transition scores. Dynamic programming selects the token path maximising their summed log scores, with a transition coefficient of 0.5 selected on synthetic data. Local sub-cell refinement then adjusts the selected state in each frame. We evaluate both frame-zero and TAP-Vid first queries against their respective cached local-decoder references.

Table A45: **Adaptation and trajectory decoding on DAVIS.** Each row has a reference with the same query set and extraction mode. Frame-zero rows use 538 queries; all-first uses 650. CogVideoX-2B, block 17, $t = 19$, 480×720 input. References use uniform heads; the fitted-weight variant replaces that aggregation. The path coefficient is 0.5, selected on synthetic data. These references differ from the final confidence-weighted readout.

Variant	Query set	Extraction	Reference $\delta\uparrow$	Variant $\delta\uparrow$
DDIM inversion	frame zero	anchor pass	59.64	59.56
fitted head weights	frame zero	anchor pass	59.64	59.31
path decoder	frame zero	cached	58.83	58.78
path decoder	all first	cached	57.36	57.29

All three alternatives yield position scores close to their references. Paired intervals for their localisation differences include zero, so these experiments do not establish either a gain or a loss. They provide no clear benefit warranting inclusion in the current readout. The comparisons concern these particular objectives, schedules and decoding rules; they leave other forms of adaptation and joint feature decoding open.

N TRACKING THROUGH OCCLUSION

Figure 1 shows examples of the complete readout.

Figure 6 shows two additional DAVIS videos, *bmx-trees* and *parkour*, chosen for repeated annotated occlusion and reappearance. In *bmx-trees*, several tracks stay close to the rider while others make large excursions. *Parkour* includes later-starting queries and substantial errors as pose and viewpoint change. These examples make the failure modes visible alongside successful matches; benchmark accuracy is assessed over all videos in Table 1.



Figure 6: **Occlusion and reappearance on DAVIS.** Predictions from the headline CogVideoX readout with its selected visibility rule. Six evenly spaced query indices are displayed per video; AJ scores all queries over all evaluated frames. Tracks appear only from their query time onward. Colours identify tracks, filled/hollow markers denote predicted visibility/occlusion, and pale rings mark visible ground truth. Lines show the preceding 12 frames. The examples include both recovered matches and errors after occlusion.