HUMAN-AWARE VIDEO GENERATION ACCELERATION FOR RECONSTRUCTION-CONSISTENT VIDEO-TO-MOTION SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-video-to-motion synthesis can scale semantic coverage with abundant, diverse 2D videos without requiring paired text–3D motion-capture data at the same scale, but generating the video intermediate is slow. We introduce a training-free, human-aware accelerator for this bottleneck. It identifies a space–time human/action tube from signals already available in a Video DiT, processes most denoising calls only on the video-latent tokens inside the tube, reuses predictions outside it, and periodically refreshes the full frame. The resulting full-frame video and tube are passed to an unchanged motion reconstructor. The routing operator is shared across architectures, while tube estimation follows each model's native signals. On Wan2.2, the selected setting combines tube routing with ROI-sparse CFG and accelerates denoising by 2.4x. Across two architecturally distinct Video DiTs, paired reconstruction and motion-quality tests indicate that the accelerated videos generally retain the local evidence needed for motion recovery.

INTRODUCTION

Direct text-to-motion models efficiently produce skeleton or parametric-body sequences in their native representations (Zhang et al., 2023; Guo et al., 2024a; Wen et al., 2025). Text-to-video-to-motion offers a complementary scaling path: large 2D video collections provide broader action and semantic coverage without requiring paired 3D motion capture for every example. A video generator first maps a prompt to RGB (Wan et al., 2025; Yuan et al., 2026), and an existing motion reconstructor then recovers a body sequence from that video (Shen et al., 2024; Goel et al., 2023). The visual intermediate can express visible props, contacts, and body–scene relations, but its generation cost is the main drawback.

In this pipeline, video is an intermediate rather than the final output. A Video DiT nevertheless denoises the entire canvas at every step, whereas the motion reconstructor depends mainly on the person, body articulation, contacts, and nearby scene context. Existing accelerators primarily exploit model-internal or temporal redundancy. We instead allocate spatial computation according to the evidence required by the downstream motion reconstructor. We call the selected space–time region a *human/action tube* and the operation *tube routing*: most denoising calls process only the video-latent tokens inside the tube, while predictions outside it are reused from periodic full-frame calls. This directly shortens the visual token sequence processed by the transformer.

Figure 1 illustrates the framework. The tube is proposed from signals already available inside the generator and conservatively enlarged beyond the visible body to retain contacts, anticipated motion, and nearby context. The same tube can guide localization for an unchanged motion reconstructor. The routing operation is shared across architectures, while the signal used to estimate the tube follows each model's native design; the Method section gives these model-specific details.

We evaluate three aspects: first, whether the tube provides a reliable reconstruction crop on the same video; second, whether acceleration changes the motion recovered from a matched same-prompt, same-seed baseline; and third, whether the accelerated motion remains locally natural and active even when it is not identical to that baseline. The Method and Experiments sections formalize these aspects as interface consistency, motion consistency, and intrinsic motion quality.

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.

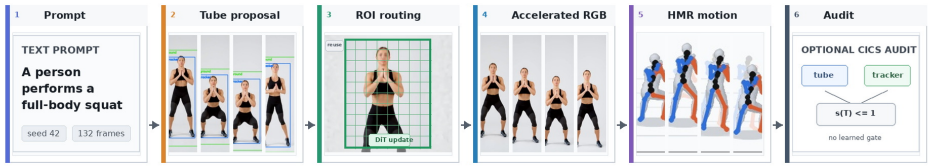

Figure 1: Framework overview. A training-free proposal estimates a human/action tube. Most denoising calls process only the tokens inside the tube, while periodic full-frame calls refresh predictions outside it. The full-frame video and a tube-guided crop are passed to an unchanged motion reconstructor. An optional audit checks crop reliability; output evaluation separately measures paired motion preservation and motion quality.

Experiments on Wan2.2 and Helios test tube routing on two different Video DiTs. On Wan2.2, the selected tube+ROI-sparse-CFG point achieves a 2.11x post-load generation speedup and a 2.40x denoising speedup while local-motion change remains below cross-seed variation. Helios is an already few-step-distilled video generator; tube routing therefore leaves a smaller but measurable 1.24x speedup.

Our contributions can be summarized as follows:

- A human-aware formulation of video-generation acceleration for text-to-video-to-motion synthesis that aligns spatial computation with the evidence required by downstream motion reconstruction.
- A training-free tube-routing operator that shortens the Video DiT token sequence, reuses exterior predictions, and composes with temporal caches.
- An evaluation across two Video DiTs and multiple motion reconstructors that separately measures reconstruction-crop reliability, paired motion preservation, and the intrinsic quality of accelerated motion.

Related Work

Video generation and acceleration. Diffusion models establish iterative denoising as a strong generative paradigm (Ho et al., 2020; Song et al., 2021); latent diffusion and DiTs improve its scalability (Rombach et al., 2022; Peebles & Xie, 2023), and video models extend it to dense space–time signals (Ho et al., 2022b;a; Blattmann et al., 2023; Bar-Tal et al., 2024; Yang et al., 2024). Their high inference cost has motivated reuse at several granularities: intermediate features can be cached, attention or residuals can be broadcast across steps, and redundant tokens can be merged or selectively updated (Ma et al., 2023; Bolya & Hoffman, 2023; Liu et al., 2024b; Zhao et al., 2024; Lv et al., 2024; Liu et al., 2024a; Zou et al., 2024). Recent video-specific work adds motion-adaptive caching and unified caching/pruning (Kahatapitiya et al., 2025; Sun et al., 2025), while structured sparse attention spans training-free and few-step-distilled designs (Xi et al., 2025; Chen et al., 2026; Ding et al., 2025). These optimize generic fidelity or model-intrinsic redundancy; our consumer-defined support is complementary and composable, including with content-agnostic token merging.

Attention and spatial control. Cross-attention exposes word–region correspondence. Prior work reads these maps for interpretation or segmentation (Tang et al., 2023; Wu et al., 2023) and modifies them to guide semantic presence or spatial layout (Chefer et al., 2023; Xie et al., 2023; He et al., 2023). We instead use warmup attention only to propose a generator-native human/action support. The distinction is operational: a localization map alone saves no computation, whereas our aligned support must shorten the transformer's visual sequence and still retain the contextual evidence needed by motion reconstruction.

Video-mediated human motion. Human video synthesis has primarily studied the forward rendering direction. Motion modules improve temporal coherence, reference- and pose-conditioned models animate a specified identity, and motion controls steer object or camera trajectories (Guo et al., 2024b; Hu et al., 2024; Xu et al., 2024; Wang et al., 2024b). These methods improve the con-

trollability of RGB output, but do not address the cost of producing that output when video is only an intermediate representation for 3D motion.

More closely related work uses pretrained image or video generators as motion priors or intermediates, extracting 3D animation from their features or outputs, or jointly modeling video and motion (Uzolas et al., 2024; Benedí San Millán et al., 2025; Lou et al., 2025; Zhao et al., 2026). This direction shows that video models contain useful human-dynamics knowledge. Converting pixels into an explicit body sequence relies on the complementary human-mesh-recovery literature: SMPL supplies the body representation and image regression estimates its parameters (Loper et al., 2015; Kanazawa et al., 2018); crop-aware methods preserve image-coordinate context (Li et al., 2022); and temporal or world-grounded systems recover coherent pose and root trajectories (Kocabas et al., 2020; Rempe et al., 2021; Goel et al., 2023; Shen et al., 2024; Shin et al., 2024; Wang et al., 2024a; Yin et al., 2024). We leave this reconstructor unchanged; our contribution is the shared tube that couples its spatial demand to generation-time computation.

Direct text-to-motion avoids RGB altogether, using diffusion, latent, discrete, or masked motion representations (Guo et al., 2022; Zhang et al., 2022; Tevet et al., 2023; Chen et al., 2023; Zhang et al., 2023; Guo et al., 2024a; Dai et al., 2024). It thus retains a clear inference-efficiency advantage. Our route is complementary: it seeks to retain the data scale, visual semantics, and inspectability of a video intermediate while narrowing its task-specific inference gap. Standard HumanML3D evaluation measures motion–text retrieval and embedding-space distribution quality (Guo et al., 2022); MotionCritic learns a scorer aligned with human motion preferences (Wang et al., 2025a), while SMP learns a reusable score-matching prior for natural motion (Mu et al., 2026). We use them as complementary preservation audits rather than calibrated measures of absolute correctness.

METHOD

TASK-SPECIFIC OBJECTIVE

Let $V = G(p, z)$ be the video generated from prompt p and noise z, and $M = R(V)$ the body pose, global orientation, and translation recovered by a backend R. Our routed generator produces $V' = G_{T,\pi}(p, z)$ using tube T and a schedule π that interleaves full-frame and region-of-interest (ROI) denoising. Rather than requiring every pixel in V' to match V, we characterize a frozen operating point by

$$\mathcal{O}(T, \pi) = \big(C(G_{T,\pi}), \\ D_{\text{motion}}(R(V'), R(V)), -Q_{\text{local}}(R(V'))\big), \tag{1}$$

where C is generation cost and D_{motion} measures reconstructed motion change, while Q_{local} measures baseline-calibrated intrinsic quality. Lower is better in all three coordinates. We use Eq. 1 to report and select operating points, rather than as an online constrained optimization problem. Matched consistency and intrinsic quality remain separate axes: a plausible action need not reproduce one stochastic baseline trajectory exactly. The criterion is task-specific: tube-exterior appearance may be updated less frequently, but context needed for articulation, contact, trajectory, or camera interpretation must remain available.

We evaluate the shared tube along two consistency axes and one quality axis. *Interface consistency* compares tube-crop and tracker-crop reconstruction on the same accelerated video. *Motion consistency* compares accelerated and matched unaccelerated reconstructions. The former can validate tube reuse at inference; the latter measures generation-induced change offline. Intrinsic quality checks the accelerated local motion without assuming exact reference identity. Formally, if $R_c(V')$ and $R_t(V')$ denote reconstructions obtained from a candidate and tracker crop, interface consistency bounds $D_{\text{motion}}(R_c(V'), R_t(V'))$, whereas motion consistency evaluates $D_{\text{motion}}(R_t(V'), R_t(V))$, which is the second coordinate in Eq. 1. The two comparisons respectively isolate crop stability and preservation of video-induced motion.

TRAINING-FREE TUBE ESTIMATION

We represent a general tube as $T_{1:F} = \{b_t\}_{t=1}^{F}$. Deployed proposals use one static swept box ($b_t = b$) to keep tensor shapes fixed; the framewise notation also covers the noncausal diagnos-

tic. A proposal defines a semantic core T_{core}; conservative expansion and patch alignment convert it into routing support

$$T_{\text{route}} = \text{Align}(\text{SafeExpand}(T_{\text{core}})) \,. \tag{2}$$

Thus, a body box may be adequate for reconstruction cropping but too restrictive for video denoising.

Our online Wan proposal reads the generator's actual softmax cross-attention, not an unnormalized query–key affinity. For visual query q and text key K_k, the probability at layer ℓ, head h, and warmup step s is

$$P_k^{\ell,h,s}(q) = \frac{\exp(q^{\ell,h,s} K_k^\top / \sqrt{d})}{\sum_j \exp(q^{\ell,h,s} K_j^\top / \sqrt{d})}. \tag{3}$$

We form $\mathcal{H}$ by case-insensitive subtoken matching against the fixed lexicon {person, human, body, man, woman, people, adult, arm(s), hand(s), leg(s), face, head, torso}. We sum P_k over $k \in \mathcal{H}$, average over selected heads, blocks, and the 12 full-frame warmup calls, and take a temporal maximum to obtain $A(x, y)$. A 3×3 patch average suppresses isolated responses. All evaluated prompts contain an explicit human descriptor; an empty $\mathcal{H}$ disables routing. Softmax probabilities are comparable across heads and blocks, whereas raw query–key scores are not.

Diffuse probability tails and boundary patches can bias a global centroid, so we use a component readout. We discard the outermost patch ring, threshold the interior map at its 90th percentile, form 8-connected components, and choose

$$C^* = \arg\max_C \sum_{(x,y)\in C} A(x, y). \tag{4}$$

We dilate C^* by three latent patches and enforce conservative width/height ranges of 35–60% and 75–100% of the canvas. Clipping and patch alignment produce one static swept support, repeated over latent time so routed tensors retain a constant shape. The procedure is prompt-conditioned, training-free, and adds neither a detector nor a learned localization module. Motion/update saliency may modulate feathering, but does not determine the component. The size constraints have distinct roles: the lower bounds prevent a sharply localized torso response from becoming an unsafe generation crop, while the upper width bound preserves a nontrivial token reduction. If the component lies near an image boundary, clipping shifts the safe box inward instead of shrinking it, preserving the requested patch count whenever the canvas permits.

We treat proposal selection and routing as separate components. **Fixed geometry** uses a centered, full-height, 80%-width strip. It is both a zero-overhead control and the replicated Helios transfer point; it is not the Wan attention component. **Reconstruction-derived support** estimates the region requested by the downstream task but is noncausal, as described next.

Reconstruction-derived support estimates the spatial extent requested by the downstream task. We run unchanged HMR2 on a matched completed video, project SMPL joints into the image, smooth their centers, and estimate width and height from the 95th-percentile joint extent with a 1.3 scale margin. This noncausal tube is an offline diagnostic rather than a deployable proposal. It approximates T_{core} and reveals when generation needs additional context.

Projected joints determine the center, scale, and motion path, whereas the generator uses their dilation. The spatial margin protects limb extrema and contacts, and temporal smoothing prevents box jitter from changing token boundaries. Motion-direction padding or a wider swept envelope can be added in Eq. 2; the diagnostic uses the fixed 1.3 margin. A larger margin can therefore be necessary even when projected joints are fully covered.

For Helios, **update energy** provides the online proposal from the preceding full-frame prediction u. Its spatial score is the normalized temporal update $\text{mean}_{c,t} |u_{c,t+1} - u_{c,t}|$ plus 0.15 times normalized response magnitude $\text{mean}_{c,t} |u_{c,t}|$, followed by 5×5 averaging. Marginal intervals retain 80% of this mass. In the conservative transfer setting, energy determines a smoothed horizontal center while the support remains 65%-width and full-height; a faster frontier variant also adapts width/height within 45–85%/65–100%. The proposal requires no training and responds to motion energy rather than explicit semantic labels.

Fixed geometry sets the full strip; softmax and update energy infer its center and possibly extent; projected HMR joints infer framewise center and scale. We retain their natural areas because area is part of the compute–coverage trade-off.

Algorithm 1: Wan Attention-Component Tube Routing

Input: denoiser f_θ, prompt p, latent z_S, shape-coupled history h_s (when present), human tokens $\mathcal{H}$, full-frame steps $\mathcal{F}$
Output: accelerated video V' and shared tube T

$T \leftarrow \emptyset$; $\bar{\epsilon} \leftarrow \emptyset$
for $s = S, \ldots, 1$ **do**
  if $s \in \mathcal{F}$ **then**
    $(\epsilon_s, \{q, K\}_s) \leftarrow \mathrm{FullForward}(f_\theta, z_s, p, s)$
    $\bar{\epsilon} \leftarrow \epsilon_s$
    if $T = \emptyset$ and warmup ended **then**
      $A \leftarrow \mathrm{MaxTime}(\mathrm{SoftmaxHuman}(\{q, K\}, \mathcal{H}))$
      $C \leftarrow \mathrm{MaxMassCC}_8(\mathrm{Interior}(A) \geq Q_{.90}(A))$
      $T \leftarrow \mathrm{Align}(\mathrm{SafeExpand}(C, 3))$
  else
    $(z_s^T, h_s^T) \leftarrow \mathrm{Crop}(z_s, h_s, T)$
    $\epsilon_s^T \leftarrow \mathrm{ROIForward}(f_\theta, z_s^T, h_s^T, p, s, \mathrm{offset}(T))$
    $\epsilon_s \leftarrow \mathrm{FeatherPaste}(\bar{\epsilon}, \epsilon_s^T, T)$
  end if
  $z_{s-1} \leftarrow \mathrm{SchedulerStep}(z_s, \epsilon_s, s)$
end for
return $\mathrm{Decode}(z_0), T$

TUBE-AWARE SPATIAL ROUTING

Algorithm 1 turns the tube proposal into computation savings. Full-frame warmup establishes scene layout and supplies attention. During an ROI step, we crop the latent and any model-history tensors to the patch-aligned tube and run the unchanged transformer on the shorter sequence. The resulting noise prediction is feathered into the most recent full-frame prediction; later full-frame calls refresh the exterior context. At the warmup–routing transition, the aligned box indexes every latent frame and all shape-coupled history tensors. The transformer therefore sees a shorter but internally consistent visual sequence; merely masking a full-frame output would retain the original compute and is not our routing operator.

Wan uses rotary positional offsets to preserve each token's full-canvas coordinate; without them, an off-center crop would be reinterpreted as a centered scene. Helios' adaptive energy variants use its full-canvas position mode, while the fixed centered control retains its native centered convention. Feathering linearly blends a narrow boundary band rather than hard-pasting the ROI. If the full sequence has N visual tokens and the tube retains fraction ρ, and $K_{\mathrm{full}}/K_{\mathrm{ROI}}$ count full/routed denoiser calls, the routed cost is approximately

$$C_{\mathrm{route}} \simeq K_{\mathrm{full}} C(N) + K_{\mathrm{ROI}} C(\rho N). \tag{5}$$

Transformer projections and MLPs decrease roughly linearly with ρ, while self-attention benefits more strongly. Warmup, refresh, text encoding, decoding, and data movement prevent end-to-end speedup from scaling as $1/\rho$.

Let P_T paste an ROI tensor into full-canvas coordinates and let M_T be a feathered tube mask. The noise prediction used by the scheduler at an ROI step is

$$\epsilon_s = M_T \odot P_T(\epsilon_s^T) + (1 - M_T) \odot \bar{\epsilon}_{s_f}^{\mathrm{ext}}, \tag{6}$$

where s_f denotes the latest full-frame call. Its exterior prediction is reused unchanged across intervening ROI timesteps and reset by the next full-frame call; the scheduler still updates the complete latent each step. Thus only the expensive denoiser forward is spatially restricted. Cropping all history tensors is likewise necessary for models that reuse latent features. Warmup establishes global layout, feathering smooths the boundary, and the reported full-frame calls restore full-canvas token interaction and refresh the exterior prediction.

Wan uses 12 full-frame warmup steps, 36 ROI steps, and two final full-frame steps. Full-canvas steps use standard classifier-free guidance (CFG). Both branches are evaluated every third ROI step; intervening steps combine the current conditional prediction with the most recent ROI unconditional prediction. We report this as a combined operating point: sparse CFG is an orthogonal implementation choice, whereas spatial savings come from the shorter token sequence and cached exterior.

Helios applies the same crop–forward–paste operation to its final pyramid stage and preserves the official 2/2/2 temporal schedule. The replicated fixed-geometry point routes all eligible final-stage calls and reuses the resized preceding-stage prediction as its exterior cache, without a same-stage refresh. The adaptive update-energy variant instead uses full-canvas coordinates and refreshes every four eligible calls.

Generic denoising accelerators can be composed with this operator. A residual cache reduces the number of transformer evaluations, while tube routing reduces the token sequence of evaluations that remain. Their composition requires no weight update: we maintain cache state separately for each CFG branch and token grid shape, so a residual computed on an ROI is never added to a full-frame sequence. We evaluate this composition with TeaCache (Liu et al., 2024a); the same separation applies to other training-free temporal caches.

Shared Reconstruction Interface

The generator exports T with V', allowing a crop-aware reconstructor to use the support directly or constrain a lightweight foreground refinement. To audit this interface, we compare either candidate crop c with the reconstructor's tracker crop t on the same accelerated video. Let d_b and d_g denote mean geodesic errors in local-joint and global orientation, and let d_r and d_v denote root-trajectory and root-velocity errors after first-frame alignment. The HMR2 Crop-Interface Contract Score is

$$s(c) = \max\left(\frac{d_b}{5^\circ}, \frac{d_g}{20^\circ}, \frac{d_r}{0.15\mathrm{m}}, \frac{d_v}{0.10\mathrm{m/frame}}\right). \tag{7}$$

A candidate satisfies the interface contract when $s(c) \leq 1$. The physical scales are fixed on known-motion development data, and the maximum implements a conjunction: good local pose cannot conceal a large orientation or trajectory change. These fixed, backend-specific scales define a dimensionless diagnostic score.

The interface exports coordinates rather than a cropped video. The generator decodes the full frame, from which HMR samples its native RGB crop. Reuse changes only the localization input, leaving HMR preprocessing, weights, and motion representation unchanged and the coordinates comparable to tracker boxes.

The diagnostic runs HMR2 with tracker, raw-support, and foreground-refined crops. Refinement searches foreground evidence within the exported support, then applies the reconstructor's standard square-crop preprocessing. Both priors are compared with the native tracker crop on V'. CICS is therefore an offline interface audit; deployment uses one refined-prior or tracker path. Matched unaccelerated reconstruction is used only to report motion preservation.

Experimental Setup

Models and Evaluation Protocol

We evaluate two architecturally distinct open Video DiT generators as a cross-architecture transfer test. Wan2.2 TI2V-5B (Wan et al., 2025) is invoked with `img=None` in text-only mode and generates 121 frames at 1280×704 and 24 fps with 50 denoising steps. Its routed configuration is evaluated on all 51 prompts at seeds 42 and 202, yielding 102 matched pairs. Helios-Distilled (Yuan et al., 2026) generates 132 requested frames at 384×640 and 24 fps (129 decoded frames). Its selected transfer point uses update energy to center a full-height, 65%-width final-stage ROI, refreshes globally every four eligible calls, and retains the official 2/2/2 schedule. A fixed centered 80%-width point is replicated over three seeds, and a faster update-energy variant also adapts the extent. Every accelerated clip is paired with the official pipeline at the same prompt and seed.

Wan's selected configuration uses the 12/36/2 schedule in Algorithm 1 and alternates baseline/candidate execution order across pairs. The frozen 102-pair acceleration run uses the softmax-component readout described in Eq. 4. After component selection and conservative expansion, the schedule-averaged token fraction is .568, including the 14 full-frame steps. These are therefore the primary quantitative results for the component method. Helios applies ROI routing only to the final pyramid stage. The fixed transfer control uses seeds 101, 202, and 303. Adaptive proposal variants

are screened on a fixed eight-prompt subset at seed 101; their parameters are then frozen for the complete 51-prompt run reported below.

The 51 prompts comprise 33 standard actions, 10 rare dances, and 8 unusual semantic or prop actions. HMR2 (Goel et al., 2023) is the primary backend; GVHMR tests cross-reconstructor transfer, and PromptHMR (Wang et al., 2025b), run on the generated videos, independently audits all 102 Wan selected and full-CFG component-only pairs. We render 60 HY-Motion sequences with known joints, using 18 to set the HMR2 scales in Eq. 7 and holding out 42. Crop shifts, a tight crop, and 0.8x/1.2x temporal resampling probe the intended channels. Similarity-aligned PA-MPJPE validates local pose. CICS retains global orientation and root motion only as an inspectable crop-interface contract; the final action-quality endpoint does not compare absolute global orientation.

The primary baseline keeps generator, prompt, seed, and temporal schedule but disables routing. A fixed-hash Wan subset of 16 prompts (8/4/4 by category) compares full-frame generation, tube routing, full-frame sparse CFG, the selected tube/ROI-sparse schedule, and TeaCache alone or composed with the tube. The sparse control evaluates both CFG branches every third step; our schedule keeps full CFG for 14 warmup/endpoint calls and sparsifies only 36 ROI calls. We refit TeaCache's fourth-order predictor on 192 forward pairs from two disjoint prompts and freeze threshold .10 on one excluded development clip. Helios operating points retain the same official temporal schedule and vary only spatial support and the explicitly reported refresh calls.

The component study includes two full-frame controls on the same Wan backbone: step-sparse CFG and TeaCache. They isolate downstream spatial routing from reducing evaluations or reusing features without spatial support.

Metrics and Timing

We group the metrics by the question they address. CICS audits crop-interface consistency on the same video using the four backend channels in Eq. 7; its scales do not enter the 95/102 output endpoint. Matched native-tracker reconstruction measures acceleration-induced motion change, and baseline-calibrated dynamics test the local quality of the accelerated output. For the cross-generator headline, we pair post-load video-generation speed with continuous HMR2 local-joint change. Dynamics use PromptHMR-video on Wan and HMR2 on Helios. For Wan, PromptHMR-video reports local SMPL-X rotations and root path after first-frame translation alignment. We compare both with unaccelerated cross-seed variation rather than transferring HMR2 thresholds; absolute global angle is excluded. CICS is a backend-specific interface diagnostic, applied in the known-motion calibration and Helios interface study rather than as the Wan quality endpoint.

Intrinsic quality uses local SO(3) angular-acceleration and jerk p95 plus the fraction of local rotation energy above 6 Hz. Robust 95th-percentile thresholds are fitted only on unaccelerated baselines from the other seed(s). We apply a $\log(1 + x)$ transform, robust median/MAD scaling, and take the maximum across the three statistics; paired ratios use a cross-seed 10th-percentile denominator floor. A candidate must also retain at least 50% of the paired angular path and active-frame fraction. We report both 1.5x and 2.0x paired-ratio sensitivity analyses. Wan additionally uses a conservative eight-frame CLIP guard ($\Delta\mathrm{CLIP} \geq -.02$); CLIP (Radford et al., 2021) supplies a conservative image–text guard. We call a pass under either matched consistency or the intrinsic-quality tests *motion acceptability*. The guard removes no clips, and this union is only a sensitivity summary, not an operating-point selector or ground-truth label.

Three pretrained audits complement these checks on Wan. The public HumanML3D evaluator reports seed-averaged R-precision, MM-Dist, Diversity, and motion-embedding FID after standard 263-D conversion at 20 fps (Guo et al., 2022; 2024a); here paired FID measures distribution shift between accelerated and matched-baseline embeddings. MotionCritic uses the SMPL input convention described in its paper after our fixed SMPL-X-to-SMPL mapping (Wang et al., 2025a). Exploratory SMP maps SMPL-X to its compact humanoid and reports negative baseline-normalized SDS loss at $t \in \{22, 15, 8\}$ over 32 common-noise repeats under the LaFAN1 benchmark/setup (Mu et al., 2026). Learned-score tests average the two seeds per prompt, leaving 51 independent units.

The headline timing is resident video generation: Wan uses complete post-load generation wall time and Helios the resident pipeline call. Model loading, HMR, and video export are excluded.

Table 1: Selected speed–motion operating points. Gen. is resident full-video generation speedup after model loading; Local is mean HMR2 local-joint angle change under matched native tracking. Dyn. A/B gives accelerated/baseline pass counts over the same N clips (PromptHMR-video for Wan; HMR2 for Helios).

Model	Setting	N	Gen.	Local	Dyn. A/B
Wan	Component + sparse CFG	102	2.112x	4.33	90/93
Helios	Energy-center 65% + refresh	51	1.240x	6.89	48/46

Denoiser-only timing is retained only to attribute component cost. Reported confidence intervals resample prompts while retaining all seeds of a sampled prompt.

All headline timing uses one NVIDIA H20. Wan runs in bfloat16; proposals average all heads from blocks 20/24/28, and pasting uses a two-patch feather. Counterbalanced resident execution reduces order effects. Diagnostic CICS charges three HMR2 paths, whereas deployment charges one. Confidence intervals use 20,000 prompt-cluster bootstrap replicates without treating frames or seeds as independent samples.

RESULTS

Before applying CICS to generated videos, we validate it on known-motion sequences. It detects crop-induced reconstruction changes and responds to temporal warps in the intended channels; full calibration results and scale sensitivity are reported in Appendix A.

WAN ACCELERATION AND MOTION PRESERVATION

Table 1 places both selected operating points under the same timing and motion definitions; Wan provides the primary speed result, while Helios tests transfer to an already distilled generator.

Primary operating point. The selected combined Wan point approximately halves token processing, while matched local-motion change remains well below cross-seed variation. Across 102 pairs, its .568 token fraction yields 2.395x denoiser and 2.112x resident-generation speed. Resident generation falls from 622.2 s to 294.6 s on average; this timing includes attention collection and component extraction during warmup. The 4.33-degree local change is .175x the 24.81-degree cross-seed mean; the dynamics pass rate changes by -2.9 points (95% CI $[-8.8, +2.9]$).

Estimator agreement. The root-path estimates depend strongly on the estimator, even though the local-motion measurements agree. Across all 102 selected pairs, PromptHMR reports 6.89 degrees local and .080 m root change (.271x/.274x cross-seed variation), versus .190 m root change with HMR2; the median per-pair PromptHMR/HMR2 root ratio is .368x. Their local angles use different body models (SMPL-X versus SMPL) and are not directly interchangeable.

Intrinsic local quality. Local dynamics remain close to baseline without a systematic increase in the measured contact artifacts: median acceleration/jerk/high-frequency ratios are 1.11/1.18/1.05, and all three contact intervals span zero. Requiring every ratio below 1.5x retains 81/102 pairs; combining matched consistency with the quality checks retains 90/102 at 1.5x and 95/102 throughout the 2.0–2.5x sensitivity range. The other seven cases fail a temporal-ratio or robust check.

Learned motion-space audits. Independent pretrained representations show small, mixed changes rather than a consistent quality decline. Table 2 summarizes these complementary audits on the same 102 pairs.

MotionCritic changes by $+.243$ (95% CI $[-.404, +.926]$); HumanML3D retrieval is mixed and Diversity retains 98.2%. Paired FID is .123 (.056x cross-seed), and matched root/Procrustes MPJPE is 54.7/33.0 mm (.212x/.201x cross-seed). SMP changes by .29% but remains supporting evidence because of its domain gap.

On the fixed 16-prompt subset, tube-only routing reaches 1.84x and outpaces TeaCache at similar local change; adding ROI-sparse CFG reaches 2.11x. Full-frame sparse CFG changes motion substantially more, while Tube+TeaCache shows that spatial routing composes with temporal caching.

Table 2: Pretrained 3D-motion audits on 102 Wan pairs; R@k denotes R-precision at rank k, and paired embedding FID measures the shift between accelerated and matched-baseline embeddings. Arrows show favorable direction.

Metric	Base	Accel.	Change
MotionCritic ↑	-5.754	-5.511	$+.243$
R@1 ↑	.217	.201	$-.016$
R@2 ↑	.334	.346	$+.012$
R@3 ↑	.420	.435	$+.015$
MM-Dist ↓	5.240	5.204	$-.036$
Diversity ↑	6.740	6.617	$.982\times$
Paired emb. FID ↓	–	–	.123
SMP quality ↑	-1.000	-1.003	$-.003$

Table 3: Component study on 16-prompt Wan subset. Gen. is post-load speedup; Local is HMR2 local-joint change.

Setting	Gen.	Local	ΔCLIP
Full frame	1.00x	0.00	$+.0000$
Full sparse CFG	1.40x	17.42	$-.0181$
Human tube	1.84x	5.94	$-.0000$
Tube + ROI-sparse CFG	2.11x	5.86	$+.0015$
TeaCache	1.62x	6.17	$-.0019$
TeaCache + tube	2.77x	6.66	$-.0039$

On the full 102-pair audit, the same component routing with full CFG reaches 2.16x denoiser speedup and realizes over 90% of the denoiser-time reduction of the 2.40x selected combined run. PromptHMR measures 6.30 degrees local and .075 m root change, .248x and .257x cross-seed variation. Sparse CFG adds .59 degrees locally (95% CI $[.31, .88]$), while its .005 m root-path increase is unresolved (95% CI $[-.005, .017]$). We treat Tube+TeaCache as a compatibility result because it was evaluated only on the 16-prompt component subset.

Helios Transfer

Spatial routing retains a smaller but measurable frontier after temporal distillation. On Helios, the balanced energy-centered point reaches 1.240x with a 6.89-degree local change. Its dynamics pass rate changes by $+3.9$ points (95% interval $[-3.9, +11.8]$); median acceleration, jerk, and high-frequency ratios are 1.014/1.010/1.002, and the 1.5x criterion retains 45/51 clips.

The energy-centered point is modestly faster than fixed geometry, while foreground refinement raises the crop-interface pass rate. The complete frontier, interface audit, proposal comparison, and estimator-sensitivity results are reported in Appendix B.

Video-Intermediate User Study

A 20-participant, category-balanced comparison of rendered motions against direct text-to-motion HY-Motion finds that it leads on standard actions, whereas our method is preferred for rare dances and rare semantic/prop actions. Full scores and protocol details are in Appendix B.

Conclusion

We introduced a training-free, human-aware accelerator that uses generator-native tubes to concentrate Video DiT denoising on evidence needed by an unchanged motion reconstructor. Model-specific proposals and coordinate handling adapt the shared crop–forward–paste operator to mainstream Video DiTs, as demonstrated on Wan2.2 and Helios. Matched, intrinsic, and pretrained audits indicate that recovered motion largely retains unaccelerated local quality, and the router composes with CFG sparsification and temporal caching. Code and configurations will be released upon acceptance.

However, gains are bounded by tube area and residual model redundancy: multi-person or large-displacement clips need wider support, and few-step distillation offers fewer routable calls. A one-step generator offers none, so the current online router cannot accelerate it. The claims concern local motion: global/root fidelity remains estimator-dependent, and prompts without a matched human descriptor fall back to full-frame generation.

REFERENCES

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

Marc Benedí San Millán, Angela Dai, and Matthias Nießner. Animating the uncaptured: Humanoid mesh animation with video diffusion models. *arXiv preprint arXiv:2503.15996*, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

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

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

Pengtao Chen, Xianfang Zeng, Maosen Zhao, Mingzhu Shen, Wei Cheng, Gang Yu, and Tao Chen. Sparse-vdit: Unleashing the power of sparse attention to accelerate video diffusion transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2957–2965, 2026. doi: 10.1609/aaai.v40i4.37287.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18000–18010, 2023.

Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. MotionLCM: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*, pp. 390–408, 2024.

Hangliang Ding, Dacheng Li, Runlong Su, Peiyuan Zhang, Zhijie Deng, Ion Stoica, and Hao Zhang. Efficient-vdit: Efficient video diffusion transformers with attention tile. *arXiv preprint arXiv:2502.06155*, 2025.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024a.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations*, 2024b.

Yutong He, Ruslan Salakhutdinov, and J. Zico Kolter. Localized text-to-image generation for free via cross attention control. *arXiv preprint arXiv:2306.14636*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022b.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024.

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

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131, 2018.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5253–5263, 2020.

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pp. 590–606. Springer, 2022.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. *arXiv preprint arXiv:2411.19108*, 2024a.

Haozhe Liu, Wentian Zhang, Jinheng Xie, Francesco Faccio, Mengmeng Xu, Tao Xiang, Mike Zheng Shou, Juan-Manuel Perez-Rua, and Jurgen Schmidhuber. Faster diffusion via temporal attention decomposition. *arXiv preprint arXiv:2404.02747*, 2024b.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015.

Yuke Lou, Yiming Wang, Zhen Wu, Rui Zhao, Wenjia Wang, Mingyi Shi, and Taku Komura. Zero-shot human-object interaction synthesis with multimodal priors. *arXiv preprint arXiv:2503.20118*, 2025.

Zhengyao Lv, Chenyang Si, Junhao Song, Zhenyu Yang, Yu Qiao, Ziwei Liu, and Kwan-Yee K. Wong. Fastercache: Training-free video diffusion model acceleration with high quality. *arXiv preprint arXiv:2410.19355*, 2024.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. *arXiv preprint arXiv:2312.00858*, 2023.

Yuxuan Mu, Ziyu Zhang, Yi Shi, Dun Yang, Minami Matsumoto, Kotaro Imamura, Guy Tevet, Chuan Guo, Michael Taylor, Chang Shu, Pengcheng Xi, and Xue Bin Peng. SMP: Reusable score-matching motion priors for physics-based character control. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2026)*, 45(4):1–21, 2026. doi: 10.1145/3811282.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11488–11499, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Proceedings*, 2024.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Wenzhang Sun, Qirui Hou, Donglin Di, Jiahui Yang, Yongjia Ma, and Jianxun Cui. UniCP: A unified caching and pruning framework for efficient video generation. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, pp. 1–7, 2025. doi: 10.1145/3743093.3770981.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 5644–5659, 2023.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023.

Lukas Uzolas, Elmar Eisemann, and Petr Kellnhofer. Motiondreamer: Exploring semantic video diffusion features for zero-shot 3d mesh animation. *arXiv preprint arXiv:2405.20155*, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haoru Wang, Wentao Zhu, Luyi Miao, Yishu Xu, Feng Gao, Qi Tian, and Yizhou Wang. Aligning human motion generation with human perceptions. In *International Conference on Learning Representations*, 2025a.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. *arXiv preprint arXiv:2403.17346*, 2024a.

Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J. Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH Conference Papers*, 2024b. doi: 10.1145/3641519.3657518.

Yuxin Wen, Qing Shuai, Di Kang, Jing Li, et al. Hy-motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025.

Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1206–1217, 2023.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, Jianfei Chen, Ion Stoica, Kurt Keutzer, and Song Han. Sparse VideoGen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025.

Jinheng Xie, Yuexiang Li, Yefeng Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. BoxDiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7452–7461, 2023.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1481–1490, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, Ziwei Liu, and Lei Yang. Whac: World-grounded humans and cameras. In *European Conference on Computer Vision*, pp. 20–37. Springer, 2024.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

Chengfeng Zhao, Jiazhi Shu, Yubo Zhao, Tianyu Huang, Jiahao Lu, Zekai Gu, Chengwei Ren, Zhiyang Dou, Qing Shuai, and Yuan Liu. Comovi: Co-generation of 3d human motions and realistic videos. *arXiv preprint arXiv:2601.10632*, 2026.

Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast. *arXiv preprint arXiv:2408.12588*, 2024.

Chang Zou, Xuyang Liu, Ting Liu, Siteng Huang, and Linfeng Zhang. Accelerating diffusion transformers with token-wise feature caching. *arXiv preprint arXiv:2410.05317*, 2024.

A CICS CALIBRATION

CICS reliably detects crop-induced reconstruction changes while retaining interpretable pose and trajectory channels. Across all 60 known-motion sequences, local-rotation distance and full CICS reach AUCs of .948 and .901, with sequence-bootstrap 95% intervals of [.911,.976] and [.851,.943]. On the 42 held-out sequences, the corresponding AUCs are .933 and .868. Varying any one physical scale by $\pm 30\%$ leaves AUC at least .883, and temporal warps affect the intended pose and trajectory channels (Table 4).

Table 4: Known-motion validation over 60 sequences. Crop AUC detects an absolute PA-MPJPE increase above 5 mm; time-warp entries report Spearman correlation with that increase.

Probe	N	Metric	Result
Crop perturbations	360	local angle ρ / AUC	.786 / .948
Crop perturbations	360	full CICS ρ / AUC	.684 / .901
Time warps	120	local / global angle ρ	.820 / .479
Time warps	120	root path / velocity ρ	.548 / .596

Table 5: Helios routing on 51 prompts (seed 101). Gen. is resident speedup (higher is faster); Active is token fraction.

Setting	Gen.	Active	Local Dyn.	A/B
Full frame	1.000x	1.000	0.00	46/46
Fixed center, 80%	1.197x	.841	6.96	50/46
Energy center, 65% + refresh	1.240x	.810	6.89	48/46
Adaptive energy extent + refresh	1.479x	.636	7.77	45/46

B SUPPLEMENTARY RESULTS

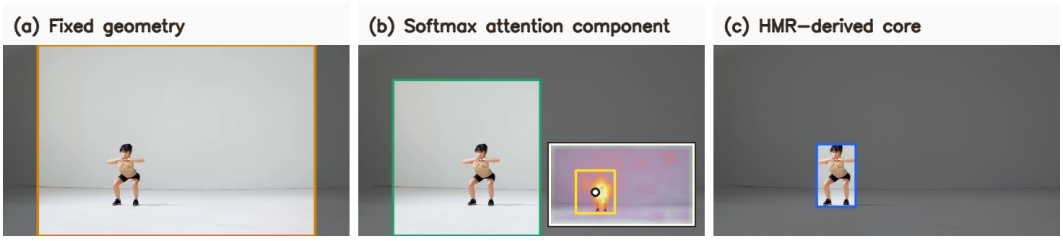

Figure 2: Off-center Wan supports: attention component, fixed geometry, and noncausal HMR core at their natural unequal areas.

Table 6: Mean preference scores against HY-Motion on $[-2, 2]$; positive values favor our method.

Action set	Text match	Naturalness	Overall
Standard	$-.43$	$-.56$	$-.44$
Rare dance	$+.40$	$+.07$	$+.37$
Rare semantic/prop	$+.28$	$+.10$	$+.23$
All	$+.08$	$-.13$	$+.05$

Additional Helios and study details. Across three seeds, the fixed 80% control reaches 1.201x, 6.84 degrees, and 147/153 versus 146/153 dynamics passes. On seed 101, foreground refinement raises CICS from 22/51 to 41/51 ($p = 6.6 \times 10^{-5}$) while retaining 49/51 local-joint passes; off-screening raw/refined counts are 18/43 and 36/43. Adaptive energy extent reaches 1.479x at 7.77 degrees (Table 5). Root motion is less stable across estimators than local rotation: in a 51-clip Helios stress audit, HMR2/GVHMR agreement is .719/.809 for root path/velocity versus .965 for local rotation, so absolute global yaw is excluded from claims.

The user study comprised 18 pairs (six standard, six rare-dance, and six unusual semantic/prop actions). Participants rated text match, naturalness, and overall preference on $[-2, 2]$; positive values favor our method (Table 6).

Timing uncertainty and component attribution. The 102-pair timing summary and the 16-prompt component study use the same prompt-cluster bootstrap as the main experiment. They sepa-

rate the contribution of spatial routing from that of sparse CFG while keeping the selected operating point fixed.

Table 7: Bootstrap intervals for timing and the Wan component study. Resident times are in seconds; Local is the matched HMR2 local-joint angle change.

102-pair resident timing			
Quantity	Baseline	Accelerated	Ratio (95% CI)
Resident wall time	622.184	294.622	$2.112\times$ [2.110, 2.114]
Denoising time	562.913	235.015	$2.395\times$

16-prompt component study			
Setting	Speedup (95% CI)	Local (95% CI)	ΔCLIP (95% CI)
Full sparse CFG	1.405 [1.401,1.408]	17.42 [14.09,20.89]	$-.0181$ [$-.0263$, $-.0099$]
Human tube	1.843 [1.839,1.849]	5.94 [3.42,8.90]	$-.0000$ [$-.0046$, $+.0045$]
Tube + sparse CFG	2.110 [2.104,2.116]	5.86 [3.42,8.76]	$+.0015$ [$-.0033$, $+.0069$]
TeaCache	1.616 [1.611,1.621]	6.17 [3.86,8.82]	$-.0019$ [$-.0047$, $+.0009$]
TeaCache + human tube	2.771 [2.759,2.784]	6.66 [4.26,9.43]	$-.0039$ [$-.0097$, $+.0021$]

Pretrained motion-space audits. Table 8 reports the paired quantities behind the main learned-audit summary. The FID comparison is between the accelerated and matched-baseline sets; the seed-split value is the corresponding unaccelerated variation on the same prompts.

Table 8: Supplementary motion-space measurements on the 102 Wan pairs. A dash means that the metric is a paired distance rather than a score for either set.

Metric	Baseline	Accelerated	Paired change or context
MotionCritic (higher)	-5.754	-5.511	$+.243$ [$-.404$, $+.926$]
R-precision @1	.217	.201	$-.016$
R-precision @2	.334	.346	$+.012$
R-precision @3	.420	.435	$+.015$
MM-Dist (lower)	5.240	5.204	$-.036$
Diversity	6.740	6.617	$.982\times$ retained
Embedding FID	–	–	.123 (seed-split: 2.209)
Root-aligned MPJPE	–	–	54.7 mm [40.6, 71.7]
PA-MPJPE	–	–	33.0 mm [24.4, 43.3]
SMP quality (higher)	-1.000	-1.003	$-.003$ [$-.0136$, $+.0096$]

The MPJPE intervals are prompt-cluster 95% intervals. The SMP score is an exploratory, baseline-normalized audit; its absolute value is not a calibrated human rating. Under the same paired protocol, the 0.3-rad jitter control lowers the SMP score by .276, providing a simple sensitivity check.

Helios interface refinement. The selected Helios route exports a support that can be refined using foreground evidence before the unchanged HMR2 crop preprocessing. Table 9 gives the complete interface comparison on the 51 accelerated clips.

Table 9: Helios crop-interface audit on 51 accelerated clips. Angles are in degrees and root quantities in metres (velocity in metres/frame).

Prior	CICS	Local-only	Body	Global	Root path	Root vel.
Raw support	22/51	33/51	4.855	7.174	.257	.080
Foreground-refined	41/51	49/51	2.624	3.448	.102	.049

The raw-to-refined pass transition is significant under an exact McNemar test ($p = 6.6 \times 10^{-5}$). This is an interface diagnostic; the Helios speed and motion-quality results above use the selected deployed route and do not charge the offline refinement as an additional model.