NOT ALL REGIONS NEED EQUAL TOKENS: MOTION-ADAPTIVE REPRESENTATION ALLOCATION FOR COMPLEX HUMAN IMAGE ANIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Complex human motion creates spatially non-uniform representation demands. Fast-changing body regions require finer representations, while slowly varying regions can be represented with fewer tokens. However, existing video Diffusion Transformers typically use a uniform patch grid, assigning the same token density to regions with substantially different representation needs. We rethink complex human image animation from this representation-allocation perspective and propose **Motion-Adaptive Token Allocation (MATA)**, which reallocates spatial representation capacity according to motion demand. MATA assigns denser representations to motion-intensive regions while compressing redundant tokens in low-motion regions, allowing additional capacity to be directed where complex motion requires it without uniformly increasing the token budget. On HyperMotionX, MATA improves generation under complex human motion, reducing FVD by 7.97% and increasing $\mathrm{PCK}_{\mathrm{bbox}}@0.05$ by 6.54 percentage points over UniAnimate-DiT 5B, with only a 1.29% increase in estimated main-DiT FLOPs. These results highlight the importance of spatially non-uniform representation allocation for complex human animation.

1 INTRODUCTION

Human image animation generates a video of a reference character following a driving motion sequence. Its central objective is to reproduce the prescribed movement while preserving the character's appearance and maintaining coherent body structure. Building on diffusion generative models (Ho et al., 2020; Rombach et al., 2022), existing methods improve reference appearance injection, pose conditioning, and temporal modeling (Peebles & Xie, 2023; Wan et al., 2025; Gan et al., 2025; Wang et al., 2025b; Cheng et al., 2025). Recent video Diffusion Transformers (DiTs) further advance generation quality through large pretrained video backbones (Gan et al., 2025; Wang et al., 2025b; Cheng et al., 2025). However, reliable motion transfer also requires handling complex articulation. Turns, inverted poses, and crossing limbs introduce rapid changes in visibility and body-part relationships. Errors at these moments can alter the depicted action even when the character and background remain visually plausible. Preserving local structure under difficult motions is therefore essential to faithful animation.

Recent work has improved complex human animation primarily by strengthening motion conditioning: enabling the generator to receive, interpret, and exploit difficult motion signals more effectively. HyperMotionX introduces a dataset and benchmark for complex human motions, together with a DiT-based model that enhances structural modeling through spatial low-frequency rotary position embeddings (Xu et al., 2025). SCAIL improves motion guidance using 3D-consistent pose representations and full-context pose injection (Yan et al., 2026b). These advances address an important question: how can complex motion be better conveyed to the generator? We pursue a complementary question: given the motion condition, is the model allocating sufficient representation capacity to the regions that need it most? Existing approaches still largely rely on a uniform spatial tokenization scheme, assigning comparable representation density across regions despite strongly non-uniform motion demands. We therefore shift the focus from improving motion conditions to reallocating representation capacity, introducing a motion-adaptive representation paradigm in which token density is matched to spatially varying motion demand.

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.

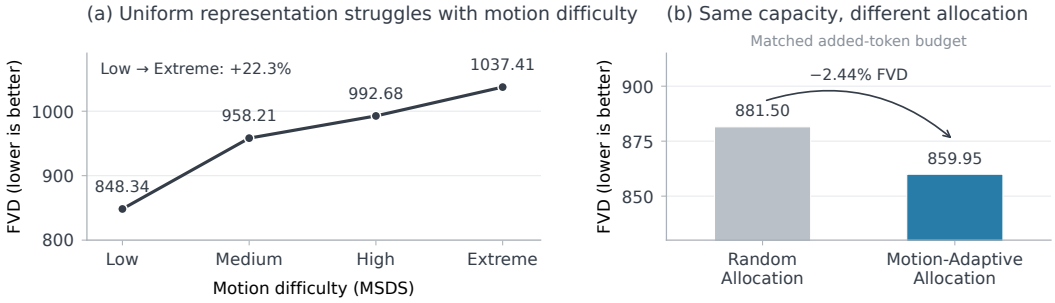

Figure 1: Motivation for motion-adaptive representation allocation. (a) UniAnimate-DiT 5B with uniform tokenization shows increasing FVD with motion difficulty, rising by 22.3% from Low to Extreme. (b) Under the same added-token budget, motion-adaptive allocation reduces FVD from 881.50 to 859.95 compared with random allocation. The results highlight both the limitation of uniform allocation and the importance of allocating representation capacity to the right regions.

This representation-side question becomes particularly important as motion complexity increases. In our pilot evaluation, UniAnimate-DiT 5B exhibits a clear degradation across four equally sized Motion Structure Difficulty Score (MSDS) bins (see Appendix B.3 for details), with FVD increasing from 848.34 in the Low bin to 1037.41 in the Extreme bin, a 22.3% increase (Figure 1a). We argue that this degradation exposes a fundamental assumption in standard DiT tokenization: a uniform patch grid allocates the same spatial token budget everywhere, despite spatially varying representation demands. Complex human motion makes this mismatch particularly consequential. A matched-budget control further shows that allocation itself matters: under the same number of added tokens, motion-adaptive placement reduces FVD from 881.50 to 859.95 compared with random placement (Figure 1b). The key is therefore not merely increasing representation capacity, but deciding where that capacity is allocated. This observation motivates our transition from uniform to motion-adaptive representation allocation.

We introduce Motion-Adaptive Token Allocation (MATA), a motion-adaptive representation allocation framework for complex human image animation. Rather than treating spatial locations as having equal representation demands, MATA redistributes representation capacity according to the motion requirements of each region while keeping the overall computational budget within two tokens of the dense baseline at packed steps. This perspective differs from reduction-oriented adaptive-token methods (Chang et al., 2026; Peruzzo et al., 2025; Kim et al., 2026): our objective is not primarily to reduce computation, but to use a constrained representation budget where it is most valuable for motion modeling. Motion cues determine how representation capacity is spatially reallocated, while compression of redundant low-motion regions serves as a budget-balancing mechanism that compensates for this redistribution. In this way, MATA converts a uniform representation scheme into a motion-adaptive one without requiring a global increase in backbone computation.

We evaluate MATA on HyperMotionX using a unified evaluation pipeline. Our 5B model improves all eight reported metrics over UniAnimate-DiT 5B, including a 7.97% reduction in FVD and a 6.54-percentage-point increase in $\mathrm{PCK}_{\mathrm{bbox}}@0.05$, while estimated main-DiT FLOPs remain within 1.29% of the dense configuration. More importantly, difficulty-stratified analysis shows that the benefit of representation reallocation is strongest for challenging motion: the full model achieves its largest pose-accuracy gain in the Extreme MSDS bin, improving $\mathrm{PCK}_{\mathrm{bbox}}@0.05$ by 8.08 percentage points over the motion-LoRA baseline. This result supports our central hypothesis that complex motion creates spatially non-uniform representation demands, and that reallocating a constrained representation budget toward those demands is more effective than treating all regions uniformly. Our contributions are threefold:

- We identify an overlooked representation-allocation problem in complex human image animation: motion induces spatially non-uniform representation demands, whereas standard DiTs allocate representation capacity uniformly across spatial regions. This mismatch motivates rethinking complex-motion generation from the perspective of representation allocation rather than motion conditioning alone.

- We propose Motion-Adaptive Token Allocation (MATA), a motion-adaptive representation paradigm that redistributes a constrained representation budget according to spatially varying motion demands. This formulation enables the model to concentrate representation capacity where it is most needed while keeping the overall backbone computation within two tokens of the dense baseline at packed steps.
- Experiments on HyperMotionX validate the proposed allocation principle: MATA achieves its strongest structural improvement on the most challenging motion split, while delivering strong overall performance with nearly unchanged backbone computation. These results validate motion-adaptive representation allocation for complex human animation under a constrained compute budget.

2 Related Work

2.1 Human Image Animation

Human image animation methods primarily improve appearance preservation, motion control, and temporal consistency through reference pathways, pose guidance, and video modeling (Hu, 2024; Xu et al., 2024; Zhu et al., 2024; Zhang et al., 2024; Tu et al., 2025; Wang et al., 2025a; Tan et al., 2024). Recent video DiTs strengthen these capabilities through large pretrained backbones (Gan et al., 2025; Wang et al., 2025b; Cheng et al., 2025). For complex articulation, the challenge extends to interpreting occlusion and changing body-part relationships. HyperMotionX addresses this through complex-motion data and enhanced spatial positional modeling (Xu et al., 2025), while SCAIL strengthens geometric guidance through 3D-consistent pose representations (Yan et al., 2026b). SCAIL-2 further explores end-to-end in-context conditioning (Yan et al., 2026a), while other recent methods learn motion representations directly from driving videos (Zeng et al., 2026; Xu et al., 2026).

These advances improve motion information and structural modeling, but recent DiT-based systems generally retain uniform spatial tokenization. Consequently, stronger guidance does not directly increase the sampling density around rapidly changing limbs and occlusion boundaries. MATA addresses this complementary issue by using pose changes to allocate denser local observations within the backbone.

2.2 Adaptive Tokenization in Diffusion Transformers

Adaptive-token methods primarily use non-uniform processing to reduce diffusion cost. Token Merging combines similar features (Bolya et al., 2022; Bolya & Hoffman, 2023); DyDiT and DiffCR route computation across tokens, layers, or timesteps (Zhao et al., 2025; You et al., 2025); and FlexiDiT and DDiT adapt patch granularity during denoising (Anagnostidis et al., 2025; Kim et al., 2026). Spatially adaptive methods further recognize unequal regional demands: D^2iT assigns finer image representations to detailed regions (Jia et al., 2025), while DynaPatch selects video patch sizes from content complexity for efficient generation (Li et al., 2026). ADAPTOR uses motion to retain more original tokens in moving regions and compresses relatively static regions to accelerate video diffusion (Peruzzo et al., 2025).

For complex human animation, uniform tokenization creates an allocation mismatch between local motion demand and representation capacity. MATA directs finer observations to complex motion regions and uses reference-aware compression to recover their budget elsewhere, improving motion structure at approximately unchanged backbone compute.

3 Method

3.1 Overview of Representation Redistribution

Motion-Adaptive Token Allocation (MATA) replaces the uniform spatial representation of a video DiT with a motion-adaptive allocation while keeping the overall computational budget within two tokens of the dense baseline at packed steps. As illustrated in Figure 2, MATA treats representation

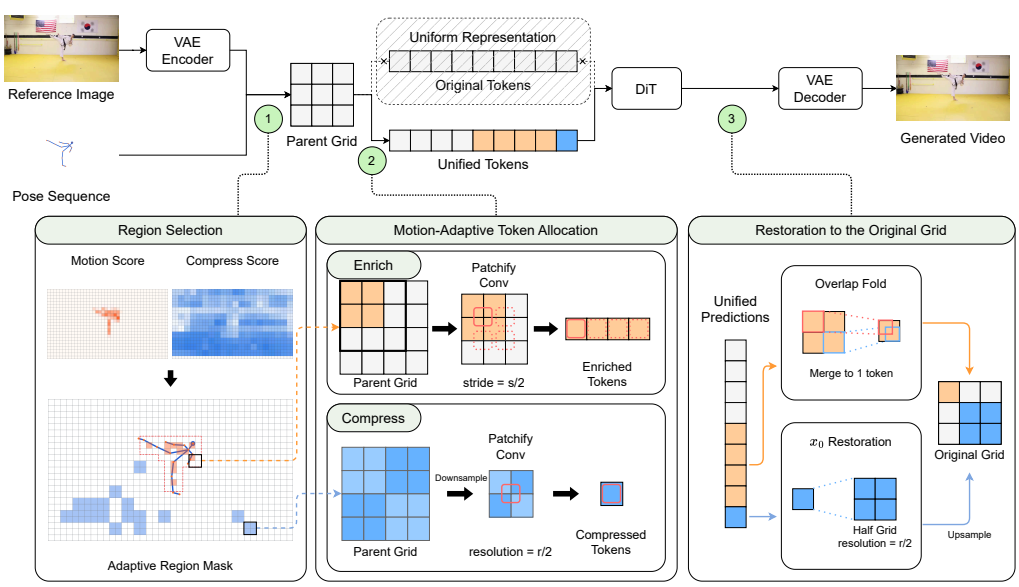

Figure 2: Overview of Motion-Adaptive Token Allocation (MATA). MATA replaces uniform spatial representation with motion-adaptive redistribution under a computational budget. Motion demand determines where representation capacity is concentrated, while redundant regions provide recoverable capacity for budget balancing. The redistributed representation is processed by the original DiT backbone and then restored to the canonical latent grid for prediction.

capacity as a shared spatial resource: motion demand determines where this capacity should be concentrated, while redundant regions provide recoverable capacity to balance the redistribution.

The framework follows a simple allocation–processing–restoration pipeline. We first estimate motion-driven representation demand together with regions suitable for budget recovery. The available representation capacity is then redistributed accordingly and processed jointly by the original DiT backbone. After backbone processing, the adaptive representation is mapped back to the canonical latent grid for prediction. Importantly, compression serves only to maintain the allocation budget; the quality-oriented decision is where representation capacity is assigned according to motion demand.

3.2 Demand-Aware Allocation Policy

Our policy selects regions for finer-grained token enrichment and regions for compression to recover the corresponding token budget. The two decisions use complementary cues: pose motion determines where additional representation is allocated, while reference content determines which regions are selected for budget recovery.

Motion-driven demand. We compute weighted absolute differences between consecutive semantic pose masks to measure local motion. These differences are temporally smoothed and aggregated to match the latent frames, then spatially pooled onto the parent-token grid. We select high-scoring regions and expand the selection to neighboring grid locations to cover surrounding body structure, subject to a fixed enrichment budget. This produces a pose-dependent allocation mask that remains fixed across denoising steps.

Recoverable capacity. Compression candidates are drawn from low-motion regions outside a spatial margin around the enrichment regions. To account for spatial detail, we downsample and upsample the clean reference latent and measure the local reconstruction error. A lower error indicates that the reference features are better preserved at reduced spatial resolution. We therefore prioritize low-error candidates for compression, recovering token capacity for motion-driven enrichment. Reference content guides compression selection, while enrichment selection is determined by the

driving pose. Channel weights, temporal smoothing, selection budgets, and the reference reconstruction score are detailed in Appendix A.1.

3.3 Budget-Preserving Token Realization

We next instantiate the allocation policy as a non-uniform token representation. The key operation is to provide finer spatial observations at high-demand locations while preserving the original DiT backbone and keeping its packed token budget close to the dense baseline.

Fine representation for high-demand regions. The baseline PatchEmbed uses a convolution kernel and spatial stride of $(1, 2, 2)$. For enrichment, we introduce a separately learnable copy of this operator, initialized from the pretrained PatchEmbed weights. We apply this copy with unit spatial stride while retaining the original parent token:

$$\begin{aligned} \mathcal{E}_\sigma &= \mathrm{Conv3D}_{k=(1,2,2),\, s=(1,1,1)}(Z_\sigma; W_E^+, b_E^+), \\ e_{i,\mathrm{in}}^{ab} &= \mathcal{E}_\sigma[:, :, f, 2y+a, 2x+b] + s^{ab}, \qquad a, b \in \{0, 1\}, \end{aligned} \tag{1}$$

where parent i is located at (f, y, x), and (W_E^+, b_E^+) are initialized from (W_E, b_E) and optimized independently. The four overlapping observations expose local structure at different spatial offsets, providing a finer representation around the selected region without replacing its original parent token. Their rotary coordinates follow the corresponding fractional spatial locations (Su et al., 2024). Boundary handling, sub-position encoding, and conditioning details are deferred to Appendix A.2.

Budget recovery. The additional representation capacity is compensated by compressing selected redundant regions. Each eligible 2×2 parent group is represented by a single token constructed from the downsampled noisy latent using the frozen baseline PatchEmbed. We select enough recovery groups to keep the packed sequence within a negligible difference from the baseline token count. Thus, compression serves specifically as a budget-balancing operation rather than an independent mechanism for improving generation quality. The exact token-balancing rule and edge cases are given in the appendix.

All retained parent tokens, fine representations, and budget-recovery tokens are processed jointly by the original DiT blocks. The backbone therefore operates on a spatially redistributed representation while retaining nearly the same overall token budget.

3.4 Restoration to the Canonical Grid

The redistributed representation exists only inside the DiT backbone. Before prediction, it must be returned to the original parent grid so that motion-adaptive internal representation remains compatible with the pretrained output head.

Restoring fine representations. After the final DiT block, the fine representations associated with each selected region are deterministically folded back to their parent locations using overlap-aware aggregation. The aggregated representation is combined with the corresponding parent state before the fine tokens are removed. This operation introduces no additional trainable parameters; the exact folding weights are specified in Appendix A.3.

x_0-consistent restoration of compressed regions. Compressed regions require reconstruction from a reduced spatial grid. Directly prolongating their velocity predictions produces inconsistent clean predictions because downsampling and prolongation do not preserve the current noisy latent. We instead perform restoration in the predicted x_0 domain.

Let $\widehat{V}_f$ denote predictions available on the canonical grid, $\widehat{V}_c$ the compressed-grid predictions, and let M_f and M_c denote the corresponding compression masks. With $\mathcal{R}$ denoting 2×2 spatial reduction and $\mathcal{P}$ its bilinear prolongation, we compute

$$\begin{aligned} \overline{V}_c &= M_c \odot \widehat{V}_c + (1 - M_c) \odot \mathcal{R}(\widehat{V}_f), \\ \widehat{Z}_{0,c} &= \mathcal{R}(Z_\sigma) - \sigma \overline{V}_c, \\ \widehat{V} &= (1 - M_f) \odot \widehat{V}_f + M_f \odot \frac{Z_\sigma - \mathcal{P}(\widehat{Z}_{0,c})}{\sigma}. \end{aligned} \tag{2}$$

Table 1: Quantitative results on 100 HyperMotionX test videos using a unified evaluator. The 5B progression isolates motion adaptation, motion-guided enrichment, and the complete budget-constrained redistribution. FID/FVD are means of per-video distances and $\mathrm{PCK}_{\mathrm{bbox}}@0.05$ is reported in percent.

Method	PSNR↑	SSIM↑	L1↓	LPIPS↓	FID↓	VFID↓	FVD↓	$\mathrm{PCK}_{\mathrm{bbox}}@0.05$(%)↑
Controlled 5B progression								
UniAnimate-DiT 5B	21.84	0.74	0.0471	0.180	90.86	15.22	959.16	59.35
Motion LoRA only	22.20	0.76	0.0448	0.170	86.84	13.78	891.94	60.60
+ Enrichment	22.27	0.75	0.0448	0.169	82.74	12.81	859.95	65.20
+ Enrichment & compression (MATA 5B)	22.20	0.75	0.0456	0.176	85.41	12.57	882.73	65.89
Comparison at the 14B scale								
HyperMotion 14B	21.88	0.74	0.0498	0.162	**78.28**	10.95	808.92	73.75
SCAIL-2 14B	19.16	0.59	0.0843	0.288	83.94	12.18	1076.00	74.43
MATA 14B	**23.62**	**0.78**	**0.0396**	**0.161**	81.67	**10.14**	**733.81**	**75.58**

This formulation restores the clean prediction consistently with the current latent while returning all predictions to the canonical grid. The derivation and comparison with direct velocity prolongation are given in Appendix A.3.

During the final denoising stage, we disable compression and return to the dense parent grid to recover fine spatial detail, while retaining the motion-adaptive fine representation. The detailed sampling schedule is provided in Appendix B.4.

4 Experiments

4.1 Experimental Setup

We train on 19,586 video–pose pairs from the Open-HyperMotionX Dataset and evaluate on 100 videos from HyperMotionX Bench (Xu et al., 2025). We adapt motion LoRA (Hu et al., 2021), the pose encoder, the enrichment PatchEmbed kernel, and sub-position embeddings with flow matching (Lipman et al., 2022) while keeping the original DiT weights frozen, and use 50-step inference. Comparisons include UniAnimate-DiT 5B, HyperMotion 14B, and pose-driven SCAIL-2 14B (Wang et al., 2025b; Xu et al., 2025; Yan et al., 2026a). Training, preprocessing, and model-specific sampling details are provided in Appendix B.

All methods are recomputed with a unified evaluator. We report PSNR, SSIM (Wang et al., 2004), L1, LPIPS (Zhang et al., 2018), FID (Heusel et al., 2017), VFID, FVD (Unterthiner et al., 2018), and $\mathrm{PCK}_{\mathrm{bbox}}@0.05$ (Bergman et al., 2022). FID/FVD are averaged over per-video distances; pooled global FVD and complete metric definitions are reported in Appendix B.6 and Appendix B.2, respectively.

4.2 Does Motion-Adaptive Allocation Improve Complex Human Animation?

Our primary question is whether changing spatial representation allocation provides benefits beyond stronger motion conditioning. We therefore place the controlled 5B progression directly alongside the comparison with existing methods.

Controlled effect of adaptive representation. The 5B progression isolates the two roles in MATA. Motion LoRA first strengthens motion conditioning without changing the spatial representation. Directing finer representation to motion-demand regions then reduces FVD from 891.94 to 859.95 and increases $\mathrm{PCK}_{\mathrm{bbox}}@0.05$ from 60.60% to 65.20%. This controlled comparison shows that motion-guided enrichment provides further gains over motion adaptation alone.

Adding compression converts this quality-oriented enrichment into the complete budget-constrained redistribution. The full model attains the highest pose accuracy (65.89%) and lowest VFID (12.57) among the controlled variants while trading part of the enrichment-only gain in FVD and LPIPS for a nearly unchanged token budget. Relative to the original UniAnimate-DiT 5B, the complete method improves all eight reported metrics, including a 7.97% reduction in FVD and a 6.54-percentage-point gain in $\mathrm{PCK}_{\mathrm{bbox}}@0.05$.

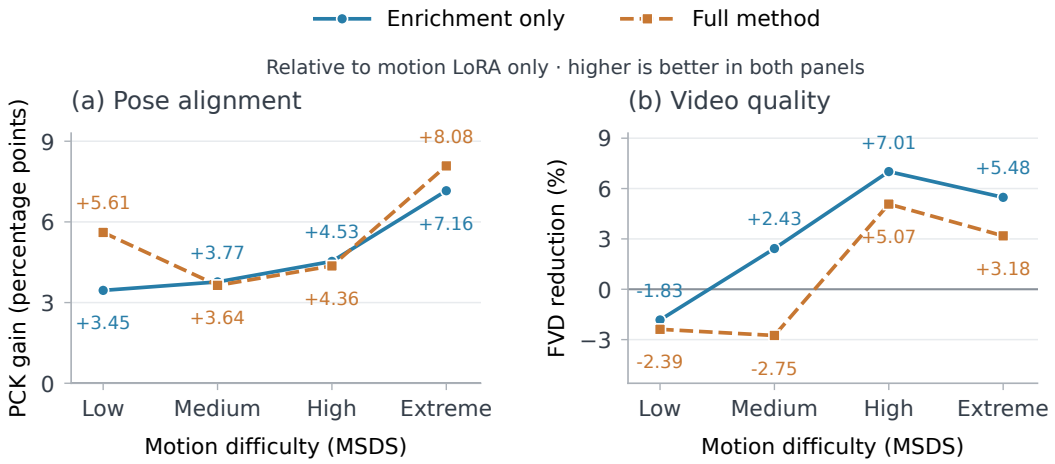

Figure 3: Difficulty-stratified gains over Motion LoRA with the 5B backbone across four MSDS bins (25 videos each). (a) $\mathrm{PCK}_{\mathrm{bbox}}@0.05$ gain in percentage points; (b) relative reduction in mean per-video FVD. Positive values indicate improvement.

Comparison at larger scale. Scaling MATA to 14B further improves all eight metrics over its 5B configuration. Compared with HyperMotion 14B, MATA 14B reduces FVD from 808.92 to 733.81 and VFID from 10.95 to 10.14, while increasing $\mathrm{PCK}_{\mathrm{bbox}}@0.05$ from 73.75% to 75.58%. It also improves PSNR, SSIM, L1, and LPIPS, with FID remaining higher than HyperMotion. Against SCAIL-2 14B, MATA improves all eight reported metrics. Pooled global FVD, which gives a different aggregation of distributional quality, is reported separately in Appendix B.6.

Benefit under challenging motion. We further examine whether finer motion-adaptive representation becomes more useful as articulation becomes more difficult. Figure 3 reports gains over the same Motion-LoRA baseline across four equally sized MSDS bins. For enrichment alone, the $\mathrm{PCK}_{\mathrm{bbox}}@0.05$ gain increases from 3.45 to 3.77, 4.53, and 7.16 percentage points from Low to Extreme. The complete method improves pose alignment in every bin and reaches its largest gain in the Extreme split (8.08 points). FVD does not follow the same monotonic trend, but improves in the High and Extreme bins. These results show that the structural gain is largest under challenging articulation.

Qualitative evidence. Figure 4 shows that the quantitative gains correspond to improved local structure under difficult configurations. In inverted poses, HyperMotion confuses the body-part layout, leaving the blue shorts in the lower image region rather than following the raised legs, whereas MATA follows the inverted configuration more faithfully. Other examples show improved head orientation, separation of crossing limbs, and preservation of clothing placement under large articulation.

4.3 Does Where We Allocate Representation Capacity Matter?

The previous results establish the benefit of adaptive representation. We next test whether its spatial placement matters, separating the quality-oriented allocation decision from the recovery-side decision used to satisfy the budget.

Motion-side allocation. To isolate token placement from token count, we replace motion-guided selection with random selection at inference using the same enrichment checkpoint and exactly the same number of added tokens. Motion-guided allocation reduces FVD from 881.50 to 859.95 and VFID from 13.43 to 12.81, while LPIPS and thresholded pose accuracy remain nearly unchanged. Thus, the improvement cannot be explained by adding more tokens alone: where the additional representation is placed affects video-level quality.

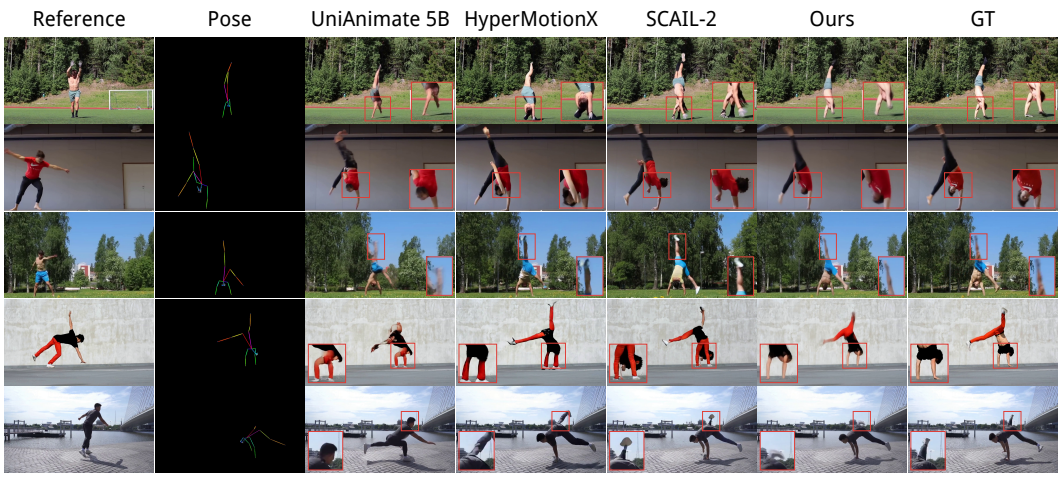

Figure 4: Qualitative comparison on challenging human motions. The examples emphasize inversion, limb crossing, and appearance preservation under large structural changes. UniAnimate-DiT uses 5B; HyperMotion, SCAIL-2, and MATA use 14B.

Table 2: Ablations of the allocation policy with the 5B backbone. Motion-side variants use the same enrichment-only checkpoint and the same number of added tokens. Recovery-side variants use the complete method and differ only in reference-based ranking.

Variant	LPIPS↓	FID↓	VFID↓	FVD↓	PCK_{bbox}@0.05(%)↑
Where to allocate capacity					
Random regions	0.169	82.93	13.43	881.50	65.38
Motion-guided regions	0.169	82.74	12.81	859.95	65.20
Where to recover capacity					
Without reference ranking	0.182	87.02	12.52	882.75	65.86
With reference ranking	0.176	85.41	12.57	882.73	65.89

Recovery-side allocation. Reference ranking has a smaller but complementary role. Removing it increases LPIPS from 0.176 to 0.182 and FID from 85.41 to 87.02, while FVD and pose accuracy remain almost unchanged. This behavior is consistent with its intended purpose: the reference cue does not decide where additional representation is needed, but helps avoid reclaiming capacity from visually detailed regions.

Complete eight-metric results and additional allocation-budget settings are reported in Appendix C.

4.4 Can Redistribution Preserve Quality under a Constrained Budget?

The final question is whether the adaptive representation can be returned to the canonical grid without negating its gains. We examine the two mechanisms that directly address this constraint: x_0-consistent restoration and late dense refinement.

Restoration consistency. Restoration domain has a large effect on visual quality. Replacing direct velocity restoration with the x_0-consistent formulation reduces FVD from 1066.42 to 882.73, VFID from 17.61 to 12.57, and FID from 100.47 to 85.41, while pose accuracy remains essentially unchanged. Visual metrics improve substantially while pose accuracy changes by only 0.02 percentage points, isolating the effect to reconstruction consistency rather than motion control.

Figure 5 illustrates the corresponding failure mode. Direct velocity restoration introduces colored block artifacts in background regions, whereas x_0-consistent restoration removes these discontinuities while preserving the generated pose.

Table 3: Ablations of restoration and late dense refinement with the 5B backbone. The complete configuration uses x_0-consistent restoration and a ten-step dense tail.

Variant	LPIPS↓	FID↓	VFID↓	FVD↓	$\mathrm{PCK}_{\mathrm{bbox}}$@0.05(%)↑
Restoration domain					
Velocity-domain restoration	0.196	100.47	17.61	1066.42	65.87
x_0-consistent restoration	0.176	85.41	12.57	882.73	65.89
Dense-refinement schedule					
Tail-0	0.180	87.20	12.87	890.76	65.77
Tail-5	0.177	86.02	12.53	883.71	65.94
Tail-10	0.176	85.41	12.57	882.73	65.89

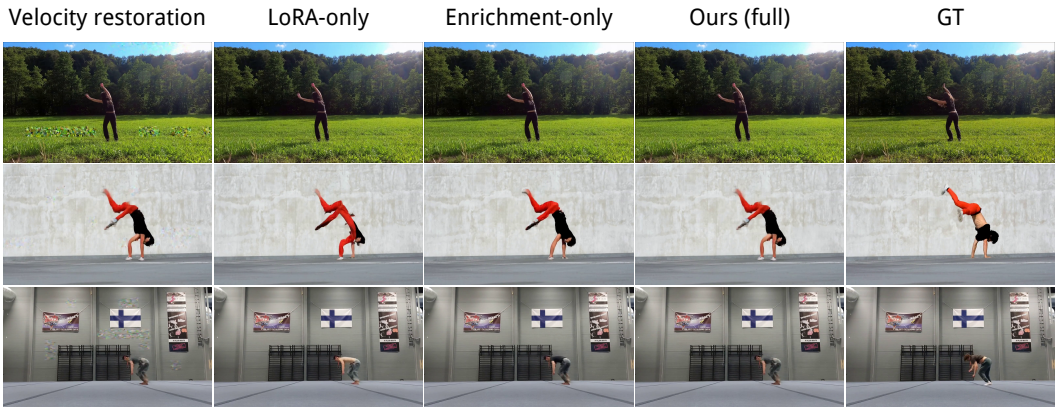

Figure 5: Qualitative diagnostics with the 5B backbone. Direct velocity restoration produces colored block artifacts in reconstructed regions, while x_0-consistent restoration suppresses these artifacts. The examples also illustrate the local structural improvement obtained from motion-adaptive enrichment.

Late dense refinement. Returning to the dense parent grid near the end of denoising further improves local reconstruction. Increasing the dense tail from zero to five steps reduces FVD from 890.76 to 883.71, with ten steps reaching 882.73. LPIPS follows the same trend, decreasing from 0.180 to 0.176. Most of the improvement is already obtained with five steps, while the ten-step setting is used as the default. Further schedule results are provided in Appendix C.4.

Computational cost. The complete redistribution requires 15.678 estimated main-DiT PFLOPs per video, compared with 15.478 PFLOPs for the dense configuration, an increase of only 1.29%. Thus, the complete method converts the quality gain from motion-adaptive representation into a nearly compute-neutral redistribution rather than relying on a uniform increase in backbone capacity. Runtime measurements and detailed FLOPs accounting are provided in Appendix B.5.

5 Conclusion

We revisited complex human image animation as a representation-allocation problem: complex motion creates spatially non-uniform representation demands, whereas standard video DiTs allocate capacity uniformly. We therefore introduced MATA, which redistributes a constrained representation budget according to motion demand while keeping backbone computation approximately unchanged. On HyperMotionX, motion-adaptive enrichment improves over motion LoRA alone, with the largest structural gain appearing on the most challenging motion split. The complete 5B model improves all eight reported metrics over UniAnimate-DiT with only 1.29% additional estimated main-DiT FLOPs, while the 14B model improves seven of eight metrics over HyperMotion 14B. These results support motion-adaptive representation allocation as an effective principle for complex human animation.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, including improvements to grammar, clarity, and readability. The research ideas, methodology, experiments, and conclusions were developed by the authors. All AI-assisted text was reviewed and revised by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The main paper presents the motion-adaptive allocation, token realization, and canonical-grid restoration used in MATA. The appendix provides the corresponding implementation details, including motion-demand selection, budget recovery, boundary handling, training and sampling settings, and evaluation protocols. It also reports supplementary results, including global FVD and complete ablation details.

REFERENCES

Sotiris Anagnostidis, Gregor Bachmann, Yeongmin Kim, Jonas Kohler, Markos Georgopoulos, Artsiom Sanakoyeu, Yuming Du, Albert Pumarola, Ali Thabet, and Edgar Schönfeld. FlexiDiT: Your diffusion transformer can easily generate high-quality samples with less compute. In *CVPR*, pp. 28316–28326. IEEE, 2025.

Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *NeurIPS*, 35:19900–19916, 2022.

Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *CVPRW*, pp. 4599–4603. IEEE, 2023.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Shuning Chang, Pichao Wang, Jiasheng Tang, Fan Wang, and Yi Yang. Sparsedit: Token sparsification for efficient diffusion transformer. *NeurIPS*, 38:91456–91483, 2026.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

Qijun Gan, Yi Ren, Chen Zhang, Zhenhui Ye, Pan Xie, Xiang Yin, Zehuan Yuan, Bingyue Peng, and Jianke Zhu. Humandit: Pose-guided diffusion transformer for long-form human motion video generation. *arXiv preprint arXiv:2502.04847*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, pp. 8153–8163, 2024.

Weinan Jia, Mengqi Huang, Nan Chen, Lei Zhang, and Zhendong Mao. D^2iT: Dynamic diffusion transformer for accurate image generation. In *CVPR*, pp. 12860–12870. IEEE, 2025.

Dahye Kim, Deepti Ghadiyaram, and Raghudeep Gadde. Ddit: Dynamic patch scheduling for efficient diffusion transformers. *arXiv preprint arXiv:2602.16968*, 2026.

Sheng Li, Connelly Barnes, Mamshad Nayeem Rizve, Hongwu Peng, Zhengang Li, Ohi Dibua, Alireza Ganjdanesh, Xulong Tang, Yan Kang, and Yifan Gong. Content-aware dynamic patchification for efficient video diffusion. In *CVPR*, pp. 35936–35945, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4172–4182. IEEE, 2023.

Elia Peruzzo, Adil Karjauv, Nicu Sebe, Amir Ghodrati, and Amir Habibian. Adaptor: Adaptive token reduction for video diffusion transformers. In *CVPR*, pp. 6419–6425, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10674–10685. ieee, 2022.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. *arXiv preprint arXiv:2410.10306*, 2024.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. In *CVPR*, pp. 21096–21106. IEEE, 2025.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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

Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *Science China Information Sciences*, 68(10):200103, 2025a.

Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. UniAnimate-DiT: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025b. URL https://arxiv.org/abs/2504.11289.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Shuolin Xu, Siming Zheng, Ziyi Wang, HC Yu, Jinwei Chen, Huaqi Zhang, Daquan Zhou, Tong-Yee Lee, Bo Li, and Peng-Tao Jiang. HyperMotionX: The dataset and benchmark with dit-based pose-guided human image animation of complex motions. *arXiv preprint arXiv:2505.22977*, 2025.

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 *CVPR*, pp. 1481–1490. IEEE, 2024.

Zhufeng Xu, Xuan Gao, Feng-Lin Liu, Haoxian Zhang, Zhixue Fang, Yu-Kun Lai, Xiaoqiang Liu, Pengfei Wan, and Lin Gao. Im-animation: an implicit motion representation for identity-decoupled character animation. *arXiv preprint arXiv:2602.07498*, 2026.

Wenhao Yan, Fengjia Guo, Zhuoyi Yang, and Jie Tang. Scail-2: Unifying controlled character animation with end-to-end in-context conditioning. *arXiv preprint arXiv:2606.10804*, 2026a.

Wenhao Yan, Sheng Ye, Zhuoyi Yang, Jiayan Teng, ZhenHui Dong, Kairui Wen, Xiaotao Gu, Yong-Jin Liu, and Jie Tang. Scail: Towards studio-grade character animation via in-context learning of 3d-consistent pose representations. In *CVPR*, pp. 4450–4460, 2026b.

Haoran You, Connelly Barnes, Yuqian Zhou, Yan Kang, Zhenbang Du, Wei Zhou, Lingzhi Zhang, Yotam Nitzan, Xiaoyang Liu, Zhe Lin, et al. Layer-and timestep-adaptive differentiable token compression ratios for efficient diffusion transformers. In *CVPR*, pp. 18072–18082. IEEE, 2025.

Yuan Zeng, Yujia Shi, Yuhao Yang, Dongxia Liu, Zongqing Lu, Wenming Yang, and Qingmin Liao. Beyond skeletons: Learning animation directly from driving videos with same2x training strategy. *arXiv preprint arXiv:2606.06903*, 2026.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595. IEEE, 2018.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

Wangbo Zhao, Yizeng Han, Jiasheng Tang, Kai Wang, Yibing Song, Gao Huang, Fan Wang, and Yang You. Dynamic diffusion transformer. In *ICLR*, volume 2025, pp. 65520–65552, 2025.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *ECCV*, pp. 145–162. Springer, 2024.

A Implementation Details

This section provides the implementation details omitted from the main paper. We follow the same hierarchy as the method: estimating spatial representation demand, realizing the redistribution under a constrained budget, and restoring the adaptive representation to the canonical grid.

A.1 Demand-Aware Allocation Policy

Motion demand. The binary semantic pose map is $P \in \{0,1\}^{5\times T\times H_0\times W_0}$, with channels for torso/head, right arm, left arm, right leg, and left leg. With channel weights $w = (1,2,2,2,2)$, adjacent-frame motion magnitude is

$$\delta_t(u) = \left[\sum_{c=1}^{5} w_c^2 \big(P_{c,t+1}(u) - P_{c,t}(u)\big)^2\right]^{1/2}. \tag{3}$$

We average every four adjacent-frame differences to match VAE frames, use the first forward difference for the reference frame, and smooth temporally with $(0.25, 0.5, 0.25)$. Adaptive average pooling maps the scores to the parent-token grid. The last parent row and column are excluded; candidates above 0.3 of the eligible frame maximum are expanded by a 3×3 neighborhood and ranked by their original motion scores. The reference anchor is excluded. Default limits are 12 selected parents per latent frame, a global selected-parent ratio of 2%, and an added fine-token ratio of 8%. Selection is discrete, receives no gradients, and is reused across denoising steps.

Recoverable capacity. Let E_f be the selected motion-demand mask and $\mathcal{D}_1(E_f)$ its one-parent-radius dilation. A 2×2 parent group g is eligible for budget recovery when $f \geq 1$, its top-left coordinates are even, it contains four valid parents, and $g \cap \mathcal{D}_1(E_f) = \varnothing$. Low motion alone does not imply low spatial complexity, so eligible groups are ranked by reconstruction error in the clean reference latent R:

$$q_g = \frac{\mathrm{RMS}_g(R - \mathcal{P}(\mathcal{R}(R)))}{\max(\mathrm{RMS}_g(R), 10^{-6})}, \tag{4}$$

where $\mathcal{R}$ is 2×2 average pooling and $\mathcal{P}$ is bilinear prolongation. Lower-error groups are preferred. Equal-score candidates are balanced across spatial quadrants with deterministic index order. The per-frame candidate cap is $\lfloor r_c \lfloor H_p/2\rfloor \lfloor W_p/2\rfloor\rfloor$ with $r_c = 0.125$. This ranking is used only for budget recovery; motion scores alone determine where additional representation is allocated.

A.2 Budget-Preserving Token Realization

Fine-token extraction and boundaries. For each selected parent, fine tokens are extracted with unit spatial stride using a separately trainable kernel initialized from the pretrained PatchEmbed kernel. The resulting channel-first embedding field has layout $B\times D\times F\times (H_\ell - 1)\times (W_\ell - 1)$. Only selected windows are evaluated. Fine-token extraction uses no padding, so excluding the last parent row and column keeps all four stride-one windows inside the latent grid. Each offset uses its fractional rotary coordinate and a learned sub-position embedding. Fine tokens receive pose information through interaction with conditioned parent tokens rather than a separate pose stream. The spatial kernel size remains 2×2; the finer representation comes from shifted overlapping observations.

Budget recovery and exact balancing. Eligible 2×2 parent groups are spatially average-pooled in the noisy latent and passed through the frozen baseline PatchEmbed to form one compressed token. Its pose condition is the average of the four parent conditions and its rotary coordinate is placed at the group center. Boundary parents that do not form complete groups remain on the canonical grid. Spatial prolongation uses bilinear interpolation with `align_corners=False`.

For K selected motion-demand parents, four fine tokens are added per parent, while each compressed group recovers three token positions. We therefore choose

$$G = \left\lfloor \frac{4K}{3} \right\rfloor, \qquad S_{\mathrm{pack}} = N + 4K - 3G = N + (K \bmod 3). \tag{5}$$

Table 4: Sequence-length and frame-rate distributions of the 100-video test set.

Frames	81	65	49	33	25
Videos	53	17	16	10	4

FPS	16	12
Videos	66	34

The packed sequence differs from the baseline by at most two tokens. If fewer than G recovery groups are available, K is reduced and eligibility is recomputed until the allocation is feasible; if none remains, the dense path is used.

A.3 Canonical-Grid Restoration

Deterministic overlap folding. The four stride-one windows overlap the original parent patch with areas in the ratio $4:2:2:1$. After the final DiT block,

$$\bar{e}_{i,\mathrm{out}} = \frac{4e^{00}_{i,\mathrm{out}} + 2e^{01}_{i,\mathrm{out}} + 2e^{10}_{i,\mathrm{out}} + e^{11}_{i,\mathrm{out}}}{9}, \qquad p^{\star}_i = p_{i,\mathrm{out}} + 0.5(\bar{e}_{i,\mathrm{out}} - p_{i,\mathrm{out}}). \tag{6}$$

The fold retains half of the propagated parent state, introduces no learned projection, and removes fine tokens before the original output head.

Clean-prediction consistency. Within a compressed region, the restoration in the main paper enforces $Z_\sigma - \sigma\widehat{V} = \mathcal{P}(\widehat{Z}_{0,c})$ for $\sigma > 0$. Direct velocity prolongation instead leaves the resampling residual $Z_\sigma - \mathcal{P}(\mathcal{R}(Z_\sigma))$. Equivalently,

$$\widehat{V} = \mathcal{P}(\overline{V}_c) + \frac{Z_\sigma - \mathcal{P}(\mathcal{R}(Z_\sigma))}{\sigma}. \tag{7}$$

This follows from $\widehat{Z}_{0,c} = \mathcal{R}(Z_\sigma) - \sigma\overline{V}_c$ and linearity of bilinear prolongation. The additional term is determined by the current noisy latent and the resampling operators. Dense refinement near the clean endpoint avoids relying on the $1/\sigma$ correction at very low noise.

B Dataset and Experimental Protocols

B.1 Data and Evaluation

We train on 19,586 paired human videos and pose-control sequences from HyperMotionX (Xu et al., 2025). Native frame rates are retained; 64.7% of videos are recorded at 25 FPS, 17.0% at 30 FPS, and 9.1% at 29.97 FPS. Each training example contains at most 81 consecutive frames with temporal stride one; shorter sequences repeat the final frame. After random cropping, inputs are resized to 832×480 or 480×832.

The test set contains 100 videos. All methods generate at 1280×736 or 736×1280 while retaining the frame rate of the pose sequence. The reference frame is excluded from visual metrics, giving 6,512 evaluated frames; $\mathrm{PCK}_{\mathrm{bbox}}@0.05$ uses all 6,612 aligned frames.

B.2 Evaluation Metrics

L1 is RGB mean absolute error in $[0, 1]$. FID uses InceptionV3-2048 frame features and is computed separately for each video. FVD uses I3D features from non-overlapping 16-frame clips resized to 224×224 and follows the same per-video aggregation:

$$\overline{D}_{\mathrm{video}} = \frac{1}{n}\sum_{i=1}^{n} D(\mathcal{F}_i^{\mathrm{GT}}, \mathcal{F}_i^{\mathrm{gen}}), \tag{8}$$

where D is the Fréchet distance between fitted feature means and covariances. VFID instead uses Kinetics-400 3D ResNet-50 features from 16-frame 112×112 clips, averages clip features within each video, and computes one global distance over video features. Incomplete trailing clips are padded by repeating the final frame; the 100-video test set contains 409 clips in total. All methods are recomputed with the same evaluator.

For pose accuracy, let (x_{tk}, y_{tk}) and $(\widehat{x}_{tk}, \widehat{y}_{tk})$ denote GT and predicted coordinates, and w_t, h_t the GT body-box width and height. We compute

$$d_{tk} = \sqrt{\left(\frac{\widehat{x}_{tk} - x_{tk}}{w_t}\right)^2 + \left(\frac{\widehat{y}_{tk} - y_{tk}}{h_t}\right)^2}, \tag{9}$$

and count a keypoint as correct when $d_{tk} < 0.05$. The main score pools valid keypoints across videos. We use X-Pose with the Swin-T configuration and `unipose_swint.pth` for 17 body keypoints; missing predictions count as incorrect and invalid normalization boxes are excluded. All five methods have 112,047 valid evaluated keypoints.

Aggregation scope. Per-video FVD measures distributional discrepancy within each matched video and then averages over the benchmark, whereas the global FVD below fits one distribution to all pooled clip features. Likewise, the reported pose score pools valid keypoints rather than averaging per-video percentages. These choices are kept fixed for every compared method.

Per-video and global FVD. The main text reports the mean of 100 matched per-video FVD values. Appendix B.6 additionally pools all 409 I3D clips across the test set and computes one global FVD. The two aggregations use the same features and preprocessing but characterize different statistical scopes.

B.3 Motion-Difficulty Stratification

MSDS computation. We compute MSDS from DWPose body-and-foot keypoints extracted from GT videos. Coordinates are normalized by image width and height, centered at the per-frame body-box center, and divided by the median body-box diagonal over the video. Invalid detections are excluded, and temporal smoothing is disabled.

The raw motion amplitude is $A^{\text{raw}} = 0.3A_{\text{joint}} + 0.4A_{\text{limb}} + 0.3A_{\text{body}}$. For each joint, its amplitude is the Euclidean norm of the coordinate-wise 90th–10th percentile spans. A_{joint} averages valid joint amplitudes; A_{limb} uses weights of 1 for wrists and ankles and 0.5 for elbows and knees. A_{body} is the corresponding percentile span of body-box area normalized by squared body scale.

We further compute four structural components. $D^{\text{raw}}_{\text{cell}}$ averages wrist, ankle, elbow, and knee displacements over four-frame intervals, measured in token-cell diagonals and clipped at 4. $D^{\text{raw}}_{\text{orient}}$ averages the change in unit bone directions between adjacent frames. $D^{\text{raw}}_{\text{topo}}$ equally combines the fraction of intersecting non-adjacent bone pairs and the mean proximity $\exp(-d^2/(2\cdot 0.2^2))$ of joint pairs without a direct bone connection, where d is their body-normalized distance. $D^{\text{raw}}_{\text{local}}$ averages $(1 - H_t)\bar{v}_t$ over adjacent-frame pairs, where $\bar{v}_t$ is mean joint displacement and H_t is the entropy of the joint-displacement distribution normalized by the logarithm of the valid joint count.

Each raw component is independently rank-normalized over the $n = 100$ test videos as $(\text{rank} - 1)/(n - 1)$, using average ranks for ties. The resulting components define

$$\begin{aligned} R_{\text{struct}} &= 0.35D_{\text{cell}} + 0.25D_{\text{orient}} + 0.25D_{\text{topo}} + 0.15D_{\text{local}}, \\ \text{MSDS} &= 0.15A + 0.85\sqrt{AR_{\text{struct}}}. \end{aligned} \tag{10}$$

Higher scores indicate larger motion coupled with greater structural difficulty. Videos are sorted by MSDS and divided into four equal-sized groups, whose membership is shared by all compared methods.

The 100 test videos are partitioned into four fixed MSDS bins of 25 videos. Score ranges are $[0.0000, 0.2779]$, $[0.2847, 0.5003]$, $[0.5051, 0.6540]$, and $[0.6702, 0.9281]$. UniAnimate-DiT 5B obtains mean per-video FVD values of 848.34, 958.21, 992.68, and 1037.41, respectively, yielding the monotonic trend used in Figure 1.

B.4 Optimization and Sampling

Training uses the flow-matching objective

$$\mathcal{L}_{\text{FM}} = \mathbb{E}\Big[w(\sigma)\left\|V_\theta(Z_\sigma, \sigma, c, I_{\text{ref}}, P^{\text{rgb}}) - (\epsilon - Z_0)\right\|_2^2\Big], \tag{11}$$

Table 5: Default allocation and sampling configuration.

Configuration	Value
Selected-parent ratio	2%
Added fine-token budget	8%
Maximum selected parents/frame	12
Recovery candidate ratio	12.5%
Training dense-path threshold	$\sigma \leq 0.56$
Inference dense tail	10/50 steps

where V_θ denotes the restored velocity on the canonical grid. The loss is evaluated in float32 after canonical-grid restoration. For the 5B configuration, an index i is sampled uniformly from a 1000-entry training grid with

$$u_i = 1 - i/1000, \qquad \sigma_i = \frac{5u_i}{1 + 4u_i}, \qquad t_i = 1000\sigma_i. \tag{12}$$

The scheduler weight is

$$y_i = \exp\left[-2\left(\frac{t_i - 500}{1000}\right)^2\right], \qquad w_i = \frac{1000(y_i - \min_j y_j)}{\sum_j (y_j - \min_k y_k)}. \tag{13}$$

Thus the grid index, rather than the shifted noise level σ, is sampled uniformly. The first latent frame is excluded when the clean TI2V anchor is enabled. The 5B backbone is Wan2.2-TI2V-5B with width 3072 and 30 DiT blocks; its VAE compresses space by 16 and time by 4. Motion LoRA targets the attention query/key/value/output projections and both feed-forward linear layers.

The 5B Stage-2 model starts from a two-epoch pose/motion-LoRA checkpoint and is trained for two additional epochs on four A100 PCIe 40GB GPUs (global batch size 4), using AdamW (Loshchilov & Hutter, 2017), bf16 mixed precision, DeepSpeed ZeRO-2, and gradient checkpointing. LoRA rank and scaling are $r = \alpha = 64$. Learning rates are 5×10^{-5} for the independently trainable enrichment PatchEmbed kernel and 2×10^{-5} for both motion LoRA and the pose module. The original DiT weights, baseline PatchEmbed, and output head remain frozen.

14B configuration. The 14B model uses three Stage-1 epochs followed by two Stage-2 epochs. It shares the Stage-2 learning rates, LoRA rank and scaling factor $r = \alpha = 64$, and inference CFG scale 1.0 with the 5B model. Other training and sampling settings follow the 5B configuration; the 14B experiments run on GPUs with larger memory capacity.

Sampling schedule. Inference uses 50-step FlowMatch sampling, shift 5, CFG 1.0, and seed 0. With $u_j = 1 - j/50$, the model-call noise levels are $\sigma_j = 5u_j/(1 + 4u_j)$ for $j = 0, \ldots, 49$. For a dense-tail length L, compression is active for the first $50 - L$ model calls and disabled thereafter, while motion-adaptive fine representation remains active throughout. Tail-0, Tail-5, and Tail-10 therefore contain exactly 50, 45, and 40 compressed calls. The last model call has $\sigma_{49} = 5/54$; the scheduler then updates to zero without another model evaluation. Compressed-region restoration rejects nonpositive noise levels.

Training separately disables compression when $\sigma \leq 0.56$; this threshold does not override the inference tail. Packed steps use $N + 4K_j - 3G_j$ tokens, whereas dense-refinement steps use $N + 4K_j$ with $G_j = 0$. The default setting uses Tail-10.

B.5 Computational Profiling

Main-DiT FLOPs count attention, cross-attention, and feed-forward matrix operations with one MAC equal to two FLOPs. They exclude text encoding, the VAE, allocation selection, and patch embedding. The reported 15.678 PFLOPs, 148.30 seconds, and 34.28 GiB for MATA 5B are means over the 100 test videos on a single A100 PCIe 40GB GPU using 50 sampling steps. Latency includes text and input encoding, denoising, and VAE decoding, but excludes video reading, comparison-grid construction, and MP4 encoding. Memory is measured with

Table 6: Main comparisons under per-video and pooled global FVD. Lower is better.

Method	FVD ↓	Global FVD ↓
UniAnimate-DiT 5B	959.16	151.26
HyperMotion 14B	808.92	119.82
SCAIL-2 14B	1076.00	154.32
MATA 5B	882.73	135.93
MATA 14B	733.81	131.48

Table 7: Complete metrics for the diagnostic comparisons summarized in the main paper.

Variant	PSNR↑	SSIM↑	L1↓	LPIPS↓	FID↓	VFID↓	FVD↓	$\mathrm{PCK}_{\mathrm{bbox}}$@0.05(%)↑
Motion-side allocation								
Random regions	22.24	0.75	0.0450	0.169	82.93	13.43	881.50	65.38
Motion-guided	22.27	0.75	0.0448	0.169	82.74	12.81	859.95	65.20
Recovery-side allocation								
No reference ranking	22.06	0.74	0.0464	0.182	87.02	12.52	882.75	65.86
Reference ranking	22.20	0.75	0.0456	0.176	85.41	12.57	882.73	65.89
Restoration domain								
Velocity domain	22.02	0.74	0.0466	0.196	100.47	17.61	1066.42	65.87
x_0 consistent	22.20	0.75	0.0456	0.176	85.41	12.57	882.73	65.89
Dense-refinement schedule								
Tail-0	22.19	0.74	0.0457	0.180	87.20	12.87	890.76	65.77
Tail-5	22.20	0.75	0.0456	0.177	86.02	12.53	883.71	65.94
Tail-10	22.20	0.75	0.0456	0.176	85.41	12.57	882.73	65.89

`max_memory_allocated`. The dense baseline requires 15.478 analytically estimated main-DiT PFLOPs.

B.6 Per-Video and Global FVD

MATA 5B improves over UniAnimate-DiT 5B under both aggregations: global FVD decreases from 151.26 to 135.93 (10.14%). Thus, this comparison is not specific to per-video averaging. At 14B, MATA has lower mean per-video FVD than HyperMotion but higher global FVD (131.48 versus 119.82), while improving over SCAIL-2 under both protocols. The two aggregations emphasize different statistical scopes—matched-video feature distributions and the pooled test-set distribution—and are therefore reported separately.

C Supplementary Experimental Results

C.1 Complete Diagnostic Metrics

The matched-token random control isolates placement from added capacity: motion guidance improves FVD and VFID while pose accuracy remains close. Reference ranking mainly protects reconstruction quality. In contrast, x_0-consistent restoration produces a large visual-quality improvement with almost unchanged pose accuracy, confirming that it addresses reconstruction consistency rather than motion control.

C.2 Allocation-Budget Sensitivity

We vary the maximum added fine-token ratio from 4% to 8% and 12%. These are caps before compensating compression, not net sequence-length increases. The corresponding operating points are summarized in Table 8.

Table 8: Allocation-budget configurations. Added fine tokens are expressed as a percentage of the original parent-token count before compensating compression. All entries are caps rather than realized token counts.

Budget	Added fine tokens	Selected parents/frame	Recovery candidate ratio
Low	4%	6	5%
Default	8%	12	12.5%
High	12%	18	20%

Table 9: Allocation-budget sensitivity. The 8% setting is the default.

Budget	PSNR↑	SSIM↑	L1↓	LPIPS↓	FID↓	VFID↓	FVD↓	$\mathrm{PCK}_{\mathrm{bbox}}$@0.05(%)↑
4%	22.28	0.75	0.0453	0.173	84.57	12.67	865.95	65.85
8%	22.20	0.75	0.0456	0.176	85.41	12.57	882.73	65.89
12%	22.29	0.75	0.0450	0.174	85.38	12.94	879.63	66.18

Table 9 reports all eight quality metrics. The 4% budget improves PSNR, L1, LPIPS, FID, and FVD over the default 8% budget, whereas 8% gives lower VFID and slightly higher pose accuracy. Increasing the budget to 12% produces mixed changes as well. Thus, increasing allocation strength does not monotonically improve generation quality. Because fine representation and budget recovery vary jointly under the packed-token constraint, this experiment characterizes the operating-point sensitivity of the complete redistribution rather than a pure token-count effect. The 8% configuration is used as the default operating point in the main experiments.

C.3 Difficulty-Resolved Diagnostics

We first resolve the matched-token random versus motion-guided comparison within the four fixed MSDS bins.

Motion-guided placement lowers mean FVD in all four bins, while the gain magnitude is non-monotonic across MSDS.

Table 11 instead compares the complete method with enrichment alone, isolating the quality cost of recovering token budget.

Budget recovery preserves most of the pose gain while introducing an FVD cost in every bin, separating its budget-control role from the quality-oriented enrichment stage.

C.4 Dense-Refinement Schedule

Table 7 compares the inference tails. Tail-0, Tail-5, and Tail-10 execute 50, 45, and 40 compressed steps, respectively, while retaining motion-adaptive fine representation throughout. Increasing the refinement tail lowers LPIPS from 0.180 to 0.177 and 0.176, and FVD from 890.76 to 883.71 and 882.73. With compression active, the clean prediction in each compressed region is constrained to the prolongation of a reduced-grid estimate (Equation 2). Disabling compression restores individual parent predictions, allowing subsequent denoising to refine spatial variation beyond that reduced representation. The LPIPS and FVD improvements therefore support releasing the compression constraint near the clean endpoint. Because motion-adaptive fine representation remains active in all three settings, this comparison specifically varies the duration of compression.

Tail-5 captures most of the FVD improvement, while Tail-10 further improves LPIPS and FID. Tail-10 is selected as the default operating point because it emphasizes late-stage perceptual refinement, whereas Tail-5 yields the best VFID and pose accuracy. The longer tail also uses more tokens during additional dense steps: its sequence length is $N + 4K_j$, compared with $N + 4K_j - 3G_j$ during budget-constrained processing. The reported full-trajectory FLOPs include this refinement cost.

Table 10: Mean per-video FVD for random and motion-guided fine-token placement within each MSDS bin. Both variants use the same checkpoint and the same number of added tokens. Lower is better.

MSDS bin	Random	Motion-guided	Relative reduction
Low	819.62	767.70	6.33%
Medium	880.93	874.89	0.69%
High	929.64	908.69	2.25%
Extreme	895.83	888.53	0.82%

Table 11: Effect of budget recovery relative to enrichment alone across MSDS bins. Positive values favor the complete method.

Bin	Pose gain (pp)	FVD reduction	FVD reduction (%)
Low	+2.15	-4.21	-0.55
Medium	-0.13	-46.44	-5.31
High	-0.17	-18.93	-2.08
Extreme	+0.92	-21.55	-2.43

D Additional Qualitative Comparisons

Figures 6–8 provide additional comparisons on challenging test sequences. All MATA results use the 14B model.

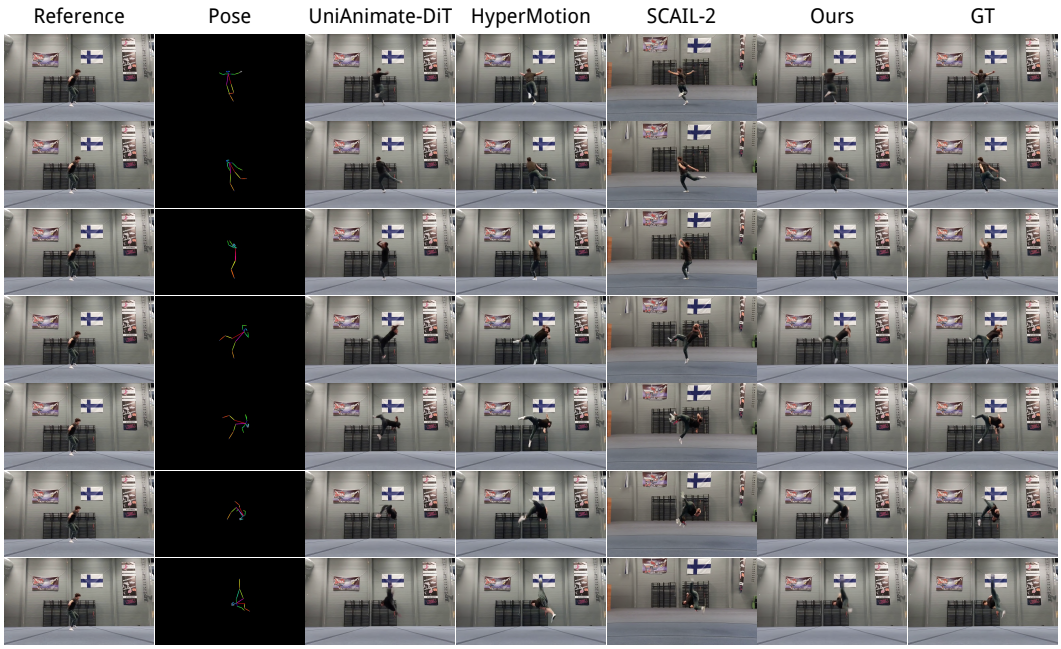

Figure 6: Additional qualitative comparison on test sequence 0021. MATA uses the 14B model.

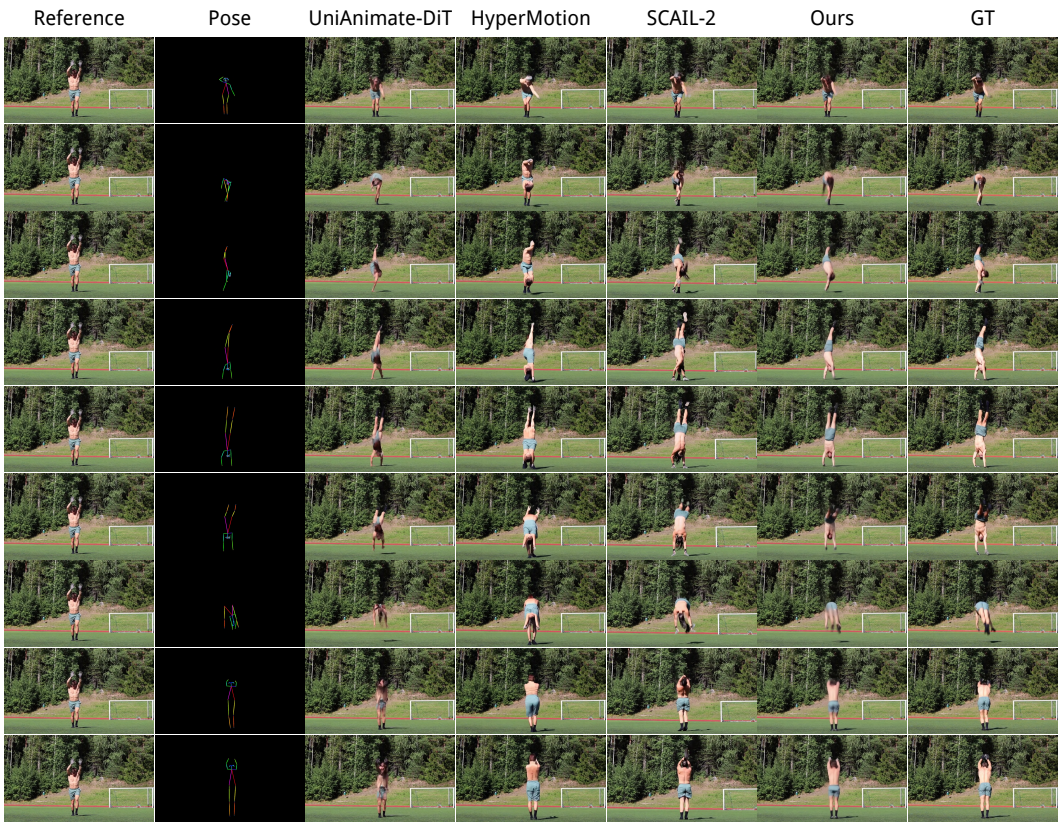

Figure 7: Additional qualitative comparison on test sequence 0037. MATA uses the 14B model.

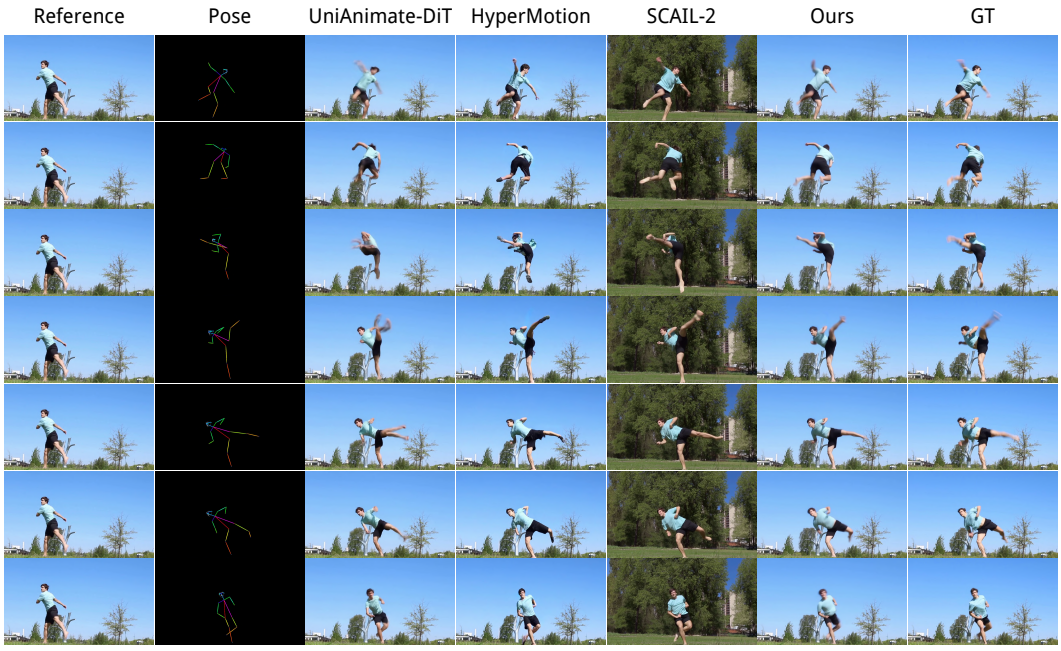

Figure 8: Additional qualitative comparison on test sequence 0059. MATA uses the 14B model.