DynaMo-Diff: Spatio-Temporal Dual-Guided Latent Diffusion for Consistent Stochastic Human Motion Prediction

1st Xiao Yu
Shandong Key Laboratory of Ubiquitous Intelligent Computing
University of Jinan, China
2734808445@qq.com

2nd Na Lyu*
Shandong Key Laboratory of Ubiquitous Intelligent Computing
University of Jinan, China
ise_lvn@ujn.edu.cn

***Abstract*—Stochastic human motion prediction (SHMP) aims to generate multiple plausible future motions from a short observation sequence. Latent diffusion models (LDMs) have recently become a competitive paradigm by performing denoising in a compact motion manifold, offering a favorable fidelity–efficiency trade-off. However, denoising purely in a compressed latent domain presents a persistent challenge in preserving articulated skeletal structures and stabilizing long-horizon trajectories, which can lead to spatio-temporal inconsistencies such as topology violations and cumulative drift. This paper proposes DynaMo-Diff, a spatio-temporal dual-guided latent diffusion framework that integrates task-aligned guidance into the denoising process. DynaMo-Diff incorporates two complementary interfaces: (i) Dynamic Graph Attention (DGA) for local spatial reasoning, which employs Top-k sparsification to adaptively model motion-dependent inter-joint dependencies; and (ii) a Trajectory Anchor Predictor (TAP) for global temporal guidance, which forecasts sparse pose-space waypoints to mitigate long-horizon error accumulation. To ensure manifold stability and computational efficiency, we adopt a decoupled two-stage training strategy with a fixed motion encoder. Extensive experiments on Human3.6M and AMASS demonstrate competitive performance and improved long-horizon stability under an efficient sampling budget. Notably, DynaMo-Diff achieves the best FDE of 0.452 on Human3.6M in our comparison while maintaining robust performance on the challenging AMASS benchmark.**

***Index Terms*—Human Motion Prediction, Latent Diffusion Models, Graph Attention Networks, Spatio-Temporal Guidance, Consistent Prediction.**

I. Introduction

Predicting future human motion is a foundational computer vision task with critical applications in human-robot collaboration and autonomous driving [4]. Given the inherent ambiguity of human behavior, research has shifted from deterministic regression to stochastic human motion prediction (SHMP). Recently, Latent Diffusion Models (LDMs) [9], [25] have emerged as a powerful paradigm, effectively balancing generative quality and efficiency by modeling motion distributions within a compact, pre-learned latent manifold.

However, denoising purely within an abstract latent space remains challenging. Since latent features are highly compressed, explicitly maintaining fine-grained skeletal topologies and long-horizon trajectory consistency is non-trivial. Standard architectures often fail to encapsulate articulated constraints [11], and even with implicit consistency scores [20] or discrete quantization [29], the lack of an *explicit* global scaffold leads to cumulative trajectory drift over extended horizons—a critical issue for stability and complex joint interactions.

To address these limitations, we propose DynaMo-Diff, a framework that introduces a Spatio-Temporal Dual-Guidance mechanism to ground the latent diffusion process with explicit structural and temporal cues. To ensure manifold stability and avoid mode collapse, we adopt a decoupled two-stage strategy with a fixed pre-trained VAE. This design allows the diffusion model to focus on learning complex temporal dynamics while receiving pose-space guidance to mitigate information loss during latent denoising.

Our framework incorporates two synergistic components: (i) a Dynamic Graph Attention (DGA) module that adaptively reconstructs joint-wise topologies within the denoiser. Unlike the fixed skeletal priors in quantized frameworks [29], DGA captures evolving dependencies for each stochastic sample to maintain kinematic integrity. (ii) a Trajectory Anchor Predictor (TAP) that forecasts sparse pose-space waypoints. In contrast to implicit temporal scores [20], TAP provides explicit scaffolds to stabilize the stochastic distribution and mitigate error accumulation without compromising diversity. By bridging implicit latent denoising and explicit pose-space constraints, DynaMo-Diff achieves a more consistent predictive distribution.

The main contributions of this work are summarized as follows:

- We propose a dual-guided latent diffusion framework that enhances motion consistency by integrating dynamic graph reasoning and explicit trajectory anchoring.
- We design a Dynamic Graph Attention (DGA) module that adaptively models inter-joint dependencies, helping to preserve structural precision across diverse actions compared to static priors.

*Corresponding author.

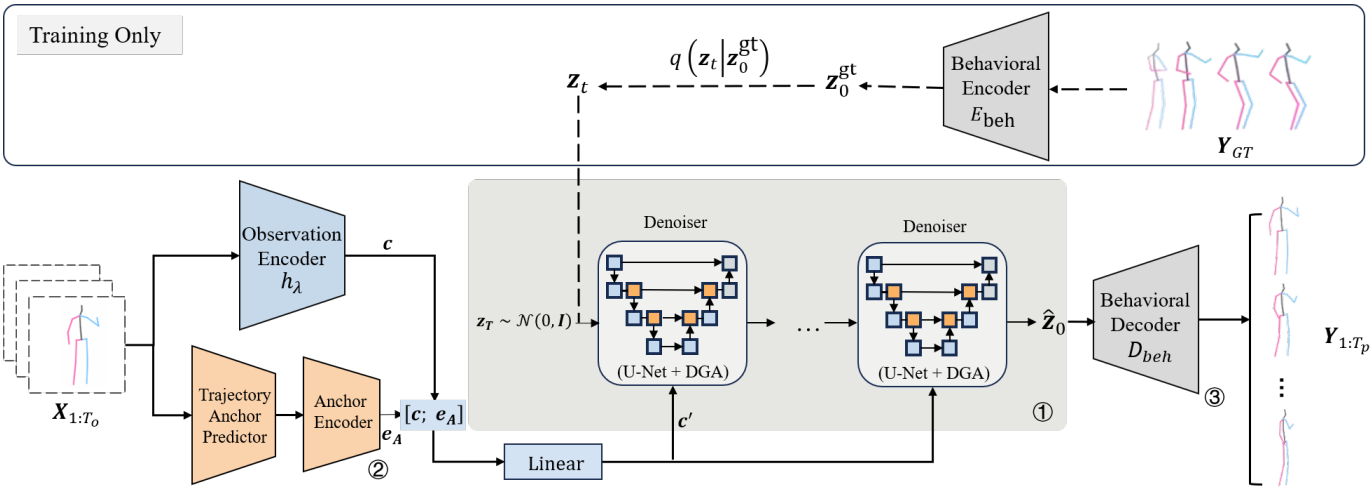

Fig. 1. Overview of DynaMo-Diff. Given prefix $\mathbf{X}_{1:T_o}$, observation context $\mathbf{c}$ and TAP-predicted anchors $\mathbf{e}_A$ are fused into enhanced conditioning $\mathbf{c}'$ to guide the U-Net+DGA denoiser. Training (dashed): Ground-truth $\mathbf{Y}_{\text{GT}}$ is mapped to latent $\mathbf{z}_0^{\text{gt}}$ for denoising, optimized via a hybrid objective ($\mathcal{L}_{diff}$, $\mathcal{L}_{tap}$, and $\mathcal{L}_{geo}$). Inference (solid): Starting from Gaussian noise $\mathbf{z}_T$, the predicted latent $\hat{\mathbf{z}}_0$ is decoded by D_{beh} to yield final motion $\mathbf{Y}_{1:T_p}$. Behavioral modules (E_{beh}, D_{beh}) are pretrained and frozen to ensure manifold stability.

- We introduce a Trajectory Anchor Predictor (TAP) mechanism that provides global temporal guidance via sparse scaffolds, improving long-horizon stability while maintaining distributional diversity.
- Extensive evaluations on Human3.6M and AMASS benchmarks demonstrate that DynaMo-Diff achieves competitive performance and stronger long-horizon stability, particularly in reducing FDE while preserving a natural generative distribution.

II. Related Work

The technical evolution of human motion prediction has transitioned from deterministic modeling to capturing the inherent stochasticity of human behavior. This section reviews the paradigms that ground our proposed framework.

A. Stochastic Human Motion Prediction

Early methods primarily utilized deterministic regression, such as RNN-based [18] or GCN-based [12] backbones, which often suffer from over-smoothing in long-horizon scenarios. To capture multi-modality, generative paradigms based on VAEs [8], [16] and GANs [14] were introduced. For instance, DLow [19] leveraged normalizing flows to diversify outputs, while more recent probabilistic forecasting models explicitly aimed to capture multimodal motion distributions [21]. However, these frameworks often face a tension between sample diversity and kinematic fidelity. DynaMo-Diff leverages the superior distribution modeling of diffusion models to achieve a more robust balance.

B. Diffusion Models for Motion Generation

Diffusion models [23], [24] have demonstrated strong generative capability across diverse domains. Building on this, their application to human motion synthesis [28] and 3D motion prediction [6] has been actively investigated. To improve efficiency, *latent diffusion models* (LDMs) [25] perform denoising in a compact manifold, enabling scalable generation. BeLFusion [9] and HumanMAC [17] established robust baselines for efficient sampling. Recently, CoMusion [20] explored implicit score-based temporal consistency, and QCDM [29] utilized discrete quantization. However, a common bottleneck remains: latent denoising lacks explicit structural grounding. DynaMo-Diff differentiates itself in two aspects: First, instead of the implicit consistency in [20], we provide *explicit* pose-space anchors as scaffolds to stabilize the *stochastic distribution*. Second, unlike the fixed graphs or quantized tokens in [29], our DGA module dynamically prunes joint-wise correlations for each generated sample.

C. Structural and Temporal Modeling

Modeling the body as a graph is standard for capturing joint dependencies [11], [30]. However, most methods rely on fixed topologies that cannot adapt to transient, motion-dependent interactions. In parallel, although sparse waypoints have been explored for temporal scaffolding [13], they are typically designed for deterministic rollouts. In contrast, our framework integrates DGA and TAP directly into the latent diffusion process. This dual-guidance ensures that the *multi-modal predictive distribution* remains grounded by both local anatomical constraints and global temporal intent, preventing the cumulative drift common in unguided latent denoising.

III. METHODOLOGY

A. Framework Overview

DynaMo-Diff follows a decoupled two-stage paradigm (Fig. 1) to integrate latent diffusion with explicit structural and temporal guidance. Given a motion prefix $\mathbf{X} \in \mathbb{R}^{T_o \times J \times C}$, we aim to predict a distribution of plausible futures $\mathbf{Y} \in$

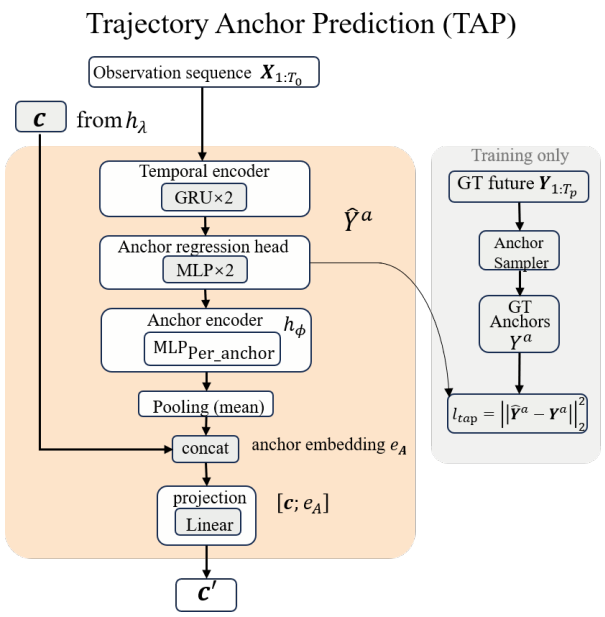

Fig. 2. Trajectory Anchor Prediction (TAP). TAP predicts sparse pose-space anchors $\widehat{\mathbf{Y}}^a$ from observation history $\mathbf{X}_{1:T_o}$. These anchors are encoded into an anchor embedding $\mathbf{e}_A$ and then fused with the observation conditioning $\mathbf{c}$ to obtain the enhanced conditioning vector $\mathbf{c}'$, which provides global temporal guidance to the diffusion process.

$\mathbb{R}^{T_p \times J \times C}$ by modeling the conditional score function within a pre-structured latent manifold $\mathcal{Z}$.

In Stage I, we employ a pre-trained motion VAE to establish a compact latent space $\mathcal{Z}$. To ensure manifold stability and avoid the "moving target" problem, the behavioral encoder E_{beh} and decoder D_{beh} are frozen during diffusion training. In Stage II, a structure-aware denoiser $\boldsymbol{\epsilon}_\theta$ operates within $\mathcal{Z}$ using a Dual-Guidance Mechanism: (i) *Global Temporal Guidance (TAP)* forecasts sparse waypoints from $\mathbf{X}$ to provide a temporal scaffold against long-horizon drift; and (ii) *Local Spatial Reasoning (DGA)* adaptively reconstructs joint topologies to maintain kinematic integrity.

B. Trajectory Anchor Prediction (TAP)

TAP provides explicit pose-space scaffolds to anchor the multimodal distribution. It forecasts N_a sparse anchors $\widehat{\mathbf{Y}}^a$ directly from history $\mathbf{X}_{1:T_o}$, which are encoded into an embedding $\mathbf{e}_A$ via an MLP and temporal mean-pooling. We use Feature-wise Linear Modulation (FiLM) [5] to integrate guidance: $\mathbf{e}_A$ and context $\mathbf{c}$ are fused into $\mathbf{c}' = \text{Linear}([\mathbf{c}; \mathbf{e}_A])$. For a feature map $\mathbf{F}$, the modulation is:

$$\text{FiLM}(\mathbf{F}; \mathbf{u}) = \gamma(\mathbf{u}) \odot \text{Norm}(\mathbf{F}) + \beta(\mathbf{u}), \qquad (1)$$

where $\mathbf{u} = [\text{Emb}(t); \mathbf{c}']$. This "soft guidance" ensures diffusion remains grounded by the temporal scaffold while preserving natural stochastic fluidity. The overall TAP architecture is illustrated in Fig. 2.

C. Dynamic Graph Attention (DGA)

DGA reconstructs articulated topology from latent features. We first project U-Net features into joint-aligned tokens $\mathbf{H} \in \mathbb{R}^{N \times d}$. To capture time-evolving interactions, we compute affinity scores $S \in \mathbb{R}^{N \times N}$ via a query-key attention mechanism inspired by scaled dot-product attention [31], with a context bias:

$$\mathbf{S}_{ij} = \frac{(\mathbf{H}_i \mathbf{W}_Q)(\mathbf{H}_j \mathbf{W}_K)^\top}{\sqrt{d_k}} + \text{MLP}_{bias}([\mathbf{c}' \parallel \text{Emb}(t)]), \quad (2)$$

where MLP_{bias} predicts a context-dependent bias matrix B. Top-k filtering and masked softmax derive the dynamic adjacency $\mathbf{A}^k_{dyn}$. The module then aggregates features via dual-stream fusion: a static branch using fixed kinematics $\mathbf{A}_{static}$ (producing $\mathbf{H}_s$) and a dynamic branch using $\mathbf{A}^k_{dyn}$ (producing $\mathbf{H}_d$). A learnable gate $\boldsymbol{\alpha}$ fuses these streams: $\tilde{\mathbf{H}} = \boldsymbol{\alpha} \odot \mathbf{H}_d + (1 - \boldsymbol{\alpha}) \odot \mathbf{H}_s$. Final output $\mathbf{H}_{out} = \text{Linear}(\mathbf{H} + \tilde{\mathbf{H}})$ allows shifting between anatomical constraints and data-driven interactions. The DGA module architecture is detailed in Fig. 3.

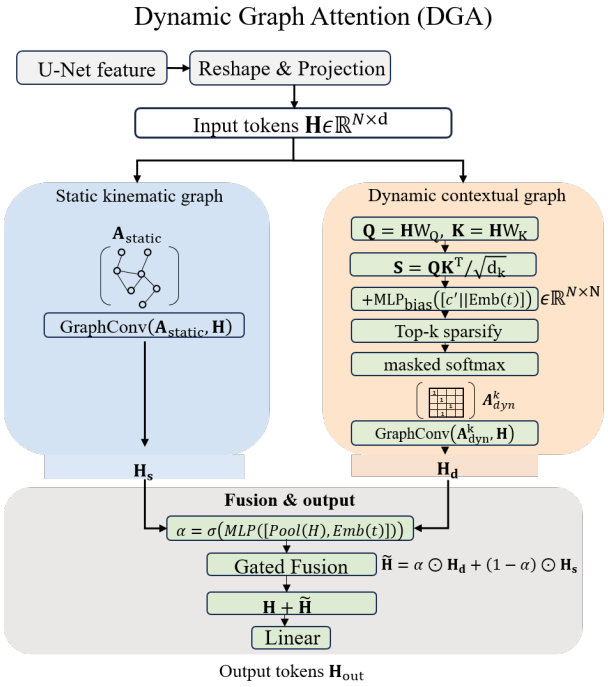

Fig. 3. Architecture of the Dynamic Graph Attention (DGA) module. DGA operates on joint tokens projected from U-Net features, $\mathbf{H} \in \mathbb{R}^{N \times d}$. A static kinematic adjacency $\mathbf{A}_{static}$ yields a structural representation $\mathbf{H}_s = \text{GraphConv}(\mathbf{A}_{static}, \mathbf{H})$. In parallel, DGA constructs a dynamic contextual graph by computing affinities $\mathbf{S}$ from token interactions and adding a context-dependent bias $\mathbf{B} \in \mathbb{R}^{N \times N}$. Top-$k$ sparsification followed by a masked softmax produces a sparse, normalized adjacency $\mathbf{A}^k_{dyn}$, which is used to obtain $\mathbf{H}_d = \text{GraphConv}(\mathbf{A}^k_{dyn}, \mathbf{H})$. A gating vector $\boldsymbol{\alpha}$ is predicted from pooled features and broadcast to fuse the two streams, $\tilde{\mathbf{H}} = \boldsymbol{\alpha} \odot \mathbf{H}_d + (1 - \boldsymbol{\alpha}) \odot \mathbf{H}_s$, and a residual projection outputs the final tokens $\mathbf{H}_{out}$.

D. Training Objective

We optimize the denoiser and guidance modules while keeping the VAE fixed. The total loss is $\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_1 \mathcal{L}_{\text{tap}} + \lambda_2 \mathcal{L}_{\text{geo}}$. $\mathcal{L}_{\text{diff}}$ is the score matching loss: $\mathbb{E}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c}, \mathbf{e}_A)\|^2$. $\mathcal{L}_{\text{tap}}$ minimizes anchor regression error: $\frac{1}{N_a J} \sum \|\widehat{\mathbf{Y}}^a_i - \mathbf{Y}^a_i\|^2$. To

TABLE I

QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON HUMAN3.6M [2] AND AMASS [3]. BEST RESULTS ARE IN **BOLD** AND SECOND-BEST ARE UNDERLINED. "-" INDICATES UNAVAILABLE OR UNREPORTED RESULTS, AND * FOLLOWS THE PRIOR EVALUATION PROTOCOL.

	Human3.6M							AMASS				
Method	**APD↑**	**ADE↓**	**FDE↓**	**MMADE↓**	**MMFDE↓**	**CMD↓**	**FID*↓**	**APD↑**	**ADE↓**	**FDE↓**	**MMADE↓**	**MMFDE↓**
HP-GAN [14]	7.214	0.858	0.867	0.847	0.858	-	-	-	-	-	-	-
DeLiGAN [15]	6.509	0.483	0.534	0.520	0.545	-	-	-	-	-	-	-
DLow [19]	11.741	0.425	0.518	0.495	0.531	4.93	1.25	13.170	0.590	0.612	0.618	0.617
GSPS [22]	14.757	0.389	0.496	0.476	0.525	10.76	2.10	12.465	0.563	0.613	0.609	0.633
DivSamp [7]	**15.310**	0.370	0.485	0.475	0.516	11.69	2.08	**24.724**	0.564	0.647	0.623	0.667
HumanMAC [17]	6.301	0.369	0.480	0.509	0.545	-	-	9.321	0.511	0.554	0.593	0.591
BeLFusion [9]	7.602	0.372	0.474	0.473	0.507	5.99	0.21	9.376	0.513	0.560	0.569	0.585
CoMusion [20]	7.632	**0.350**	0.458	0.494	0.506	**3.20**	**0.10**	10.848	0.494	0.547	**0.469**	**0.466**
QCDM [29]	7.820	0.353	0.470	0.484	0.506	-	-	-	-	-	-	-
Ours	7.680	0.358	**0.452**	**0.468**	**0.498**	5.85	0.19	9.650	**0.488**	**0.526**	0.551	0.568

mitigate artifacts like foot sliding, $\mathcal{L}_{\text{geo}}$ penalizes reconstruction and joint velocity in pose space:

$$\mathcal{L}_{\text{geo}} = \frac{1}{T}\sum_{t=1}^{T}(\|\widehat{\mathbf{Y}}_t - \mathbf{Y}_t\|_2^2 + \lambda_{vel}\|\mathbf{v}_t - \hat{\mathbf{v}}_t\|_2^2), \quad (3)$$

where $\widehat{\mathbf{Y}} = D_{\text{beh}}(\hat{\mathbf{z}}_0)$. We use the Straight-Through Estimator (STE) [1] for the non-differentiable Top-k selection.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

We evaluate DynaMo-Diff on two benchmarks using 0.5s observations to predict 2.0s of motion.

a) Datasets: For *Human3.6M* [2], we follow [9] using a 16-joint skeleton at 25 fps, training on subjects S1, S5–S8 and testing on S9, S11. A 15-joint protocol (excluding root) is adopted for deterministic evaluations. *AMASS* [3] serves as a cross-dataset testbed, training on 11 datasets and testing on 7 unseen ones (e.g., DanceDB, GRAB) featuring a 21-joint configuration.

b) Evaluation Metrics: We assess accuracy via ADE and FDE ($K = 50$), diversity via APD, and distributional quality via FID on Human3.6M. For AMASS, we report ADE/FDE together with multi-modal metrics (MMADE/MMFDE) to account for domain ambiguity.

B. Implementation Details

a) Model Configuration: DynaMo-Diff adopts a decoupled strategy: a behavioral VAE establishes a 128-dimensional latent manifold $\mathcal{Z}$, with E_{beh} and D_{beh} frozen during Stage II. The denoiser is an 8-layer conditional U-Net (hidden dim 512) integrating DGA modules, while TAP utilizes a dual-layer GRU and an MLP-based regression head.

b) Conditioning and FiLM Rationale: Anchor embeddings $\mathbf{e}_A \in \mathbb{R}^{128}$ are fused with context $\mathbf{c}$ into $\mathbf{c}'$ to modulate the U-Net via Feature-wise Linear Modulation (FiLM). Unlike rigid coordinate constraints that suppress stochastic fluidity, FiLM provides a "soft-guidance" mechanism. This allows the model to interpret anchors as adaptive scaffolds, effectively grounding long-horizon trajectories without compromising high-frequency motion details.

c) Training Protocol: We optimize Stage II via Adam (LR 5×10^{-4}, EMA 0.999) using a square-root noise schedule ($T = 10$) and loss weights $\lambda_{tap} = \lambda_{geo} = 1.0$. Training requires ∼12 hours on an NVIDIA RTX 3090 for Human3.6M.

TABLE II

EVALUATION OF DETERMINISTIC PREDICTION ACCURACY (MPJPE IN MM) ON HUMAN3.6M USING THE 15-JOINT PROTOCOL. DYNAMO-DIFF DEMONSTRATES STRONGER STABILITY IN LONG-HORIZON FORECASTING (320–1000 MS).

Time (ms)	80	160	320	400	560	1000
LTD [12]	12.1	25.0	51.0	61.3	78.5	114.3
MSR [10]	12.1	25.6	51.6	62.9	81.1	114.2
ST-DGCN [11]	**10.3**	**22.7**	47.4	58.5	76.9	110.3
QCDM [29]	10.4	23.5	47.3	56.4	75.2	105.1
Ours	10.4	23.2	**47.1**	**56.2**	**75.0**	**104.5**

C. Comparison with State-of-the-Art Methods

a) Stochastic Motion Prediction: As shown in Table I, DynaMo-Diff achieves competitive overall performance on both benchmarks. On Human3.6M, we achieve the best FDE of 0.452, outperforming CoMusion (0.458), while obtaining an FID of 0.19. This improvement stems from our dual-guidance mechanism, which grounds latent denoising in explicit temporal intent and skeletal topology, thereby improving long-horizon stability (see Fig. 4). Competitive metrics on AMASS (ADE 0.488, FDE 0.526) further demonstrate robust generalization to unseen motion distributions.

b) Deterministic Prediction Analysis: We evaluate accuracy across horizons against deterministic baselines (LTD [12], MSR [10], ST-DGCN [11], and QCDM [29]) using the 15-joint protocol (Table II). While GCN-based models like ST-DGCN excel at short intervals (≤80ms), the advantage of DynaMo-Diff becomes more pronounced over longer horizons, reaching the best MPJPE of 104.5 mm at 1000 ms. This suggests that TAP-generated scaffolds serve as effective global anchors, helping reduce the "averaging effect" and trajectory collapse typical of long-horizon deterministic forecasting.

TABLE III
ABLATION STUDY OF DYNAMO-DIFF ON HUMAN3.6M AND AMASS. S/D/T: STATIC/DYNAMIC/TAP. ENC: ENCODER. †VARIANT (F) WAS EVALUATED ON HUMAN3.6M ONLY.

	Components				Human3.6M			AMASS	
Variant	S	D	T	Enc	ADE↓	FDE↓	APD↑	FDE↓	MMFDE↓
(a) Baseline	-	-	-	Fixed	0.372	0.474	7.602	0.560	0.585
(b) w/ Static	✓	-	-	Fixed	0.369	0.472	7.615	0.555	0.575
(c) w/ Dynamic	-	✓	-	Fixed	0.365	0.470	7.625	0.551	0.572
(d) w/ DGA	✓	✓	-	Fixed	0.363	0.468	<u>7.640</u>	0.548	0.580
(e) w/ TAP	-	-	✓	Fixed	0.370	<u>0.458</u>	7.635	<u>0.545</u>	<u>0.570</u>
(f) w/ Concat	-	-	✓	Fixed	0.372	0.461	7.520	-	-
(g) Joint Training	✓	✓	✓	Fine-tune	**0.356**	0.455	7.210	0.535	0.585
(h) Full Model	✓	✓	✓	Fixed	<u>0.358</u>	**0.452**	**7.680**	**0.526**	**0.568**

D. Ablation Study

Table III validates our design choices on Human3.6M and AMASS. Dual-Guidance: DGA (Variants b–d) enhances short-horizon precision by capturing dynamic topologies (ADE: 0.372 → 0.363), while TAP (Variant e) anchors long-horizon stability (0.458 FDE). Comparing (e) and (f) suggests that FiLM-based guidance better avoids the motion stiffness induced by rigid concatenation (FDE 0.461) while preserving natural stochasticity. On AMASS, dual-guidance (Variant h) achieves the lowest MMFDE (0.568), highlighting its importance for reducing trajectory divergence in unseen domains.

Decoupled Strategy: Freezing the behavioral manifold is crucial for manifold stability. Jointly training the encoder (Variant g) triggers a catastrophic diversity collapse (APD: 7.680 → 7.210) and hinders generalization (0.585 MMFDE). Our decoupled paradigm allows the diffusion model to explore diverse futures within a stable latent space $\mathcal{Z}$ without being biased toward the training mean.

E. Hyperparameter Sensitivity and Robustness

Sensitivity analysis (averaged over multiple seeds to account for stochastic variance) confirms $N_a = 8$ and $k = 8$ as optimal. Deviations in N_a lead to either trajectory drift (low density) or manifold stiffness (high density), while k balances essential joint dependencies against spurious correlations from kinematically distant joints. Efficient few-step sampling has been widely studied in diffusion modeling [26], [27]. Regarding efficiency, FDE reaches its performance nadir at $T = 10$ (0.452), with marginal gains (< 0.003m) up to $T = 50$ despite a 5× latency increase. Optimal guidance strength $\lambda_{tap} = 1.0$ effectively balances explicit anchoring with the learned temporal smoothness of the motion manifold. Finally, TAP exhibits high noise tolerance on Human3.6M (Table IV, $< 2.5\%$ FDE increase at $\sigma = 0.05$m), confirming that FiLM-based soft-guidance treats anchors as informative scaffolds that rectify estimation inaccuracies while preserving a realistic motion distribution.

F. Trade-off and Efficiency Analysis

We evaluate the computational efficiency of DynaMo-Diff on an NVIDIA RTX 3090 (Table V). Our model provides a favorable trade-off between speed and stability, incurring only a marginal 25ms overhead compared to BeLFusion [9] (345ms vs. 320ms) while achieving a 4.6% average FDE improvement (Fig. 4). While CoMusion [20] is faster due to its simplified architecture, DynaMo-Diff provides stronger skeletal consistency and competitive FID, which are important for high-fidelity human-robot interaction. Notably, compared to HumanMAC [17], our model is 46% more parameter-efficient and over 3.6× faster, thanks to the Top-k sparsification in DGA and the lightweight temporal encoder in TAP. These results suggest that DynaMo-Diff is computationally feasible for latency-sensitive robotics and VR applications.

TABLE IV
ROBUSTNESS OF TAP TO ANCHOR NOISE ON HUMAN3.6M (σ IN METERS). DYNAMO-DIFF REMAINS STABLE UNDER MODERATE ESTIMATION ERRORS.

Noise Level σ	ADE↓	FDE↓	APD↑
0 (Full Model)	0.358	0.452	7.680
0.01	0.360	0.455	7.675
0.05	0.365	0.463	7.662
0.10	0.371	0.475	7.640

TABLE V
COMPARISON OF MODEL PARAMETERS AND INFERENCE RUNTIME ON HUMAN3.6M ($K = 50$ SAMPLES). DYNAMO-DIFF OFFERS A COMPETITIVE SPEED-ACCURACY BALANCE FOR REAL-TIME APPLICATIONS.

Method	Params (M) ↓	Time (ms) ↓
HumanMAC [17]	28.40	1255
BeLFusion [9]	**13.19**	320
CoMusion [20]	18.79	**179**
Ours	<u>15.20</u>	<u>345</u>

G. Discussion and Limitations

DynaMo-Diff demonstrates clear improvements in long-horizon forecasting stability and skeletal consistency. However, two primary limitations remain. First, the reliance on deterministic anchors can be problematic in scenarios with extreme, rapid stochasticity (e.g., sudden directional reversals), where a single scaffold may not capture the full distribution of plausible outcomes. Second, while the fixed encoder ensures manifold stability, it may limit the model's ability to adapt to highly out-of-distribution motion patterns not seen during VAE training. Future research could focus on probabilistic anchor prediction to further enhance distributional coverage in complex environments.

V. CONCLUSION

We propose DynaMo-Diff, a dual-guided latent diffusion framework for consistent stochastic human motion prediction. By integrating Dynamic Graph Attention and a Trajectory Anchor Predictor, we bridge the gap between abstract latent denoising and pose-space plausibility. Extensive evaluations on Human3.6M and AMASS benchmarks show that our model

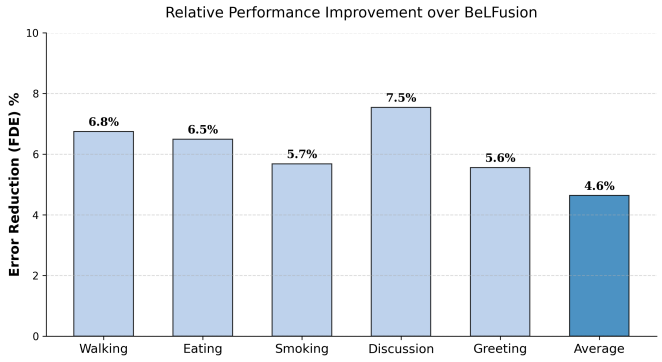

Fig. 4. Summary of Stability Gains. Relative FDE reduction on Human3.6M vs. BeLFusion across typical action categories. The results show that our dual-guidance mechanism consistently stabilizes long-horizon forecasting. The 'Average' bar summarizes the overall improvement across the evaluated Human3.6M action categories, including those not explicitly plotted here.

achieves strong accuracy and high motion diversity. These findings highlight the importance of explicit structural grounding in stabilizing diffusion-based motion synthesis, paving the way for more robust and realistic human-centric AI systems.

REFERENCES

[1] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013.

[2] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36, no. 7, pp. 1325–1339, 2014.

[3] N. Mahmood et al., "AMASS: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5442–5451.

[4] K. Lyu, H. Chen, Z. Liu, B. Zhang, and R. Wang, "3d human motion prediction: A survey," *Neurocomputing*, vol. 489, pp. 345–365, 2022.

[5] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 32, no. 1, 2018, pp. 3942–3951.

[6] H. Ahn, E. V. Mascaro, and D. Lee, "Can we use diffusion probabilistic models for 3d motion prediction?," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9837–9843.

[7] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li, "Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space," in *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 2022, pp. 5162–5171.

[8] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3332–3341.

[9] G. Barquero, S. Escalera, and C. Palmero, "BeLFusion: Latent diffusion for behavior-driven human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 2317–2327.

[10] L. Dang, Y. Nie, C. Long, Q. Zhang, and G. Li, "MSR-GCN: Multi-scale residual graph convolution networks for human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 11467–11476.

[11] T. Ma, Y. Nie, C. Long, Q. Zhang, and G. Li, "Progressively generating better initial guesses towards next stages for high-quality human motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6437–6446.

[12] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9489–9497.

[13] S. Xu, Y.-X. Wang, and L.-Y. Gui, "Diverse human motion prediction guided by multi-level spatial-temporal anchors," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 251–269.

[14] E. Barsoum, J. Kender, and Z. Liu, "Hp-gan: Probabilistic 3d human motion prediction via gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1418–1427.

[15] S. Gurumurthy, R. K. Sarvadevabhatla, and R. V. Babu, "Deligan: Generative adversarial networks for diverse and limited data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 166–174.

[16] X. Bie et al., "Hit-dvae: Human motion generation via hierarchical transformer dynamical vae," *arXiv preprint arXiv:2204.01565*, 2022.

[17] L.-H. Chen et al., "Human-mac: Masked motion completion for human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 9544–9555.

[18] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[19] Y. Yuan and K. Kitani, "DLow: Diversifying latent flows for diverse human motion prediction," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 346–364.

[20] J. Sun and G. Chowdhary, "CoMusion: Towards consistent stochastic human motion prediction via motion diffusion," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, pp. 18–36.

[21] T. Salzmann, M. Pavone, and M. Ryll, "Motron: Multimodal probabilistic human motion forecasting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6457–6466.

[22] W. Mao, M. Liu, and M. Salzmann, "Generating smooth pose sequences for diverse human motion prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13309–13318.

[23] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.

[24] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8162–8171.

[25] R. Rombach et al., "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.

[26] C. Lu, Y. Zhou, F. Bao, J. Chen, and J. Zhu, "DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 5775–5787.

[27] C. Lu et al., "DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models," *Machine Intelligence Research*, vol. 22, no. 4, pp. 730–751, 2025.

[28] M. Zhang et al., "MotionDiffuse: Text-driven human motion generation with diffusion model," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 46, no. 6, pp. 4115–4128, 2024.

[29] B. Huang, X. Li, C. Hu, and H. Li, "Stochastic human motion prediction using a quantized conditional diffusion model," *Knowledge-Based Systems*, vol. 309, p. 112823, 2025.

[30] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 32, no. 1, 2018, pp. 7444–7452.

[31] A. Vaswani et al., "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.