UMO: Unified In-Context Learning Unlocks Motion Foundation Model Priors

Xiaoyan Cong[1*], Zekun Li[1*], Zhiyang Dou[2], Hongyu Li[1], Omid Taheri[4], Chuan Guo[3], Abhay Mittal[3], Sizhe An[3], Taku Komura[5], Wojciech Matusik[2], Michael J. Black[4], and Srinath Sridhar[1†]

[1]Brown University [2]Massachusetts Institute of Technology [3]Meta Reality Lab
[4]Max-Planck Institute for Intelligent Systems [5]University of Hong Kong

Project Page: `https://oliver-cong02.github.io/UMO.github.io`

Abstract. Large-scale foundation models (LFMs) have recently made impressive progress in text-to-motion generation by learning strong generative priors from massive 3D human motion datasets and paired text descriptions. However, how to effectively and efficiently leverage such single-purpose motion LFMs, *i.e.*, text-to-motion synthesis, in more diverse cross-modal and in-context motion generation downstream tasks remains largely unclear. Prior work typically adapts pretrained generative priors to individual downstream tasks in a task-specific manner. In contrast, our goal is to unlock such priors to support a broad spectrum of downstream motion generation tasks within a single unified framework. To bridge this gap, we present **UMO**, a simple yet general unified formulation that casts diverse downstream tasks into compositions of atomic per-frame operations, enabling in-context adaptation to unlock the generative priors of pretrained DiT-based motion LFMs. Specifically, UMO introduces **three learnable frame-level meta-operation embeddings** to specify per-frame intent and employs lightweight temporal fusion to inject in-context cues into the pretrained backbone, with negligible runtime overhead compared to the base model. With this design, UMO finetunes the pretrained model, originally limited to text-to-motion generation, to support diverse previously unsupported tasks, including temporal inpainting, text-guided motion editing, text-serialized geometric constraints, and multi-identity reaction generation. Experiments demonstrate that UMO consistently outperforms task-specific and training-free baselines across a wide range of benchmarks, despite using a single unified model. Code and models will be publicly available.

Keywords: Human Motion Generation · Motion Foundation Model

1 Introduction

3D human motion generation [4, 8, 14, 29, 36, 45, 64, 75] has witnessed substantial progress across diverse tasks, yet the field remains fragmented, with each task

* Equal contribution.
† Corresponding author.

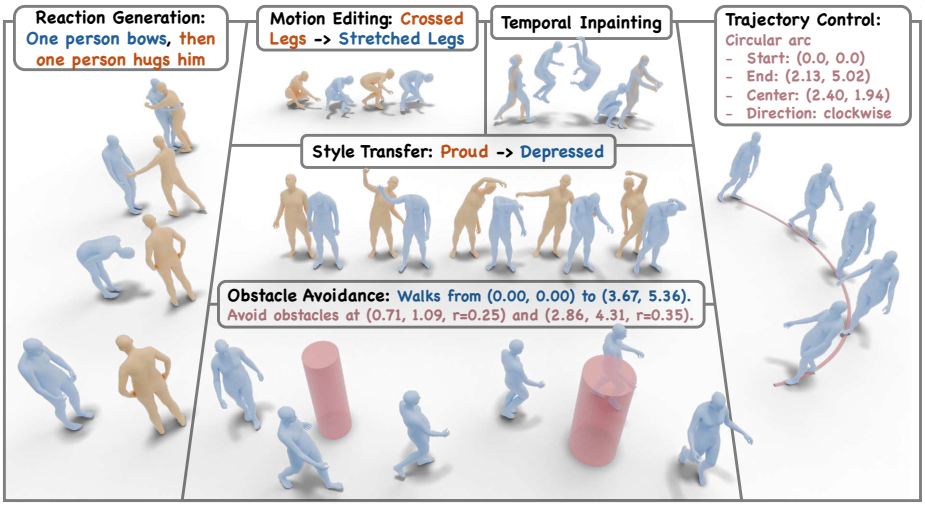

Fig. 1: UMO casts diverse downstream tasks as compositions of our novel atomic per-frame operations, unlocking the generative priors of a pretrained text-to-motion LFM without any architectural modification in a broad spectrum of applications.

typically addressed by a specialized architecture, preventing knowledge sharing and cross-task generalization. In other generative domains [9,20,40,55,57], it has been widely demonstrated that unifying diverse tasks under a shared formulation can effectively unlock powerful generative priors from pretrained large foundation models [11, 27, 28, 53]. However, such a unified framework remains largely absent in the motion domain, despite being critical for improving generalization and consolidating scarce per-task data into a unified training framework.

Recent efforts in building motion LFMs [12,30,38,56] have significantly scaled up both model capacity and training data, learning general generative priors for human motion. However, these models primarily focus on basic text-to-motion (T2M) generation, leaving their rich generative priors untapped for broader applications. This raises a natural question: *how can we develop a unified and efficient framework that unlocks pretrained motion LFM priors to support diverse unseen downstream tasks?*

We introduce **UMO**, a simple yet effective unified framework that unlocks generative priors of pretrained DiT-based motion LFMs for diverse

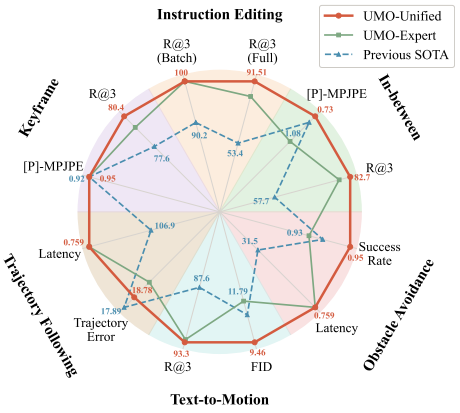

Fig. 2: UMO achieves state-of-the-art performance and efficiency across diverse tasks. Latency, Traj. Err, [P]-MPJPE, and FID are inverted so that the outermost boundary is consistently optimal.

downstream tasks, as summarized in Tab. 2. Unlike prior task-specific solutions [4,7,19,48,49] that require dedicated architectures or conditioning modules for each task, UMO expresses all cross-modal and in-context motion generation tasks under a unified formulation without modifying the pretrained backbone. Our key insight is that any motion-related task can be decomposed at the frame level into exactly one of three mutually exclusive meta-operations with respect to the source motion context: it is either preserved (identity), generated without reference motion (context-free generation), or generated based on existing content (context-based generation). Regardless of whether a task involves temporal inpainting, text-guided motion editing, spatial control, or multi-identity reaction generation, the per-frame intent always reduces to one of these atomic meta-operations. Based on this observation, we introduce three learnable frame-level meta-operation embeddings (`[preserve]`, `[generate]`, and `[edit]`) that specify per-frame intent to the generative backbone. On the language conditioning side, all task conditions, from semantic descriptions and editing instructions to precise 3D spatial constraints, are expressed as text and encoded by the same pretrained LLM, requiring no task-specific conditioning module.

Specifically, we adopt HY-Motion [56] as our motion LFM backbone, a large-scale DiT-based T2M model that provides robust generative priors. By augmenting each source motion frame with its corresponding meta-operation embedding and injecting the resulting task-aware sequence into the pretrained backbone via lightweight temporal fusion, our framework extends the T2M-only LFM to naturally cover a broad spectrum of previously unsupported tasks (see Fig. 1), including temporal inpainting [7] (*e.g.*, prediction, backcasting, in-betweening, and keyframe control) and text-guided motion editing via either fine-grained instructions [1] or high-level style prompts [24, 77]. Furthermore, UMO enables geometric constraints [26, 45, 54] in text-format without optimization and multi-identity reaction generation [2, 13, 22, 63] in a single unified model. Notably, all these tasks lie beyond the base model's original text-to-motion training scope, yet are handled by a single unified model through two key technical contributions: (1) frame-level meta-operation embeddings that explicitly specify per-frame generative intent, and (2) a lightweight and effective temporal fusion mechanism that injects this task-aware conditioning into the DiT backbone, requiring no task-specific modules or architectural changes.

Comprehensive experiments show that the generative prior learned from a *single pretraining objective* (text-to-motion) can be efficiently unlocked to support *various downstream motion generation tasks*, shown in Fig. 2. We also conduct systematic ablations among different conditioning architectures, including temporal fusion [9, 41], sequence concatenation [5, 51], AdaLN [43], and ControlNet [74]. Achieving state-of-the-art performance across these tasks with a single unified model demonstrates both the transferability and generalization of motion priors learned from large-scale T2M pretraining, and the effectiveness of our in-context motion conditioning design.

In summary, our contributions are:

– **A unified in-context formulation grounded in three atomic meta-operations.** We identify that per-frame intent in any in-context motion task

reduces to one of three mutually exclusive operations—preserve, generate, or edit. The corresponding learnable embeddings, combined with unified language conditioning, express diverse downstream tasks without architectural changes to the LFM backbone.
- **A lightweight and efficient conditioning design with systematic validation.** We propose temporal fusion to inject in-context cues at the input embedding level with only 0.207M additional parameters and negligible runtime overhead, and validate this choice against alternative architectures.
- **Evidence of broad prior transferability.** A single unified UMO consistently outperforms task-specific baselines across temporal completion, text-guided editing, geometric constraints, and multi-identity reaction generation, all beyond the base model's original training scope.

2 Related Work

2.1 Diverse Motion Generation Downstream Tasks

Motion generation downstream tasks can be organized around a fundamental distinction: whether the model operates solely from external signals, or must additionally comprehend and respect existing motion context. **Cross-modal motion generation**, *i.e.*, $(C_0, C_1, \cdots) \mapsto X$, requires mapping from one or more conditioning modalities into plausible motion sequences. Representative modalities include audio [3, 15, 42, 64], 3D environments [8, 21, 25, 31], spatial constraints [26, 45, 54, 60], 3D objects [44, 65–68], videos [34, 71, 73, 75], or their combinations. **In-context motion generation**, *i.e.*, $(X, C) \mapsto X$, requires balancing faithfulness to the source motion with responsiveness to the condition, based on overall comprehension. This category can be further divided along two axes. *Homogeneous* tasks operate on a single motion entity and require maintaining self-consistency under textual or temporal instructions, including motion editing [1, 4, 24, 32] and temporal completion [7, 19, 49]. *Heterogeneous* tasks involve multiple motion entities and demand inter-entity coordination, such as reaction generation [46, 58, 63].

Despite the growing diversity of these tasks, existing approaches mainly develop task-specific architectures with dedicated training pipelines, resulting in fragmented solutions that cannot share knowledge across tasks. A few recent efforts [19, 39] attempt to unify multiple tasks within a single framework, yet they are typically limited by insufficient data for unified task learning from scratch and fail to outperform the specialist task models. In this work, we instead unlock the powerful motion priors in large-scale pretrained T2M models via a lightweight supervised fine-tuning strategy, which effectively adapts a single foundation model to a diverse set of downstream tasks.

2.2 Text-to-Motion Foundation Models

Early text-to-motion (T2M) models explored a variety of generative paradigms at relatively small scale, like diffusion-based methods [10, 52, 78], masked modeling approaches [16,17], and next-token-prediction [59,72]. However, these models

were typically trained on limited datasets (*e.g.*, HumanML3D [18]), restricting the richness of the learned motion priors. More recently, inspired by scaling laws, current works leverage more scalable architecture designs such as decoder-only transformer [12, 30, 38] and diffusion transformer [34, 56] to achieve impressive generation quality in open-vocabulary settings, substantially scaling up both data and model capacity.

3 Method

3.1 Preliminary

Flow Matching. We adopt the rectified flow formulation [35, 37]. Let $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be Gaussian noise and $\mathbf{x}_1$ the clean motion. The interpolation path and target velocity are: $\mathbf{x}_t = (1-t)\,\mathbf{x}_0 + t\,\mathbf{x}_1, \mathbf{v}_t = \mathbf{x}_1 - \mathbf{x}_0, t \in [0, 1]$. A neural network $\mathbf{v}_\theta$ is trained to regress this velocity: $\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[\|\mathbf{v}_\theta(\mathbf{x}_t, \mathbf{c}, t) - \mathbf{v}_t\|^2\right]$, where $\mathbf{c}$ denotes the conditions. At inference, motion is generated by solving the ODE $\mathrm{d}\mathbf{x}/\mathrm{d}t = \mathbf{v}_\theta(\mathbf{x}_t, \mathbf{c}, t)$ from $t=0$ to $t=1$.

HY-Motion. HY-Motion [56] is the first DiT-based motion foundation model pretrained on over 3,000 hours of motion data. It employs a multimodal DiT (MMDiT) [11] architecture ($460M/1B$ parameters) with text conditioning via an LLM encoder (Qwen3-8B [69]) and a CLIP encoder [47].

Motion Representation. Following HY-Motion, each motion is represented as a vector $\mathbf{f} \in \mathbb{R}^{201}$, comprising global root translation (3D), root orientation in 6D continuous rotation, 21 local joint rotations (each 6D), and 22 local joint positions (each 3D). All motions are resampled to 30 fps and normalized to zero mean and unit variance. A motion sequence of T frames is denoted $\mathbf{x} \in \mathbb{R}^{T\times 201}$.

3.2 Unified In-Context Motion Generation Formulation

Our key insight is: for any target motion sequence to be synthesized, every frame stands in exactly one of three mutually exclusive relationships with respect to the source motion context: (i) the frame should be preserved exactly as provided, (ii) the frame should be generated from scratch, or (iii) the frame should be edited based on the source motion. These three relationships are collectively exhaustive: regardless of whether a task involves temporal completion, semantic editing, spatial control, or cross-identity conditioning, the per-frame intent always reduces to one of these three atomic meta-operations.

Frame-Level Meta-Operation Embeddings. We introduce three learnable frame-level meta-operation embeddings, `[preserve]` (`P`), `[generate]` (`G`), and `[edit]` (`E`), each in $\mathbb{R}^{201}$, that represent fine-grained per-frame semantic intent, refining the input embedding of the generative backbone: `[preserve]`: retain this frame unchanged; `[generate]`: synthesize this frame from scratch; `[edit]`: modify the source motion content. To produce a T-frame motion sequence, each frame i is assigned a meta-operation embedding $\tau_i \in \{\texttt{P}, \texttt{G}, \texttt{E}\}$ and paired with

Table 1: Text prompt templates. All task instructions are expressed as text, enabling a unified neural-symbolic interface without any architectural modification.

Category	Template	Example
Description	`<motion description>`	A man kicks something with his left leg.
Editing Instruction	`<editing instruction>`	Speed up your motion.
Parameterized Trajectory	`{type:<curve_type>, params:attribute1:value1, attribute2:value2, ...}`	{type:cubic_bézier, params:{start:[0.0,0.0], end:[5.22,3.77], P0:[0.0,0.0], P1:[-0.23,3.95], P2:[5.44,-0.17], P3:[5.22,3.77]}}
Spatial Constraint	`A person walks from (`x_1`,`y_1`) to (`x_2`,`y_2`). Avoiding N obstacles at (`x`,`y`, `r`), ..., where `r` is the safety radius in meters.`	A person walks from (0.00, 0.00) to (3.96, 6.19). Avoiding 3 obstacles at (2.47, 3.04, 0.44), (2.78, 3.82, 0.45), (2.97, 4.68, 0.39), where r is the safety radius in meters.

a source motion $\mathbf{s}_i \in \mathbb{R}^{201}$, set to the original motion $\mathbf{m}_i$ for preservation and editing or to $\mathbf{0}$ when no source reference exists.

Unified Language Conditioning Design. While the meta-operation embeddings unify what to do at each frame on the motion side, we further unify how to do it on the language conditioning side: all task conditions, from semantic descriptions and editing instructions to precise 3D spatial constraints, are expressed as language and encoded by the same pretrained LLM, without any task-specific conditioning module. As summarized in Table 1, we design four prompt templates spanning from natural language to structured specifications: *Description* serves as the standard semantic interface; *Editing Instruction* expresses the desired modification; *Parameterized Trajectory* serializes trajectories into structured language; and *Spatial Constraint* describes scene-level 3D constraints as compositional language. This design is also inherently extensible: supporting a new constraint type requires only a new prompt template, with no architectural changes.

The Parameterized Trajectory template defines a hierarchy of parameterized curves (line, arc, sinusoid, Bézier, *etc.*) via a structured format `{type: <curve_type>, params:{...}}`, where the parameters uniquely determine the curve's geometric properties such as start/end points, control points, curvature, amplitude, and frequency. The Spatial Constraint template composes a natural-language sentence specifying the start and goal positions together with a list of obstacles, each described by its center coordinates and safety radius, *e.g.*, "A person walks from (x_1, y_1) to (x_2, y_2), Avoiding N obstacles at $(x, y, r), \ldots$". Both templates encode precise 3D geometric specifications entirely as language, requiring no specialized spatial conditioning module.

Task Instantiation via Composition. Together, as summarized in Table 2, the meta-operation embeddings and language conditions form a complete unified interface: all downstream tasks are fully specified by a per-frame $(\mathbf{s}, \boldsymbol{\tau})$ configuration plus a language condition. Notably, the three meta-operations form a complete and minimal basis for frame-level intent, and the language conditioning provides an equally open-ended condition space. Any downstream task, including those not yet conceived, can be instantiated by simply defining a new $(\mathbf{s}, \boldsymbol{\tau})$ pattern and a corresponding prompt template, with no architectural changes.

Table 2: Unified task formulation. Each task is specified by the per-frame source motion $\mathbf{s}$ and meta-operation embedding $\boldsymbol{\tau} \in \{\mathtt{P}, \mathtt{G}, \mathtt{E}\}$. $\star$ denotes variant configurations: $(\mathbf{s}=\mathbf{m}, \mathtt{E})$ for editing; $(\mathbf{s}=\mathbf{0}, \mathtt{G})$ for generation.

Task	Configuration ($\mathbf{s}$ top / $\boldsymbol{\tau}$ bottom)
Text-to-Motion / Trajectory Following / Obstacle Avoidance	$s_i = \mathbf{0}, \quad \tau_i = \mathtt{G} \quad \forall\, i$
Instruction-Based Editing / Reaction / Stylization / Trajectory Following / Obstacle Avoidance	$s_i = \mathbf{m}_i, \quad \tau_i = \mathtt{E} \quad \forall\, i$
Prediction	$[\mathbf{m}_1, \ldots, \mathbf{m}_k, \star, \ldots, \star]$ $[\mathtt{P}, \ldots, \mathtt{P}, \star, \ldots, \star]$
Backcasting	$[\star, \ldots, \star, \mathbf{m}_{T-k+1}, \ldots, \mathbf{m}_T]$ $[\star, \ldots, \star, \mathtt{P}, \ldots, \mathtt{P}]$
In-betweening	$[\mathbf{m}, \ldots, \mathbf{m}, \star, \ldots, \star, \mathbf{m}, \ldots, \mathbf{m}]$ $[\mathtt{P}, \ldots, \mathtt{P}, \star, \ldots, \star, \mathtt{P}, \ldots, \mathtt{P}]$
Keyframe Infilling	$[\mathbf{m}, \star, \ldots, \star, \mathbf{m}, \star, \ldots, \star, \mathbf{m}]$ $[\mathtt{P}, \star, \ldots, \star, \mathtt{P}, \star, \ldots, \star, \mathtt{P}]$

3.3 In-Context Conditioning Architecture

Given the formulation above, the remaining design question is how to encode and inject the in-context motion signal $(\mathbf{s}, \boldsymbol{\tau})$ into the DiT backbone. We first describe the encoding process, then compare four injection architectures.

In-Context Motion Encoder. The meta-operation embedding is added to the source to form the task-aware in-context input:

$$\tilde{\mathbf{s}} = \mathbf{s} + \mathrm{Emb}(\boldsymbol{\tau}), \tag{1}$$

where $\mathbf{s} = [\mathbf{s}_1, \ldots, \mathbf{s}_T]$ and $\boldsymbol{\tau} = [\tau_1, \ldots, \tau_T]$. The task-aware input $\tilde{\mathbf{s}}$ is then encoded by an MLP encoder E_{ctx}, initialized as a copy of the pretrained input encoder E_{in} of HY-Motion, inheriting its representation capacity while subsequent finetuning specializes it for noise-free, task-conditioned motion context.

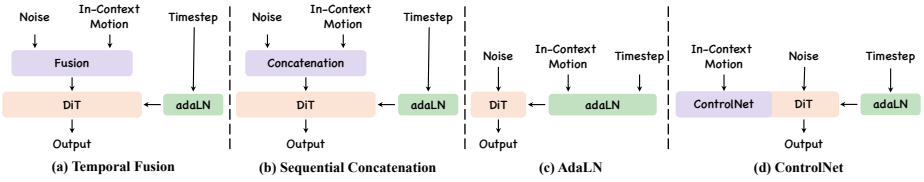

Fig. 3: Architecture Alternatives for In-Context Motion Conditioning.

Architecture Alternatives. Given the encoded features $E_{\mathrm{ctx}}(\tilde{\mathbf{s}})$, we consider four representative in-context injection architectures, illustrated in Fig. 3, spanning from input-level fusion to layer-wise modulation:
(a) *Temporal Fusion* adds $E_{\mathrm{ctx}}(\tilde{\mathbf{s}})$ element-wise to $E_{\mathrm{in}}(\mathbf{x}_t)$ before the DiT blocks, *i.e.*, $\mathbf{x}'_t = E_{\mathrm{in}}(\mathbf{x}_t) + E_{\mathrm{ctx}}(\tilde{\mathbf{s}})$, where $\mathbf{x}_t$ denotes the noisy motion at timestep t.

This fuses in-context motion at the input level with minimal parameter overhead (only 0.207M) while preserving full per-frame granularity.
(b) *Sequential Concatenation* concatenates $E_{\text{ctx}}(\tilde{\mathbf{s}})$ with $E_{\text{in}}(\mathbf{x}_t)$ along the sequence dimension ($2T$ tokens), with an asymmetric attention mask that makes the in-context motion a read-only conditioning signal. This preserves per-frame granularity but significantly increases computational cost.
(c) *AdaLN* compresses $E_{\text{ctx}}(\tilde{\mathbf{s}})$ into a global vector via attention pooling and injects it through adaptive layer normalization alongside the timestep embedding. This provides efficient layer-wise modulation but sacrifices per-frame granularity by collapsing the entire source motion into a single vector.
(d) *ControlNet* [74] freezes the backbone and clones its double-stream blocks into a parallel trainable branch, injecting in-context motion through zero-initialized residual connections. While this preserves the original backbone weights, it introduces substantial parameter overhead (234M) and restricts the backbone from efficiently adapting to in-context conditioning.
We systematically ablate all four architectures in Sec. 4.2.

4 Experiment

We first ablate the architecture design for injecting in-context motion features (Sec. 4.2), then systematically evaluate how the generative priors of HY-Motion can be exploited for various downstream tasks and how far these priors can be pushed in more chandelling scenarios, organizing our evaluation along a progressive difficulty axis: in-domain adaptation (Sec. 4.3), out-of-domain adaptation (Sec. 4.4), and stress tests that probe the boundaries of both the text encoder and the motion backbone (Sec. 4.5).

4.1 Experimental Setup

Datasets. We use HumanML3D [18] for text-to-motion generation and all temporal inpainting subtasks (prediction, backcasting, in-betweening, and keyframe infilling), MotionFix [1] for instruction-based motion editing, Inter-X [62] and InterHuman [33] for reaction generation. For parameterized trajectory control and obstacle avoidance, we construct a dedicated dataset of 2,000 sequences per task. Details are provided in the supplementary material.

Evaluation Metrics. For text-to-motion generation and temporal inpainting, we follow the standard evaluation protocol [17, 18, 59] and report: Fréchet Inception Distance (FID), measuring the distributional distance between generated and real motions; R-Precision (R@1/2/3) and Multimodal Distance (MM-Dist), measuring text-motion semantic alignment. For temporal inpainting, we additionally report Mean Per-Joint Position Error (MPJPE, cm) over the full sequence, and `[P]`-MPJPE (cm) restricted to frames assigned `[preserve]` tokens, measuring how faithfully the model respects the conditioning frames. For instruction-based motion editing, following [1, 24, 32], we report Generated-to-Target retrieval R@1/2/3 and average rank (AvgR) under both batch-level and

Table 3: Ablation Study on Four Architecture Choices on Keyframe Infilling. ΔP, ΔF, and ΔL denote the additional parameters, FLOPs, and latency over the HY-Motion base model when generating a 360-frame motion sequence with 50-step Euler ODE solver. **Bold** indicates best, underline indicates second best. We convert generated results to 272-representation to calculate the FID and R-Precision via official evaluator from MotionStreamer [59]. We downsample the results from 30 fps to 20 fps to calculate [P]-MPJPE to match the evaluation Tab. 5.

Architectures	[P]-MPJPE ↓	FID ↓	R@1 ↑	R@2 ↑	R@3 ↑	ΔP (M) ↓	ΔF (G) ↓	ΔL (s) ↓
ControlNet	5.19	6.520	0.688	0.835	0.889	234.2	85.12	0.49
AdaLN	11.1	8.860	**0.746**	**0.875**	**0.922**	4.400	1.660	0.02
Sequential Concat	2.04	11.77	0.680	0.831	0.887	**0.207**	198.6	0.89
Temporal Fusion	**0.95**	**0.476**	0.692	0.842	0.896	**0.207**	**0.140**	**0.01**

full-set protocols to evaluate editing accuracy. For geometric-constrained generation, we report Trajectory Error measuring the average deviation from the target trajectory, inference Latency, and obstacle avoidance Success Rate.

Implementation Details. We employ HY-Motion-Lite (460M) [56] as the pretrained base model and train all models on 4 NVIDIA B200 GPUs with a batch size of 256 and a learning rate of 5×10^{-5}. The unified model is jointly trained on all tasks across all datasets for $100k$ steps, while each expert model is trained on its respective task for $6k$ steps. During inference, we use a 50-step Euler ODE solver with a classifier-free guidance scale of 2.0. All evaluation experiments are conducted on NVIDIA A6000 GPUs.

Baselines. We compare UMO with state-of-the-art methods for each task [1,6,7, 17, 19, 23, 24, 26, 32, 33, 45, 46, 52, 54, 56, 59, 60, 70, 72, 76], all following their official open-source implementations. Notably, to evaluate the vanilla performance of HY-Motion in temporal inpainting tasks, we implement a baseline based on the classic training-free inversion strategies [49, 50]. Details are provided in the supplementary material.

4.2 Architecture Ablation

We ablate all four architecture variants described in Sec. 3.3 on the keyframe infilling subtask of temporal inpainting, with quantitative results in Tab. 3. The results reveal a clear hierarchy. Temporal Fusion achieves the best [P]-MPJPE and FID with minimal overhead (ΔP=0.207M, ΔL=0.01s), balancing faithfulness to the source with responsiveness to the condition. Although AdaLN attains the highest R-Precision but critically fails at per-frame constraint, it largely ignores the [preserve] token according to its worst [P]-MPJPE among all variants, which validates that compressing the entire source motion into a single global vector fundamentally cannot convey frame-level intent like temporal fusion. Sequential Concatenation and ControlNet significantly introduce more extra parameters and runtime overhead, yet yield suboptimal performance. The qualitative comparison in Fig. 4 corroborates these findings, where ControlNet

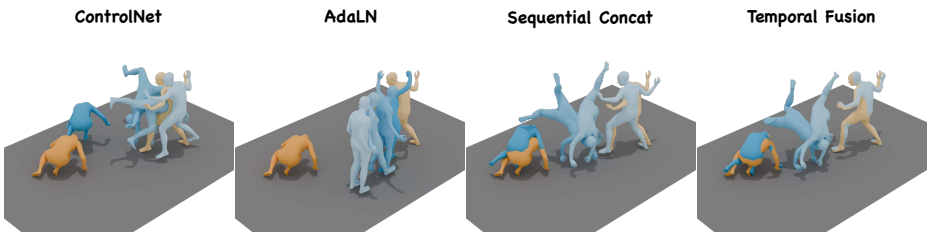

Fig. 4: Ablation of In-context Conditioning Architectures in Keyframe Infilling.

Table 4: Comparison of Text-to-Motion on HumanML3D. **Bold** indicates best, underline indicates second best. We used official evaluator from MotionStreamer [59].

Methods	FID ↓	R@1 ↑	R@2 ↑	R@3 ↑	MM-D ↓
MDM [52]	23.454	0.523	0.692	0.764	17.42
T2M-GPT [72]	12.475	0.606	0.774	0.838	16.81
MoMask [17]	12.232	0.621	0.784	0.846	16.14
MotionStreamer [59]	11.790	0.631	0.802	0.859	16.08
HY-Motion [56]	61.035	0.667	0.818	0.876	17.53
UMO-Expert	17.04	0.763	0.889	0.931	15.49
UMO-Unified	**9.460**	**0.774**	**0.892**	**0.933**	**15.22**

exhibits noticeable motion drifting and AdaLN fails to respect the keyframe positions. To this end, we adopt Temporal Fusion as the default architecture in all subsequent experiments, as it achieves the best trade-off among performance, efficiency, and simplicity. Hereafter, UMO-Expert and UMO-Unified denote Temporal Fusion models trained for a single purpose and jointly across all tasks, respectively. More ablations in various tasks can be found in the supplementary.

4.3 In-Domain Task Adaptation

With Temporal Fusion established as our default architecture, we now evaluate UMO along a progressive difficulty axis, beginning with text-to-motion generation on HumanML3D, which is most closely aligned with the pretraining objective. As shown in Tab. 4, directly evaluating HY-Motion on HumanML3D [18] yields high FID despite competitive R-Precision, exposing a distribution gap between its pretraining corpus and HumanML3D rather than a lack of semantic capability, which confirms the necessity of appropriate fine-tuning. With lightweight adaptation, UMO-Expert closes this gap dramatically, demonstrating that fine-tuning successfully activates the semantic priors of the pretrained model. UMO-Unified further advances all metrics, achieving a state-of-the-art FID that surpasses even dedicated text-to-motion models. This shows a key advantage of our unified multi-task formulation: jointly training across diverse tasks provides complementary supervision that strengthens each individual task, rather than diluting performance through task competition.

Table 5: Comparison of Temporal Inpainting on HumanML3D based on the MotionLab [19] evaluator. We benchmark four subtasks in temporal inpainting. **Bold** indicates best, underline indicates second best. For fair comparison, we downsample UMO and HY-Motion generated results from 30 FPS to 20 FPS and convert them to 263-representation [18] to match the evaluation protocol of all metrics.

Methods	MPJPE ↓	[P]-MPJPE ↓	FID ↓	R@1 ↑	R@2 ↑	R@3 ↑	MM-D ↓
Prediction							
HY-Motion	13.8	0.02	18.23	0.344	0.515	0.619	4.829
CondMDI [7]	11.2	**0.38**	1.592	0.146	0.244	0.314	6.619
MotionLab [19]	12.3	0.81	2.844	0.179	0.283	0.374	6.170
UMO-Expert	10.3	0.58	0.087	0.536	0.734	0.832	2.836
UMO-Unified	**10.2**	0.54	**0.056**	**0.553**	**0.744**	**0.834**	**2.823**
Backcasting							
HY-Motion	13.6	2.28	9.299	0.386	0.573	0.682	4.174
CondMDI [7]	10.9	2.18	2.408	0.131	0.226	0.297	6.895
MotionLab [19]	12.1	3.78	1.469	0.231	0.352	0.448	5.413
UMO-Expert	9.80	1.66	0.182	0.521	0.728	0.814·	3.067
UMO-Unified	**9.55**	**1.61**	**0.057**	**0.563**	**0.742**	**0.839**	**2.808**
In-Betweening							
HY-Motion	13.7	0.97	21.97	0.319	0.474	0.577	5.282
CondMDI [7]	10.0	1.08	0.957	0.155	0.252	0.333	6.619
MotionLab [19]	10.0	1.37	2.124	0.268	0.405	0.495	5.117
UMO-Expert	8.75	0.91	0.340	0.501	0.704	0.802	3.025
UMO-Unified	**8.55**	**0.73**	**0.050**	**0.541**	**0.731**	**0.827**	**2.849**
Keyframe Infilling							
HY-Motion	13.9	1.03	36.82	0.206	0.319	0.394	6.618
CondMDI [7]	2.72	**0.92**	0.803	0.160	0.277	0.366	6.471
MotionLab [19]	4.01	1.44	0.117	0.495	0.673	0.776	3.131
UMO-Expert	**2.06**	0.95	0.075	0.496	0.698	0.794	3.032
UMO-Unified	2.67	0.95	**0.040**	**0.514**	**0.707**	**0.804**	**2.974**

4.4 Out-of-Domain Tasks Adaptation

We evaluate UMO on two categories of out-of-domain tasks, where out-of-domain manifests differently: temporal inpainting challenges the backbone with an in-context generation structure entirely absent from pretraining, while instruction-based editing additionally shifts the language distribution, requiring the text encoder to interpret editing instructions rather than motion descriptions.

Temporal Inpainting. Tab. 5 reports results across four subtasks: prediction, backcasting, in-betweening, and keyframe infilling. Training-free solution [49, 50] of HY-Motion fails substantially across all subtasks, confirming that naively repurposing the pretrained model is insufficient. With proper fine-tuning, both UMO-Expert and UMO-Unified consistently surpass CondMDI [7] and MotionLab [19] across all four subtasks in generation quality (FID, R-Precision, MM-Dist). On frame preservation ([P]-MPJPE), UMO achieves competitive results

Table 6: Comparison of Instruction-Based Editing on the MotionFix Dataset. Our method outperforms all other baselines on generated-to-target retrieval. We report R@1, R@2 and R@3 as percentages. ↑ indicates higher values are better, ↓ indicates lower values are better, **bold** indicates best, and underline indicates second best.

Methods	Generated-to-Target (Batch)				Generated-to-Target (Full)			
	R@1↑	R@2↑	R@3↑	AvgR↓	R@1↑	R@2↑	R@3↑	AvgR↓
Ground Truth	100.0	100.0	100.0	1.00	64.36	88.75	95.56	1.74
Baselines								
MDM	4.03	7.56	10.48	15.55	0.10	0.10	0.10	–
MDM-BP	39.10	50.09	54.84	6.46	8.69	14.71	18.36	180.99
TMED	62.90	76.51	83.06	2.71	14.51	21.72	28.73	56.63
SimMotionEdit	70.62	82.92	88.12	2.38	25.49	39.33	49.21	23.49
motionReFit	66.33	80.05	84.98	2.64	14.13	23.52	30.53	54.06
MotionLab	72.65	82.71	87.89	2.20	–	–	–	–
PartMotionEdit	73.96	85.83	90.21	1.92	27.27	45.06	53.36	16.24
UMO-Expert	**98.08**	**99.90**	**100.0**	**1.02**	60.91	82.13	91.31	1.76
UMO-Unified	**98.08**	**99.90**	**100.0**	**1.02**	**61.70**	**82.23**	**91.51**	**1.75**

through learned `[preserve]` embeddings alone, without any hard replacement of conditioning frames, which is a fundamentally different mechanism from those used in CondMDI [7] and HY-Motion baseline. More details are in Appendix.

Instruction-Based Motion Editing. As shown in Tab. 6, both UMO-Expert and UMO-Unified achieve near-perfect batch-level retrieval (R@3=100.0%), substantially outperforming all prior task-specific methods. Under the more challenging full-set protocol, UMO-Unified slightly outperforms UMO-Expert in R-Precision, suggesting that exposure to diverse tasks during joint training enhances the model's fine-grained editing capability. The qualitative comparison in Fig. 5 further illustrates that UMO applies precise, instruction-aligned edits while faithfully preserving unmodified semantics, confirming the effectiveness of the `[edit]` embedding in conveying fine-grained editing intent. Notably, despite never encountering editing instructions during pretraining, the general motion priors can be effectively repurposed for instruction following through our in-context formulation, an early indication of the emergent capability we further explore in the next sections.

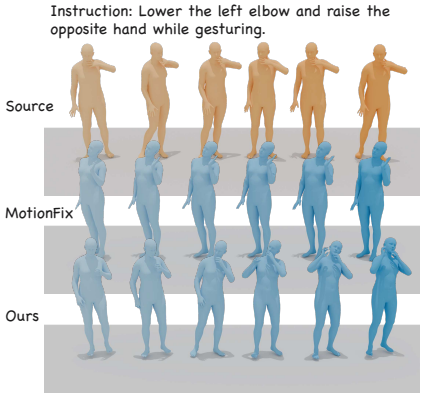

Fig. 5: Qualitative Comparison of Text-guided Motion Editing.

4.5 Stress Test: Ability Emergence

Beyond out-of-domain adaptation, we further probe the boundaries of pretrained priors with two stress tests that target fundamentally different sources of emer-

gence. The first is geometric-constrained generation, where structured spatial specifications are in-domain inputs for pretrained LLMs but represent a completely foreign input distribution to the DiT backbone. The second is dual-identity reaction generation, a fundamentally different task topology, where single-person motion priors must generalize to two-person interaction. In both cases, the question is whether targeted fine-tuning can unlock latent competencies already implicitly embedded in large-scale pretraining.

Geometric Constraints. As described in Sec. 3.2, we encode parameterized trajectories, including line, arc, sinusoid, Bézier, etc, and spatial constraints, *i.e.*, obstacles, entirely as structured language prompts, encoded by the same pretrained LLM encoder without any specialized spatial conditioning module. This design introduces a fundamental challenge: while the LLM encoder may have encountered mathematical notation and coordinate data in its pretraining corpus, the DiT backbone has never been exposed to the resulting geometric features as motion conditioning signals. Consequently, zero-shot evaluation of HY-Motion fails completely (Traj. Err=325.1cm), as the model cannot interpret this entirely new distribution of language features. We hypothesize that the LLM encoder has already internalized a latent capacity for structured geometric understanding through its pretraining on corpora rich in mathematical notation, code, and coordinate data. Targeted fine-tuning, in this view, need not teach geometric reasoning from scratch but rather activate this dormant capacity and align it with the DiT's generative space. Tab. 7 strongly confirms our hypothesis. Among feed-forward methods, UMO-Unified matches OmniControl in trajectory error while being around 90 times faster. Although optimization-based methods achieve higher precision, they incur latency that is orders of magnitude

Table 7: Geometric-Constrained Generation. We evaluate trajectory following and obstacle avoidance. Traj. Err: average trajectory error (cm). Latency: average inference time (s) per sample. Succ.: obstacle avoidance success rate (%). **Bold**: best, underline: second best. • means feed-forward solutions, • means optimization-based solutions.

	Traj. Err ↓	Latency ↓	Succ. ↑
• HY-Motion	325.1	0.755	–
• CondMDI	70.46	38.44	–
• OmniControl	17.89	68.10	–
• TLControl w/o Opt.	61.15	**0.019**	–
• TLControl w/ Opt.	**2.930**	3.243	–
• MaskControl	3.064	31.50	0.93
• DNO	–	174.9	0.88
• UMO-Expert	21.29	0.759	0.92
• UMO-Unified	18.78	0.759	**0.95**

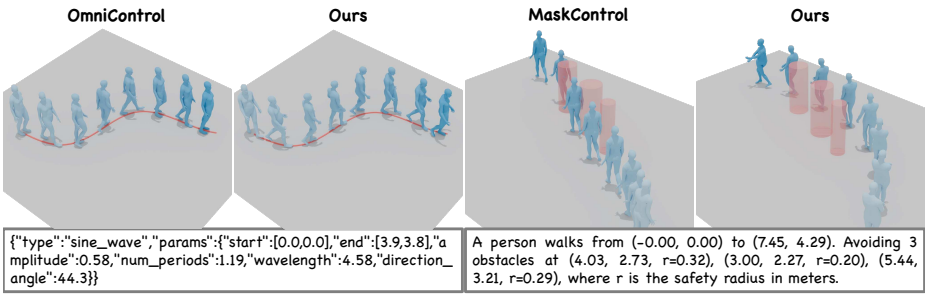

Fig. 6: Qualitative Comparison of Text-serialized Geometric Constraints.

greater. Fig. 6 further illustrates that UMO produces geometrically faithful motions. Together, these results suggest that large-scale T2M pretraining implicitly encodes transferable priors that extend into structured spatial reasoning, and since our language conditioning design requires only a new prompt template to incorporate new constraint types, this points to a scalable path toward broader geometric tasks.

Dual Identity Reaction Generation. Our second stress test probes whether single-person motion priors can generalize to two-person interaction, a setting entirely absent from pretraining. As shown in Tab. 8, UMO achieves the best FID across all methods, though InterMask [22] retains a higher R@1, suggesting a trade-off between distributional fidelity and text-motion alignment. Notably, HY-Motion was never trained on two-person data, yet produces reactions with higher overall motion quality than dedicated multi-person models. This surprising emergence suggests that core motion priors, such as body dynamics, joint coordination, and temporal coherence, are inherently transferable from single-person to multi-person scenarios, and our in-context formulation can effectively activate them for dual-identity generation.

Table 8: Reaction Generation on InterHuman. **Bold**: best, underline: second best.

	FID ↓	R@1 ↑
InterGen	52.89	0.194
InterMask	2.990	**0.462**
UMO-Expert	<u>2.130</u>	0.428
UMO-Unified	**2.055**	<u>0.431</u>

5 Limitations

While UMO demonstrates strong generalization across diverse tasks, several limitations remain. First, the three frame-level meta-operation embeddings operate at the whole-body level and do not support finer-grained part-level control; extending the formulation to a part-specific dimension is a natural direction for future work. Second, in this paper, we focus on language conditioning and does not handle audio signals such as music or speech. Incorporating such modalities requires a multimodal foundation model with audio-language encoding [61].

6 Conclusion

We present UMO, a unified in-context adaptation framework that unlocks pretrained motion foundation model priors for diverse downstream tasks within a single model. The core insight is that any frame-level intent can be expressed as one of three mutually exclusive meta-operations, preserve, generate, or edit, whose compositions compactly specify tasks ranging from temporal inpainting and instruction-based editing to geometric-constrained generation and dual-identity reaction generation. Equipped with only three learnable embeddings and a lightweight temporal fusion module, UMO consistently outperforms task-specific baselines across all benchmarks without any backbone modification. Two findings are particularly noteworthy: encoding geometric constraints purely in language space achieves results comparable to dedicated spatial conditioning

methods, and single-person motion priors transfer effectively to two-person interaction, providing strong evidence that large-scale T2M priors are broadly transferable given an appropriate in-context interface.

Acknowledgements. This research was supported by NSF CAREER grant #2143576, and a sponsored research award from Meta Inc.

References

1. Athanasiou, N., Ceske, A., Diomataris, M., Black, M.J., Varol, G.: MotionFix: Text-driven 3d human motion editing. In: SIGGRAPH Asia 2024 Conference Papers (2024)
2. Cen, Z., Pi, H., Peng, S., Shuai, Q., Shen, Y., Bao, H., Zhou, X., Hu, R.: Ready-to-react: Online reaction policy for two-character interaction generation. arXiv preprint arXiv:2502.20370 (2025)
3. Chen, C., Zhang, J., Lakshmikanth, S.K., Fang, Y., Shao, R., Wetzstein, G., Fei-Fei, L., Adeli, E.: The language of motion: Unifying verbal and non-verbal language of 3d human motion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6200–6211 (2025)
4. Chen, L.H., Lu, S., Dai, W., Dou, Z., Ju, X., Wang, J., Komura, T., Zhang, L.: Pay attention and move better: Harnessing attention for interactive motion generation and training-free editing. arXiv preprint arXiv:2410.18977 (2024)
5. Chen, R., Shi, M., Huang, S., Tan, P., Komura, T., Chen, X.: Taming diffusion probabilistic models for character control. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–10 (2024)
6. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18000–18010 (2023)
7. Cohan, S., Tevet, G., Reda, D., Peng, X.B., van de Panne, M.: Flexible motion in-betweening with diffusion models. In: ACM SIGGRAPH 2024 conference papers. pp. 1–9 (2024)
8. Cong, P., Wang, Z., Dou, Z., Ren, Y., Yin, W., Cheng, K., Sun, Y., Long, X., Zhu, X., Ma, Y.: Laserhuman: Language-guided scene-aware human motion generation in free environment. arXiv preprint arXiv:2403.13307 (2024)
9. Cong, X., Yang, H., Wang, A., Wang, Y., Yang, Y., Zhang, C., Ma, C.: Viva: Vlm-guided instruction-based video editing with reward optimization. arXiv preprint arXiv:2512.16906 (2025)
10. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9760–9770 (2023)
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
12. Fan, K., Lu, S., Dai, M., Yu, R., Xiao, L., Dou, Z., Dong, J., Ma, L., Wang, J.: Go to zero: Towards zero-shot motion generation with million-scale data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13336–13348 (2025)

13. Fan, K., Tang, J., Cao, W., Yi, R., Li, M., Gong, J., Zhang, J., Wang, Y., Wang, C., Ma, L.: Freemotion: A unified framework for number-free text-to-motion synthesis. In: European Conference on Computer Vision. pp. 93–109. Springer (2024)
14. Feng, Y., Dou, Z., Chen, L.H., Liu, Y., Li, T., Wang, J., Cao, Z., Wang, W., Komura, T., Liu, L.: Motionwavelet: Human motion prediction via wavelet manifold learning. arXiv preprint arXiv:2411.16964 (2024)
15. Ghosh, A., Zhou, B., Dabral, R., Wang, J., Golyanik, V., Theobalt, C., Slusallek, P., Guo, C.: Duetgen: Music driven two-person dance generation via hierarchical masked modeling. In: ACM SIGGRAPH (2025)
16. Guo, C., Hwang, I., Wang, J., Zhou, B.: Snapmogen: Human motion generation from expressive texts. arXiv preprint arXiv:2507.09122 (2025)
17. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1910 (2024)
18. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, T., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: CVPR (2022)
19. Guo, Z., Hu, Z., Soh, D.W., Zhao, N.: Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13869–13879 (2025)
20. Han, Z., Jiang, Z., Pan, Y., Zhang, J., Mao, C., Xie, C., Liu, Y., Zhou, J.: Ace: All-round creator and editor following instructions via diffusion transformer. arXiv preprint arXiv:2410.00086 (2024)
21. Hwang, I., Zhou, B., Kim, Y.M., Wang, J., Guo, C.: Scenemi: Motion in-betweening for modeling human-scene interaction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6034–6045 (2025)
22. Javed, M.G., Guo, C., Cheng, L., Li, X.: Intermask: 3d human interaction generation via collaborative masked modeling. arXiv preprint arXiv:2410.10010 (2024)
23. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36**, 20067–20079 (2023)
24. Jiang, N., Li, H., Yuan, Z., He, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Dynamic motion blending for versatile motion editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22735–22745 (2025)
25. Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1737–1747 (2024)
26. Karunratanakul, K., Preechakul, K., Aksan, E., Beeler, T., Suwajanakorn, S., Tang, S.: Optimizing diffusion noise can serve as universal motion priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1334–1345 (2024)
27. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
28. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
29. Li, J., Cao, J., Zhang, H., Rempe, D., Kautz, J., Iqbal, U., Yuan, Y.: Genmo: A generalist model for human motion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11766–11776 (2025)

30. Li, Z., An, S., Tang, C., Guo, C., Shugurov, I., Zhang, L., Zhao, A., Sridhar, S., Tao, L., Mittal, A.: Llamo: Scaling pretrained language models for unified motion understanding and generation with continuous autoregressive tokens. arXiv preprint arXiv:2602.12370 (2026)
31. Li, Z., Zhou, R., Sajnani, R., Cong, X., Ritchie, D., Sridhar, S.: GenHSI: Controllable generation of human-scene interaction videos. arXiv preprint arXiv:2506.19840 (2025)
32. Li, Z., Cheng, K., Ghosh, A., Bhattacharya, U., Gui, L., Bera, A.: Simmotionedit: Text-based human motion editing with motion similarity prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27827–27837 (2025)
33. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. International Journal of Computer Vision **132**(9), 3463–3483 (2024)
34. Lin, J., Wang, R., Lu, J., Huang, Z., Song, G., Zeng, A., Liu, X., Wei, C., Yin, W., Sun, Q., et al.: The quest for generalizable motion generation: Data, model, and evaluation. arXiv preprint arXiv:2510.26794 (2025)
35. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
36. Liu, S., Guo, C., Zhou, B., Wang, J.: Ponimator: Unfolding interactive pose for versatile human-human interaction animation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12068–12077 (2025)
37. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
38. Lu, S., Wang, J., Lu, Z., Chen, L.H., Dai, W., Dong, J., Dou, Z., Dai, B., Zhang, R.: Scamo: Exploring the scaling law in autoregressive motion generation model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27872–27882 (2025)
39. Luo, M., Hou, R., Li, Z., Chang, H., Liu, Z., Wang, Y., Shan, S.: M^3gpt: An advanced multimodal, multitask framework for motion comprehension and generation. Advances in Neural Information Processing Systems **37**, 28051–28077 (2024)
40. Mao, C., Zhang, J., Pan, Y., Jiang, Z., Han, Z., Liu, Y., Zhou, J.: Ace++: Instruction-based image creation and editing via context-aware content filling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1958–1966 (2025)
41. Mou, C., Sun, Q., Wu, Y., Zhang, P., Li, X., Ye, F., Zhao, S., He, Q.: Instructx: Towards unified visual editing with mllm guidance. arXiv preprint arXiv:2510.08485 (2025)
42. Ng, E., Zhang, S., Chen, Z., Zollhoefer, M., Richard, A.: Sarah: Spatially aware real-time agentic humans. arXiv preprint arXiv:2602.18432 (2026)
43. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
44. Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., Jiang, H.: Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2878–2888 (2025)
45. Pinyoanuntapong, E., Saleem, M., Karunratanakul, K., Wang, P., Xue, H., Chen, C., Guo, C., Cao, J., Ren, J., Tulyakov, S.: Maskcontrol: Spatio-temporal control for masked motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9955–9965 (2025)

46. Ponce, P.R., Barquero, G., Palmero, C., Escalera, S., Garcia-Rodriguez, J.: in2in: Leveraging individual information to generate human interactions. arXiv preprint arXiv:2404.09988 (2024)
47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
48. Sawdayee, H., Guo, C., Tevet, G., Zhou, B., Wang, J., Bermano, A.H.: Dance like a chicken: Low-rank stylization for human motion diffusion. arXiv preprint arXiv:2503.19557 (2025)
49. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023)
50. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
51. Tevet, G., Raab, S., Cohan, S., Reda, D., Luo, Z., Peng, X.B., Bermano, A.H., van de Panne, M.: Closd: Closing the loop between simulation and diffusion for multi-task character control. arXiv preprint arXiv:2410.03441 (2024)
52. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
53. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
54. Wan, W., Dou, Z., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Tlcontrol: Trajectory and language control for human motion synthesis. In: European Conference on Computer Vision. pp. 37–54. Springer (2024)
55. Wei, C., Liu, Q., Ye, Z., Wang, Q., Wang, X., Wan, P., Gai, K., Chen, W.: Univideo: Unified understanding, generation, and editing for videos. arXiv preprint arXiv:2510.08377 (2025)
56. Wen, Y., Shuai, Q., Kang, D., Li, J., Wen, C., Qian, Y., Jiao, N., Chen, C., Chen, W., Wang, Y., et al.: Hy-motion 1.0: Scaling flow matching models for text-to-motion generation. arXiv preprint arXiv:2512.23464 (2025)
57. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
58. Wu, Q., Dou, Z., Guo, C., Huang, Y., Feng, Q., Zhou, B., Wang, J., Liu, L.: Text2interact: High-fidelity and diverse text-to-two-person interaction generation. arXiv preprint arXiv:2510.06504 (2025)
59. Xiao, L., Lu, S., Pi, H., Fan, K., Pan, L., Zhou, Y., Feng, Z., Zhou, X., Peng, S., Wang, J.: Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10086–10096 (2025)
60. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. arXiv preprint arXiv:2310.08580 (2023)
61. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 (2025)
62. Xu, L., Lv, X., Yan, Y., Jin, X., Wu, S., Xu, C., Liu, Y., Zhou, Y., Rao, F., Sheng, X., et al.: Inter-x: Towards versatile human-human interaction analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22260–22271 (2024)

63. Xu, L., Zhou, Y., Yan, Y., Jin, X., Zhu, W., Rao, F., Yang, X., Zeng, W.: Regennet: Towards human action-reaction synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1759–1769 (2024)
64. Xu, S., Dou, Z., Shi, M., Pan, L., Ho, L., Wang, J., Liu, Y., Lin, C., Ma, Y., Wang, W., et al.: Mospa: Human motion generation driven by spatial audio. arXiv preprint arXiv:2507.11949 (2025)
65. Xu, S., Li, D., Zhang, Y., Xu, X., Long, Q., Wang, Z., Lu, Y., Dong, S., Jiang, H., Gupta, A., Wang, Y.X., Gui, L.Y.: Interact: Advancing large-scale versatile 3d human-object interaction generation. In: CVPR (2025)
66. Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In: ICCV (2023)
67. Xu, S., Ling, H.Y., Wang, Y.X., Gui, L.Y.: Intermimic: Towards universal whole-body control for physics-based human-object interactions. In: CVPR (2025)
68. Xu, S., Wang, Z., Wang, Y.X., Gui, L.Y.: Interdreamer: Zero-shot text to 3d dynamic human-object interaction. In: NeurIPS (2024)
69. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
70. Yang, Y., Zhang, Z., Chen, J., Wu, Z.: Partmotionedit: Fine-grained text-driven 3d human motion editing via part-level modulation. arXiv preprint arXiv:2512.24200 (2025)
71. Yu, C., Zhai, W., Yang, Y., Cao, Y., Zha, Z.J.: Hero: Human reaction generation from videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10262–10274 (2025)
72. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14730–14740 (2023)
73. Zhang, L., Li, Z., Li, T., Cao, Z., Xu, R., Long, X., Wang, W., Wang, J., Liu, Y., Wang, W., et al.: Egoreact: Egocentric video-driven 3d human reaction generation. arXiv preprint arXiv:2512.22808 (2025)
74. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
75. Zhao, C., Shu, J., Zhao, Y., Huang, T., Lu, J., Gu, Z., Ren, C., Dou, Z., Shuai, Q., Liu, Y.: Comovi: Co-generation of 3d human motions and realistic videos. arXiv preprint arXiv:2601.10632 (2026)
76. Zhong, C., Hu, L., Zhang, Z., Xia, S.: Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 509–519 (2023)
77. Zhong, L., Xie, Y., Jampani, V., Sun, D., Jiang, H.: Smoodi: Stylized motion diffusion model. In: European conference on computer vision. pp. 405–421. Springer (2024)
78. Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., Liu, Y., Komura, T., Wang, W., Liu, L.: Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In: European Conference on Computer Vision. pp. 18–38. Springer (2024)