Ego-HMB: Human Motion Bridging from Egocentric Images via A Motion Bridge Diffusion Model

Gaoge Han[1,2] **Yongkang Cheng**[1,2] **Shaoli Huang**[2] **Jing Zhang**[5]
Mingming Gong[3,1*] **Fakhri Karray**[1] **Tongliang Liu**[4,1*]

[1]**Mohamed bin Zayed University of Artificial Intelligence**
[2]**AgiBot** [3]**The University of Melbourne**
[4]**The University of Sydney** [5]**Wuhan University**

hangaoge@gmail.com

Abstract

In this paper, we move beyond the conventional setting of motion in-betweening that relies on structured pose inputs, and instead propose a new paradigm that is conditioned on egocentric (first-person) visual observations, thereby reducing the requirements on both users and upstream systems; we term this task Egocentric Human Motion Bridging. This is motivated by realistic consumer scenarios in VR/AR and robotics, where users typically have access only to raw egocentric imagery (e.g., from wearable devices) without reliable human pose annotations or motion capture signals. Unlike conventional motion interpolation in human pose-conditioned settings or motion generation conditioned on a complete temporal language or image context, the egocentric perspective introduces large viewpoint variations, partial body visibility, and severe temporal ill-posedness due to cross-modal endpoint constraints. To this end, we propose Ego-HMB, a unified diffusion-based model that internalizes both endpoint human pose estimation capability and temporally consistent 3D motion infilling ability within a single model through two training stages. Specifically, we first introduce Egocentric-to-Canonical Pretraining (pretraining stage), where we design a hierarchical curriculum with progressively increasing temporal spans. Then, we present a novel Motion Bridge Diffusion Training with endpoint and physics awareness as a fine-tuning stage, where our bridge diffusion diffuse from neighboring poses to enhance temporal continuity while preserving the global constraints between the two endpoints of a 3D motion sequence. Extensive experiments and comparisons with one-stage baselines (a generative model conditioned directly on two egocentric images) and two-stage baselines (estimation models followed by interpolation models) demonstrate that our approach consistently outperforms these baselines. Code is available at `https://github.com/gaogehan/Ego-HMB`.

1 Introduction

Human motion generation based on deep learning and diffusion models has achieved substantial success in VR/AR and robotics, significantly reducing the cost of motion creation for both professionals and general users while enabling diverse forms of interactive entertainment. Existing approaches can be broadly categorized into several classes: (1) multimodal-to-motion generation (typically conditioned on text or audio); (2) motion estimation (usually from video or monocular images); and (3) motion in-betweening/interpolation (given starting and ending poses). These tasks generally require dense user inputs or fully specified 3D pose sequences.

*Co-corresponding authors: Mingming Gong and Tongliang Liu. Emails: `mingming.gong@unimelb.edu.au` and `tongliang.liu@sydney.edu.au`, respectively.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

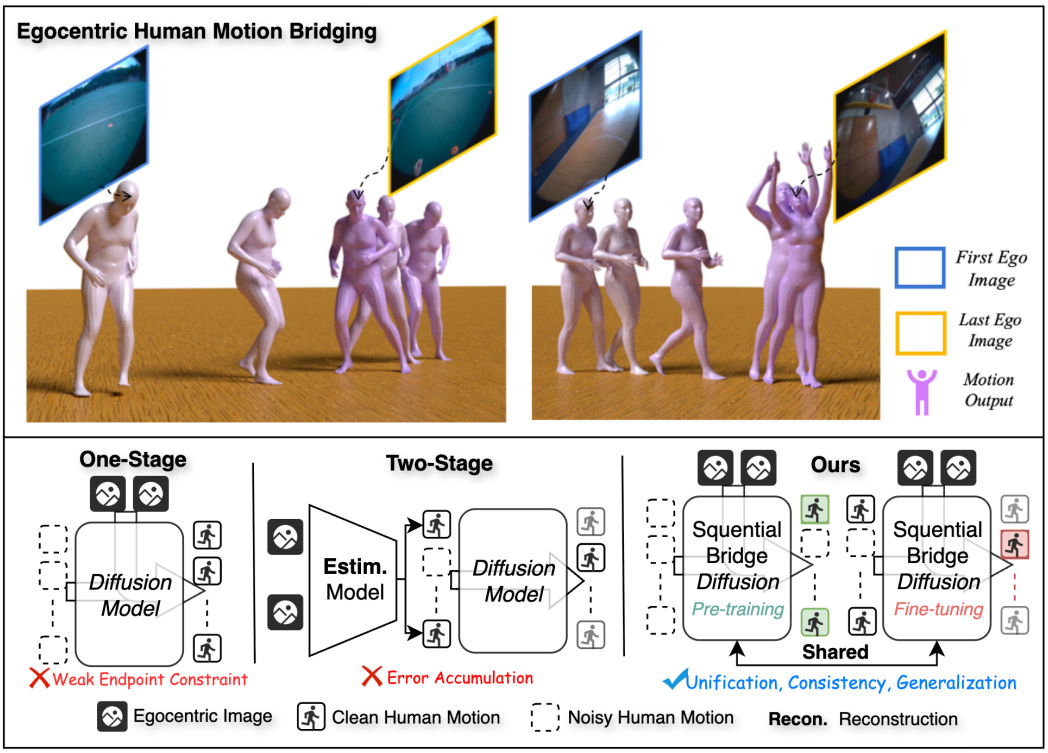

Figure 1: Egocentric Human Motion Bridging and Methods Comparison. Top: Egocentric human motion bridging from first-person observations to produce coherent intermediate motions; Bottom: Comparison of one-stage, two-stage, and Brownian bridge diffusion pipelines with improved training consistency and reduced error accumulation.

However, in common VR/AR and robotic application scenarios, users who intend to obtain motion sequences often have access only to sparse image representations, rather than complete pose trajectories. Generating full motion sequences from such sparse image inputs is therefore practically significant, as it reduces both user expression cost and storage overhead (by relying on sparse images instead of dense motion data).

To address this realistic requirement, we propose a new task termed egocentric human motion bridging task. In essence, the objective is to generate a coherent motion segment from two endpoint egocentric images. This task differs from existing paradigms in several aspects: 1) Unlike other cross-modal generation tasks, such as text- or scene-to-motion generation, where the conditioning signal usually describes a holistic description of the entire motion to be generated, our task is constrained by implicit start and end conditions. 2) It is not standard motion in-betweening/interpolation, since users are not required to provide complete 3D starting and ending poses. The notion of bridging refers to leveraging image modalities distinct from motion representations to connect human motion endpoints and synthesize the intermediate motion segment.

Therefore, how to effectively address the motion bridging task remains insufficiently and unsystematically explored in the existing literature. Based on current technical paradigms, two intuitive solution paths can be derived, both relying on the adaptation or composition of existing model frameworks, and thus serving as strong baselines for comparison with our method, as shown in Fig. 1.

The first approach treats sparse ego images as conditioning signals and trains a conditional generative model directly. For example, diffusion-based text/scene-to-motion models such as [30, 25] can be adapted by replacing text or audio inputs with ego-image features. However, this strategy requires large-scale paired data, is vulnerable to temporal inconsistency in high-dimensional sequence generation, and lacks explicit structural constraints to ensure accurate alignment with the desired start and end states. The second approach follows a two-stage pipeline: first estimates 3D endpoint poses from input images [34], and then uses a motion in-betweening model [28, 25] to generate

intermediate motion. While modular and intuitive, this design suffers from error accumulation, as inaccuracies in endpoint estimation can propagate and cause trajectory drift. It also introduces higher structural complexity, model size, and inference cost than an end-to-end formulation.

Our core motivation for this task is twofold. First, sparse endpoint ego observations make the problem highly ill-posed, while self-occlusion and limited body visibility further increase endpoint pose uncertainty. Thus, strong pose priors are needed to constrain the solution space beyond explicit conditioning. Second, the generated intermediate motion must jointly satisfy endpoint consistency, temporal continuity and plausibility, avoiding discontinuities or unstable body structures.

To address these challenges, we propose a unified bridge diffusion-based generative model trained in two stages, enabling a single framework to internalize both endpoint human pose estimation and physically consistent 3D motion generation, as presented in Fig. 1. Our method has two key features: 1) we propose egocentric-to-canonical endpoint pretraining with a time-span scheduling strategy, which strengthens the model's endpoint pose priors and enables more accurate endpoint fitting and alignment; and 2) we introduce the first motion bridge diffusion model that diffuses from neighboring motion frames. This design preserves end-to-end structural consistency while promoting physically continuous and plausible motion generation. Overall, our method avoids the strong coupling difficulty and error accumulation inherent in prior approaches.

To rigorously validate the effectiveness of our approach, we implement the two aforementioned strategies as strong baselines and conduct systematic ablation studies on all major components of our framework. Experiments are performed on the EE4D-Motion dataset [25], a large-scale motion capture benchmark comprising 208 hours of high-quality mocap data. Extensive quantitative and qualitative evaluations consistently demonstrate that our method achieves significant improvements over competing approaches across multiple metrics.

Our contributions can be summarized as follows:

- We introduce a new task, egocentric human motion bridging task, which aims to generate a coherent 3D motion segment from sparse egocentric image observations, addressing a practically relevant yet underexplored problem setting.
- We propose a unified bridge diffusion-based framework that jointly internalizes endpoint pose estimation and temporally continuous 3D motion completion through a two-stage training paradigm, enabling structurally consistent and physically plausible motion generation within a single model.
- We conduct comprehensive comparisons against strong baselines, including both direct conditional generation (one-stage) and estimation–interpolation pipelines (two-stage), and perform systematic ablation studies. Extensive quantitative and qualitative results demonstrate the clear improvements of our approach.

2 Related Work

Human Motion Task: Human Motion Modeling has gradually become a prominent research direction in computer vision and computer graphics. On the one hand, human motion exhibits strong nonlinearity, complex temporal dependencies, and multimodal characteristics, which pose significant challenges for modeling. On the other hand, it has broad application value in animation synthesis, virtual reality, human–computer interaction, robotic control, and behavior understanding.

Human motion tasks typically focus on generating or predicting plausible sequences of human joint movements under given conditions. The core challenge lies in modeling high-dimensional, structured, and strongly constrained spatiotemporal data. Depending on the task formulation, existing research can be broadly categorized into motion prediction [20, 2, 33, 29], motion generation [11, 10, 21, 35, 31, 27, 16], motion interpolation [12, 28, 6, 15]. Besides, motion generation tasks can be categorized according to the input modality, with common settings including text-to-motion [10, 11, 37, 31], audio-to-motion [5, 4, 3], and scene-to-motion [32, 25].

In this work, we focus on the growing interest in leveraging first-person perspectives to enhance perception for embodied intelligence and VR/AR applications, enabling the generation of human motion. To this end, we introduce a new task termed Egocentric Human Motion Bridging. Specifi-

cally, the goal is to directly generate intermediate motion sequences from sparse egocentric image observations.

Bridge Diffusion Models: The Bridge Diffusion Models (BDMs) [17, 36, 23, 22] are a recently proposed variant of diffusion models that incorporates endpoint constraints into the generative process by leveraging the Brownian bridge stochastic process. Unlike standard denoising diffusion probabilistic models (DDPM) [14], which learn to reverse a diffusion trajectory from pure Gaussian noise to the data distribution, BDM constructs a conditional stochastic path between two sources. This formulation explicitly enforces boundary consistency during generation.

So far, Bridge Diffusion has primarily been explored in image transformation and interpolation tasks. Brownian Bridge Diffusion Model (BBDM) [17] introduces a brownian bridge diffusion framework for image-to-image translation. By treating the source and target images as the two endpoints of a stochastic bridge, the model learns a conditional diffusion trajectory that smoothly transforms one image domain into another. Denoising Diffusion Bridge Models (DDBMs) [36] model distribution-to-distribution translation via diffusion bridges. By learning the bridge score and solving the associated stochastic differential equation, they offer a unified framework related to score-based diffusion and OT/flow matching, with improved results on image translation. TLB-VFI [22] applies the Brownian bridge formulation in latent space for video frame interpolation. In this setting, the first and last frames serve as fixed endpoints, and intermediate frames are generated along a time-aware stochastic bridge.

In this work, we are the first to extend Bridge Diffusion to human motion modeling. We resolve the incompatibility of Bridge Diffusion with multi-frame sequence data. We formulate human motion bridging as a cross-modal stochastic bridge that connects sparse egocentric image observations to a complete motion sequence. Unlike image translation or completion, our task requires inferring full-body intermediate motion trajectories from sparse visual inputs under strong structural and temporal constraints. To this end, we design a tailored adaptation of the bridge diffusion framework to generate temporally coherent, kinematically plausible human motion while respecting observational guidance.

3 Method

In this section, we first formulate our new egocentric human motion bridging task and then detail our technological innovations: 1) a short-to-long curriculum endpoint pretraining with progressively increasing temporal spans, which facilitates scaling the pose estimation capability from image-frame pairs by leveraging a larger amount of paired data; and 2) the first motion bridging diffusion model, which generates continuous motion by diffusing from neighboring frames rather than from pure noise, thereby enabling the recovery of intermediate motion from endpoint poses and improving the modeling of smooth temporal transitions.

3.1 Problem Formulation

Given an egocentric image sequence $I_{1:N} = (I_1, I_2, \ldots, I_N)$ acquired by a head-mounted camera, each frame satisfies $I_i \in \mathbb{R}^{H\times W\times 3}$. The corresponding 3D human motion is denoted by $X_{1:N} = (X_1, X_2, \ldots, X_N)$, where each pose X_i is represented with SMPL-X [26] parameters as $X_i = (R_i^r, t_i^r, \theta_i, \beta_i)$. Here, $R_i^r \in \mathbb{R}^3$ and $t_i^r \in \mathbb{R}^3$ denote the global root rotation and translation, $\theta_i \in \mathbb{R}^{21\times 3}$ contains the local joint rotations, and $\beta_i \in \mathbb{R}^{10}$ captures the body shape. Following standard practice, we treat body shape as time-invariant, so $\beta_i = \beta_j$ for all i, j.

Egocentric Human Motion Bridging Task aims to generate a plausible motion segment that connects two temporally sparse egocentric observations. Formally, the task is to model $p(X_{1:N} \mid I_s, I_e)$, where I_s and I_e denote the egocentric images captured at the start and end of the target interval. The output is a motion sequence $X_{1:N}$ that bridges the two observations, such that the generated motion is consistent with both endpoint images. The term *bridging* highlights precisely this property: the model must rely on egocentric visual cues, instead of explicit motion parameters, to bridge two motion endpoints and synthesize the missing motion segment. This task is particularly challenging because the input is highly ill-posed. Egocentric images only provide partial first-person observations of the surrounding scene, while the wearer's body is largely invisible. As a result, the model must infer the

latent human pose and dynamics from limited visual evidence, while ensuring temporal consistency with both endpoint observations.

3.2 Egocentric-to-Canonical Endpoint Pretraining

In the pretraining stage, we learn a reliable mapping from egocentric observations at the first and last frames to canonical 3D human poses, treating the task as egocentric motion estimation. By decomposing each motion sequence into multiple endpoint pairs with different temporal gaps, we increase the paired training data and reduce the ambiguity caused by limited visibility, self-occlusion, and depth uncertainty.

We pretrain the model with a conditional bridge diffusion objective on endpoint pose estimation. Specifically, given a motion window $X^{1:N}$, egocentric image features (O_s, O_e), we construct a pretraining sample by selecting a pair of subframes (X^s, X^e) with temporal gap $g = e - s$, where g is progressively scheduled from short to long values. The selected endpoint frames are placed at the two ends of a fixed-length window, yielding $(\tilde{X}_s, \tilde{X}_e)$. Only the corresponding conditioning inputs $(\tilde{O}_s, \tilde{O}_e)$ remain visible at the two endpoint positions, while all intermediate frames are masked out. The model is then trained to learn $p_\theta(\tilde{X}_{(s,e)} \mid \tilde{O}_{(s,e)})$ under the same bridge diffusion framework. In our implementation, the temporal gap follows a curriculum $g \in \{8, 16, 32, 64\}$ and progressively enlarge it until reaching the full-length setting, so the model first learns short-range endpoint consistency and then gradually handles larger pose variation across time. Importantly, this pretraining formulation allows each original motion window to be converted into multiple endpoint-pair supervision samples with different temporal spans, effectively increasing the amount of usable training data by several times. As a result, the model sees much richer ego-to-pose correspondences during training, which helps alleviate the ill-posed nature of egocentric 3D human pose and motion estimation. This provides a stronger initialization for the subsequent full-sequence modeling stage.

3.3 Human Motion Bridging Diffusion Finetuning

We first briefly review standard diffusion models and conventional bridge diffusion, and then introduce our motion bridging diffusion model for human motion generation from heterogeneous conditions.

Standard Diffusion. Denoising diffusion models learn a generative process by progressively corrupting clean data with Gaussian noise and then reversing this process. Given a clean sample x_0, the forward process is defined as

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t} x_0,\, (1 - \bar{\alpha}_t)\mathbf{I}\big)\,, \tag{1}$$

A neural network is then trained to recover the clean signal, or equivalently the injected noise, from the noisy sample x_t at timestep t. In the standard DDPM formulation, the model predicts the noise term ϵ, and generation starts from a Gaussian sample $x_T \sim \mathcal{N}(0, \mathbf{I})$ followed by iterative denoising.

Standard Bridge diffusion. Standard bridge diffusion defines a stochastic path between a source sample x_0 and a paired target sample y. Its forward process is

$$x_t = (1 - m_t)x_0 + m_t y + \sqrt{\delta_t}\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{2}$$

and the model is trained to predict the bridge residual $m_t(y - x_0) + \sqrt{\delta_t}\epsilon$. During sampling, it starts from the conditional endpoint $x_T = y$ and iteratively recovers x_{t-1} from x_t using the predicted residual and the reference y. Therefore, standard bridge diffusion is naturally suited to paired same-modality translation, where the target-space reference y is available throughout the reverse process.

Motivation. Inspired by bridge diffusion, we note that its bridging idea is particularly natural for human motion. A motion sequence is inherently transitional, with each frame evolving smoothly from nearby states, so modeling motion as a stochastic bridge provides a suitable inductive bias for temporal continuity and coherent kinematics.

Why Standard Bridge Diffusion is Insufficient for Motion Sequences? Although this formulation is natural for image-to-image translation, it is not directly suitable for human motion generation in

our setting. First, the original bridge objective depends explicitly on a same-modality reference y in the output space. In our task, however, the available conditions are heterogeneous, such as egocentric image observations and camera trajectory. Second, human motion is a structured temporal sequence rather than a single static sample, so the bridge should respect frame-wise temporal relations instead of treating the whole sequence as an unstructured paired object. **Solutions**: To make bridge diffusion suitable for motion sequences, we introduce two key modifications. First, instead of defining a global bridge between two fixed endpoints, we construct a local frame-wise bridge: splitting the sequence around its temporal center, each frame is paired with a nearby temporal neighbor, which is more natural for motion due to its local smoothness and short-range temporal dependency. Second, inspired by MDM, we predict the clean motion x_0 rather than the bridge noise, while still designing the forward process so that the terminal state approaches a standard Gaussian distribution. This allows direct supervision in motion space, which is more compatible with geometric losses, and also enables a unified generation process conditioned on heterogeneous modalities, without requiring a same-modality output reference as in standard bridge diffusion.

Motion Bridge Diffusion Model. As shown in 1, we formulate bridge diffusion directly in motion space while using external modalities only as conditions. Let $X^{1:N} = \{X^1, \dots, X^N\}$ denote a motion sequence, where X^i is the motion representation at frame i. We additionally define zero-padded boundaries $X^0 = X^{N+1} = 0$. For each frame i, we construct a local bridge pair

$$x_0^i = X^i, \qquad y^i = X^{i+s_i}, \tag{3}$$

where

$$s_i = 2 \cdot \mathbb{I}_{i>N/2} - 1. \tag{4}$$

This means that frames in the first half of the sequence are bridged toward their previous neighbors, while frames in the second half are bridged toward their next neighbors. The boundary terms are naturally handled by the zero padding. In this way, the bridge is defined between temporally adjacent motion states rather than between two global samples, which makes it more suitable for modeling motion dynamics.

For the forward process, based on this local bridge construction, we define the forward noising process as

$$x_t^i = \sqrt{\bar{\alpha}_t}\left((1-m_t)x_0^i + m_t y^i\right) + \sqrt{1-\bar{\alpha}_t}\,\epsilon^i, \qquad \epsilon^i \sim \mathcal{N}(0, \mathbf{I}), \tag{5}$$

where $m_t \in [0,1]$ is the bridge coefficient and $\bar{\alpha}_t$ follows the standard diffusion noise schedule. In our implementation, m_t increases linearly with t. This process can be interpreted as first defining a motion-dependent bridge mean between the current frame and its local temporal reference, and then applying standard Gaussian diffusion around this mean.

This design is important because it preserves a key property of conventional diffusion models: as t approaches the final timestep T, $\bar{\alpha}_t \to 0$, and thus

$$x_T^i \approx \epsilon^i \sim \mathcal{N}(0, \mathbf{I}). \tag{6}$$

Therefore, this allows sampling to start from a modality-agnostic Gaussian prior, which is crucial for unified generation from heterogeneous input instead of a human pose.

During sampling, the process starts from Gaussian noise. At each timestep t, the model predicts the clean motion $\hat{x}_0 = \epsilon_\theta(x_t, t, c)$ and uses it to compute the reverse transition to x_{t-1}. Repeating this process from $t = T$ to $t = 1$ gradually removes noise and yields the final motion sample x_0.

Algorithm 1: Motion Bridging Diffusion Model Training

repeat
- Sample motion sequence data $X^{1:N} \sim q(X^{1:N})$, with $X^0 = X^{N+1} = 0$;
- Sample a pair of egocentric images $(o^1, o^N) \sim q(O^1, O^N)$;
- Sample timestep $t \sim \mathrm{Uniform}(1, \dots, T)$;
- **for** $i = 1, \dots, N$ **do**
  - Define $s_i = 2 \cdot \mathbb{1}_{i>N/2} - 1$;
  - Define paired data $(x_0^i, y^i) = (X^i, X^{i+s_i})$;
  - Sample Gaussian noise $\varepsilon^i \sim \mathcal{N}(0, I)$;
  - Define condition $c^i = \mathbb{1}_{i=1} \cdot o^1 + \mathbb{1}_{i=N} \cdot o^N + 0$;
  - Forward diffusion: $x_t^i = \sqrt{\bar{\alpha}_t}\left((1-m_t)x_0^i + m_t y^i\right) + \sqrt{1-\bar{\alpha}_t}\,\varepsilon^i$;
- Take a gradient descent step on: $\nabla_\theta \sum_{i=1}^N \left\| x_0^i - \epsilon_\theta(x_t^i, t, c^i) \right\|^2$;

until *converged*;

4 Experiments

In this section, we present in detail the strong baselines constructed for the new egocentric human motion bridging task, as well as the technical implementation, evaluation metrics, and dataset details. These experimental designs are intended to demonstrate the unified formulation and effectiveness of our method on this task.

4.1 Implementation

We adopt a conditional bridge diffusion model to model the entire human motion sequence. The denoising network is based on a Transformer decoder architecture with 12 layers, 12 attention heads, and a latent dimension of 768. The input to the network consists of noisy motion tokens, while the conditioning information includes egocentric image features. The image features, together with the diffusion timestep embedding, form context tokens that are injected into the denoising process through cross-attention.

For the egocentric human motion bridging task, we retain only the conditioning information from the first and last frames, while masking out all trajectory and image conditions for the intermediate frames. This forces the model to recover the complete intermediate motion based solely on the endpoint constraints. The training objective is to predict the clean motion representation itself, and an MSE loss is computed over the valid frames. We adopt a head-centric motion representation, rather than the pelvis-centric representation commonly used in motion synthesis literature, which makes it better suited to egocentric devices. We use DINOv2 [24] fine-grained features to extract representations from egocentric images, since they have been shown to be more effective than CLIP in [25].

4.2 Dataset

We use EE4D-Motion [25], a version of the large-scale EgoExo4D [8] dataset annotated with SMPL-X [1]. It is a 3D human motion dataset comprising 208 hours of time-synchronized 3D motion data, egocentric videos, and associated annotations. The dataset captures real-world human activities performed by participants wearing Project Aria glasses, covering diverse scenarios such as cooking, dance, sports, music, healthcare, and bike repair. To construct the training set, we extract 8-second video clips sampled at 10 fps ($\mathrm{N} = 80$) with a sliding window of 2 seconds, resulting in 143K training samples. For validation, we use the same clip length and frame rate but sample clips every 20 seconds, yielding 4,400 validation samples.

4.3 Metrics

In the egocentric human motion bridging task, the goal is to generate the intermediate motion while ensuring that the poses at the start and end points are as aligned as possible with the ground truth. Accordingly, our evaluation is divided into two parts: estimation metrics for the endpoint frames and generation metrics for the intermediate motion.

Table 1: **Model Efficiency Comparison**: Comparison of model complexity and inference latency for generating a 1-second clip using 1000 diffusion steps on an NVIDIA A100 GPU.

Method	Params	Latency ↓
MDM [31]	87M	1.66s
UEM [18]	89M	1.82s
EgoAllo [18]+UEM	140M	3.21s
Ours	**86M**	**1.58s**

We evaluate endpoint pose estimation quality using a diverse set of metrics in the 3D geometric space. **MPJPE** measures the mean per-joint position error (in meters) over 22 body joints. **MPJPE-PA** further aligns the predicted and ground-truth motions on a per-frame basis using Procrustes analysis before computing MPJPE, thereby reflecting the accuracy of local pose estimation. **MPJPE-B** reports the mean per-joint position error (in meters) for the body joints. **MPJPE-H** reports the mean per-joint position error (in meters) for the hand joints.

For the motion generation part, we consider contact-related motion quality metrics. **Foot Sliding** [11] measures the amount of undesired foot motion when the foot is close to the ground. **Foot Contact** [11] computes the average distance (in meters) between the foot and the ground, capturing both floating artifacts and ground penetration. In addition, we adopt semantic-level motion evaluation metrics.

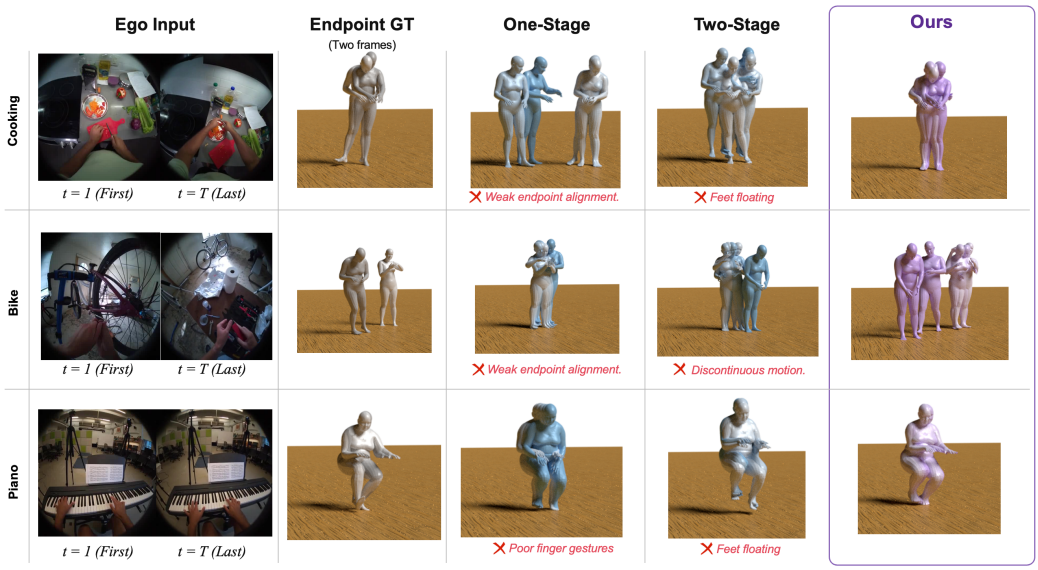

Figure 2: **Qualitative Comparison with Strong Baselines.** The first two columns show the input start (bright white) / end (dark gray) egocentric images and the corresponding ground-truth endpoint poses. The next two columns present the one-stage baseline (UEM) and the two-stage baseline (EgoAllo+UEM), respectively, followed by our method in the last column. Further comparisons demonstrating physical continuity are available in the **supplementary material video**.

Table 2: **Endpoint Results**: Comparison with strong baselines for endpoint pose estimation.

Method	MPJPE(↓)	MPJPE-PA(↓)	MPJPE-B(↓)	MPJPE-H(↓)
MDM [31]	0.742	0.286	0.701	0.865
Genmo [19]	0.524	0.213	0.537	0.646
UEM [25]	0.486	0.162	0.463	0.579
EgoAllo [34] + UEM	0.421	0.137	0.398	0.512
Ours	**0.327**	**0.091**	**0.309**	**0.409**

Semantic Similarity measures the consistency between generated motions and ground-truth motions in a semantic latent space, analogous to CLIP-score [13]. Concretely, we use the motion encoder from TMR [9] to project motions into the latent space and compute the cosine distance between their embeddings. **FID** measures the distributional gap between generated and real motions in the same latent space, similar to the standard FID used in image generation. **Motion Jitter** measures the average variation in motion within a fixed-length temporal window; lower values indicate smoother motion.

4.4 Baselines

Since there are no direct baselines for the egocentric human motion bridging task, we construct strong baselines based on state-of-the-art motion modeling paradigms. Specifically, we compare our method against two categories of strong baselines: one-stage methods (MDM [31], UniEgoMotion (UEM) [25], Genmo [19]) and two-stage methods (EgoAllo [34] with UniEgoMotion [25]).

One-stage methods directly model motion generation in a conditional manner, whereas two-stage methods first estimate the 3D poses at the start and end frames and then feed them into an in-betweening model to predict the intermediate motion sequence. For a fair comparison, we retrain all baseline methods on the EE4D-Motion [25, 8] dataset and keep the same input and output conditions to adapt them to our proposed human motion bridging task.

Table 3: **Motion Generation Results**: Comparison with strong baselines on EE4D-Motion for motion generation quality and diversity.

Method	Foot Slide(↓)	Foot Contact(↓)	Semantic Sim.(↑)	FID(↓)	Jitter(↓)	Diversity
MDM [31]	7.318	0.087	0.049	0.132	0.419	9.325
Genmo [19]	5.210	0.053	0.076	0.087	0.378	9.274
UEM [25]	4.912	0.049	0.074	0.071	0.352	9.457
EgoAllo [34] + UEM	4.865	0.047	0.075	0.069	0.347	9.411
Ours	**3.446**	**0.031**	**0.082**	**0.047**	**0.176**	9.403

Table 4: **Cross-dataset Generalization on Nymeria [7].** Models trained on EE4D-Motion are evaluated directly on Nymeria without training or fine-tuning on Nymeria.

Method	Endpoint MPJPE(↓)	Foot Slide(↓)	Foot Contact(↓)	Semantic Sim.(↑)	FID(↓)	Jitter(↓)	Diversity
MDM [31]	0.875	6.452	0.104	0.061	0.164	0.537	9.347
Genmo [19]	0.643	4.895	0.081	0.071	0.126	0.418	9.445
UEM [25]	0.530	3.875	0.062	0.072	0.114	0.352	9.576
EgoAllo [34] + UEM	0.481	3.676	0.053	0.076	0.093	0.324	9.501
Ours	**0.369**	**2.748**	**0.023**	**0.086**	**0.059**	**0.153**	9.533

4.5 Results

Quantitative Results: Tables 2 and 3 compare our method with one-stage and two-stage baselines on endpoint estimation and motion generation, respectively. MDM, Genmo, and UEM directly generate motion in a one-stage manner, while EgoAllo+UEM adopts a two-stage pipeline that first estimates endpoint poses and then performs motion generation. Our conditional unified motion bridge diffusion model achieves the best endpoint accuracy and motion quality scores while maintaining comparable Diversity (9.403). In particular, the lower endpoint MPJPE-based errors indicate more accurate endpoint pose estimation, while the improved Foot Slide, Foot Contact, Semantic Similarity, FID, and Jitter scores demonstrate more plausible, temporally stable motion generation. Besides, Table 1 demonstrates that our unified bridge diffusion design achieves better model inference performance compared with other models.

We further evaluate cross-dataset generalization by directly applying models trained on EE4D-Motion to Nymeria, without training or fine-tuning on the target dataset (Table 4). Our method achieves the lowest endpoint MPJPE (0.369), FID (0.059), and Jitter (0.153), together with the best foot-related metrics and Semantic Similarity. Diversity remains comparable to the baselines at 9.533. These zero-shot results support the transfer of our endpoint estimation and motion generation capabilities beyond the training dataset.

Qualitative Results: Figure 2 compares motion sequences generated from the same start and end egocentric images. The one-stage baseline exhibits weak endpoint alignment and implausible hand configurations, while the two-stage baseline improves the global structure but still produces physical artifacts such as floating feet. Our method better aligns the endpoints with the reference poses and preserves more plausible body configurations and contact patterns throughout the sequence. These observations agree with the quantitative results in Tables 2 and 3, supporting the benefit of jointly modeling endpoint estimation and motion bridging. Additional comparisons of temporal continuity are provided in the supplementary video.

4.6 Ablation Study

Table 5 presents a controlled 2×2 comparison of standard diffusion and bridge diffusion, each with and without E2C endpoint pretraining. All variants use the same architecture and experimental protocol, allowing us to examine the two components separately.

Effect of E2C Pretraining. E2C reduces endpoint MPJPE from 0.468 to 0.362 for standard diffusion and from 0.431 to 0.327 for bridge diffusion. Its benefit therefore holds under both diffusion formulations. Figure 3 provides complementary visual evidence: compared with the variant without E2C, the pretrained model produces endpoint body configurations that more closely match the ground

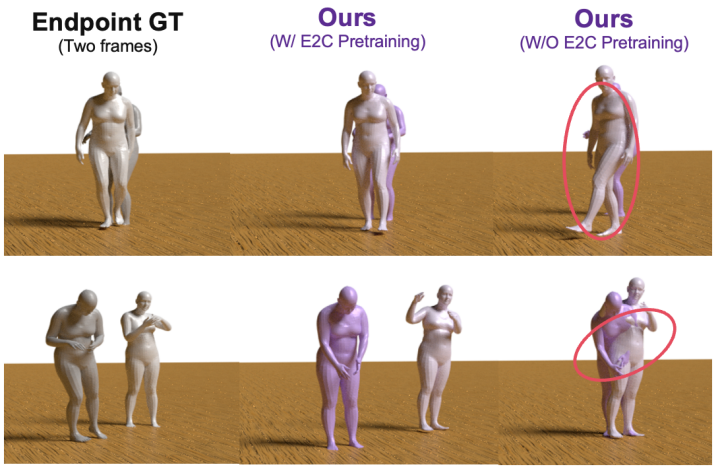

Figure 3: Visual Comparison of E2C Pretraining Ablation.

Table 5: **Ablation Study.** A controlled comparison of standard diffusion and bridge diffusion with and without E2C endpoint pretraining. Endpoint MPJPE evaluates endpoint-pose accuracy; the remaining metrics evaluate motion generation quality.

Method	Endpoint MPJPE(↓)	Foot Slide(↓)	Foot Contact(↓)	Semantic Sim.(↑)	FID(↓)	Jitter(↓)
Standard Diffusion w/o E2C	0.468	4.921	0.058	0.052	0.097	0.325
Bridge Diffusion w/o E2C	0.431	3.684	0.038	0.074	0.056	0.184
Standard Diffusion w/ E2C	0.362	4.847	0.056	0.057	0.094	0.308
Bridge Diffusion w/ E2C	**0.327**	**3.446**	**0.031**	**0.082**	**0.047**	**0.176**

truth. The highlighted discrepancies without pretraining illustrate the ambiguity of recovering body poses from egocentric views, in which much of the body is unobserved. E2C also improves all reported motion generation metrics under both diffusion formulations.

Effect of Bridge Diffusion. With E2C fixed, replacing standard diffusion with bridge diffusion reduces Foot Slide from 4.847 to 3.446, FID from 0.094 to 0.047, and Jitter from 0.308 to 0.176, while increasing Semantic Similarity from 0.057 to 0.082. The same trend holds without E2C, indicating that the generation improvements are not solely attributable to endpoint pretraining. Together, Figure 3 and Table 5 show complementary benefits: E2C strengthens endpoint alignment, while bridge diffusion improves motion quality and temporal smoothness. Combining both yields the best results across all metrics reported in the ablation.

5 Conclusion and Limitation

In this paper, we introduced Egocentric Human Motion Bridging, a new task that generates coherent 3D human motion segments from sparse egocentric endpoint observations. To address the challenges of partial body visibility, endpoint ambiguity, and temporal ill-posedness, we proposed Ego-HMB, a unified bridge diffusion framework that jointly internalizes endpoint pose estimation and temporally consistent motion completion through a two-stage training strategy. Extensive experiments demonstrate that our method outperforms both direct conditional generation and two-stage estimation–interpolation baselines, producing more accurate, continuous, and plausible motions. A current limitation of this work is that we focus on human motion bridging from egocentric images, which are commonly available in VR/AR and robotics scenarios, while other modalities such as text and audio are not yet explored. Nevertheless, our framework can be naturally extended to incorporate these additional modalities as conditioning signals. Another promising direction is to further adapt the generated motion representations for direct humanoid robot motion control, enabling broader embodied applications.

Acknowledgments and Disclosure of Funding

Mingming Gong was supported by ARC DP240102088 and WIS–MBZUAI 142571.

References

[1] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016.

[2] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9544–9555, 2023.

[3] Yongkang Cheng and Shaoli Huang. Holegest: Decoupled diffusion and motion priors for generating holisticly expressive co-speech gestures. In *2025 International Conference on 3D Vision (3DV)*, pages 748–757. IEEE, 2025.

[4] Yongkang Cheng, Shaoli Huang, Xuelin Chen, Jifeng Ning, and Mingming Gong. Didiffges: Decoupled semi-implicit diffusion models for real-time gesture generation from speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2464–2472, 2025.

[5] Yongkang Cheng, Mingjiang Liang, Shaoli Huang, Gaoge Han, Jifeng Ning, and Wei Liu. Conditional gan for enhancing diffusion models in efficient and authentic global gesture generation from audios. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2164–2173. IEEE, 2025.

[6] Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH 2024 conference papers*, pages 1–9, 2024.

[7] Ma et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *ECCV*, 2024.

[8] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[9] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022.

[10] Gaoge Han, Shaoli Huang, Mingming Gong, and Jinglei Tang. Hutumotion: Human-tuned navigation of latent motion diffusion models with minimal feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2031–2039, 2024.

[11] Gaoge Han, Mingjiang Liang, Jinglei Tang, Yongkang Cheng, Wei Liu, and Shaoli Huang. Reindiffuse: Crafting physically plausible motions with reinforced diffusion model. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2218–2227. IEEE, 2025.

[12] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

[13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[15] Inwoo Hwang, Bing Zhou, Young Min Kim, Jian Wang, and Chuan Guo. Scenemi: Motion in-betweening for modeling human-scene interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6034–6045, 2025.

[16] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023.

[17] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 1952–1961, 2023.

[18] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023.

[19] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11766–11776, 2025.

[20] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks for 3d human motion prediction. In *European conference on computer vision*, pages 18–36. Springer, 2022.

[21] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023.

[22] Zonglin Lyu and Chen Chen. Tlb-vfi: Temporal-aware latent brownian bridge diffusion for video frame interpolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16260–16269, 2025.

[23] Zonglin Lyu, Ming Li, Jianbo Jiao, and Chen Chen. Frame interpolation with consecutive brownian bridge diffusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3449–3458, 2024.

[24] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[25] Chaitanya Patel, Hiroki Nakamura, Yuta Kyuragi, Kazuki Kozuka, Juan Carlos Niebles, and Ehsan Adeli. Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10318–10329, October 2025.

[26] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[27] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022.

[28] Jia Qin, Youyi Zheng, and Kun Zhou. Motion in-betweening via two-stage transformers. *ACM Trans. Graph.*, 41(6):184–1, 2022.

[29] Jianwei Tang, Hong Yang, Tengyue Chen, and Jian-Fang Hu. Stochastic human motion prediction with memory of action transition and action characteristic. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1883–1893, 2025.

[30] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[31] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[32] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12206–12215, 2021.

[33] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Xinchao Wang, and Yanfeng Wang. Auxiliary tasks benefit 3d skeleton-based human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9509–9520, 2023.

[34] Brent Yi, Vickie Ye, Maya Zheng, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. *arXiv preprint arXiv:2410.03665*, 2024.

[35] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024.

[36] Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *The Twelfth International Conference on Learning Representations*, 2024.

[37] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pages 18–38. Springer, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction largely reflect the paper's contributions and scope. The claims are aligned with the proposed method and the reported experimental results. However, the authors should ensure that any broad statements about generality or superiority are clearly qualified, especially where the evidence is limited to the evaluated datasets or experimental settings.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper mentions some limitations of the work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theoretical results, formal assumptions, or theorem-style claims.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper discloses the key information needed to reproduce the main experimental results, including the experimental setup, datasets, baselines, evaluation metrics, and implementation details relevant to the main claims and conclusions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper currently does not provide open access to the data and code. However, the authors plan to consider releasing the code and relevant data after acceptance, along with sufficient instructions to help reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the main training and test details necessary to understand the results, including the data splits, evaluation protocol, hyperparameters, and optimizer settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports appropriate statistical information where applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resources used for the experiments, including the type of compute workers, memory requirements, and execution time where relevant.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: There is no societal impact of the work performed

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: The paper does not use existing assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.