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

Gaoge Han Yongkang Cheng Shaoli Huang Jing Zhang Mingming Gong Fakhri Karray Tongliang Liu

01 Egocentric human motion bridging

- **Motivation.** Recover motion in VR/AR and robotics without reliable 3D poses or motion capture.
- **Task.** Generate a complete SMPL-X motion sequence from only two endpoint egocentric images.
- **Constraints.** Match both endpoints and preserve continuity despite partial body visibility and self-occlusion.

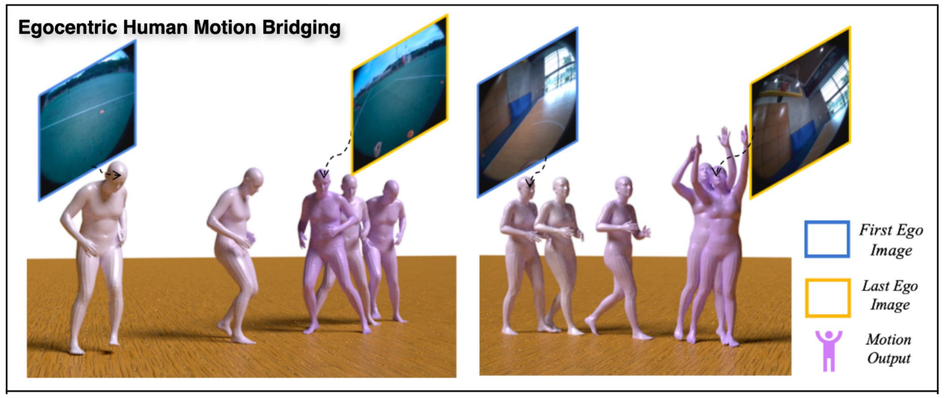


Figure 1 (top). Sparse endpoint images condition the complete human motion segment.

02 A unified model with two training stages

1 E2C pretraining

- **Pose priors.** Learn ego-image-to-canonical-pose correspondences.
- **Curriculum.** Increase temporal gaps through 8, 16, 32 and 64 frames.

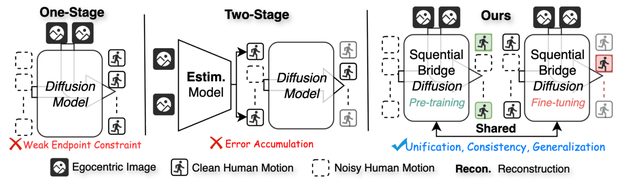


Figure 1 (bottom). Unified training avoids the error accumulation of estimation–interpolation pipelines.

- **Unified generation.** One model internalizes endpoint estimation and motion completion.

2 Motion bridge diffusion

- **Local pairing.** Use previous neighbors in the first half and next neighbors in the second half.

$$x_0^i = X^i, \qquad y^i = X^{i+s_i}, \qquad s_i = 2 \cdot \mathbb{I}_{i>N/2} - 1.$$

- **Forward process.** Diffuse around the mean of adjacent motion states.

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

- **Sampling.** Start from Gaussian noise.
- **Denoising.** Predict clean motion using DINOv2 endpoint features.

03 Evaluation on EE4D-Motion

143K training clips 4,400 validation clips 80 frames / 8 seconds

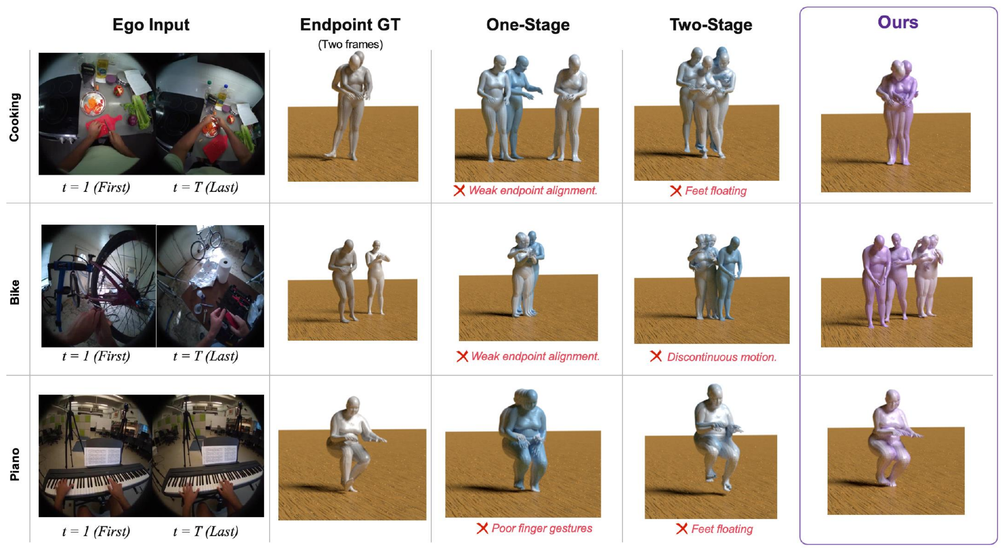


Figure 2. Endpoint alignment and motion quality across cooking, bike repair and piano.

Table 2. 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**

Table 3. Motion generation quality

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

22.3%
lower endpoint MPJPE
Relative to EgoAllo + UEM

49.3%
lower motion jitter

Table 5. Component ablation

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**

- **E2C** strengthens endpoint pose priors.
- **Motion bridge** improves generation quality and temporal smoothness.
- **Together** the components yield the best results across all ablation metrics.

86M parameters
1.58 s per 1-second motion clip
1,000 diffusion steps on an NVIDIA A100