HandWorld: Hand-Centric Unified Video Action Generation

Supplementary Material

A. Training

A.1. Training Strategy

As described in the main paper, all training objectives in HandWorld can be expressed as special cases of the unified loss in Equation 5 under different condition configurations $\mathbf{C}_t$ and target states x_1. In practice, we adopt a three-stage training strategy to stabilize optimization and progressively couple the video and action domains.

In the first stage, we train the video DiT together with the shared condition network, while keeping the action conditions fully observed and using only the video generation objective. Concretely, all action tokens are visible to the condition network, whereas the video sequence is truncated to only the first frame as input, and the model is trained to generate the remaining frames using the video loss $\mathcal{L}_{\text{video}}$. This stage encourages the condition network to extract useful temporal context from the action domain and adapt the video DiT to egocentric HOI data under relatively simple conditions.

In the second stage, we train the action DiT and the shared condition network, now using the action loss $\mathcal{L}_{\text{action}}$ while keeping the video conditions fully observed. All video frames are provided as input to the condition network, and only the first quarter of the action sequence (13 frames) is retained as observed action history. The remaining action tokens are predicted by the action DiT. This stage focuses on learning accurate action dynamics conditioned on rich visual context and further refines the cross-domain representation.

In the third stage, we freeze both the video and action DiTs and train only the shared condition network under a mixed multi-task setting. At each iteration, we randomly sample one of three tasks: egocentric video generation, action forecasting, or joint prediction of both domains under partial observations. For the selected task, we apply the corresponding condition configuration and jointly optimize both $\mathcal{L}_{\text{video}}$ and $\mathcal{L}_{\text{action}}$. In addition to the task-specific condition pattern, we introduce an additional random masking strategy. We mask the currently observed condition tokens with probability 0.2 and replace them with learnable mask tokens. This encourages the condition network to robustly infer temporal context and learn more generalizable cross-domain condition signals.

For optimization, we use AdamW with a learning rate of $1e^{-4}$ in the first two stages and $2e^{-5}$ in the third stage. Each stage is trained for $20K$ steps. We adopt a global batch size of 16 (batch size 1 per GPU) using the DeepSpeed framework for memory-efficient distributed training.

Table 4. Effect of denoising steps. τ represents the number of denoising steps. K represents the number of samples.

Steps	Avg Distance		Final Distance	
	$K=1$	$K=5$	$K=1$	$K=5$
$\tau=10$	0.055	0.044	0.063	0.052
$\tau=16$	0.047	0.041	0.049	0.048
$\tau=25$	0.040	0.041	0.053	0.046
$\tau=50$	0.044	0.039	0.051	0.045
$\tau=100$	0.046	0.038	0.051	0.046

B. Evaluation

B.1. Hand Action Prediction

We analyze the impact of different denoising steps. We report both the average and final distance metrics following a best-of-K evaluation protocol. The results in Table 4 indicate that our model maintains competitive performance even with only 16 sampling steps. Benefiting from our decoupled DiTs design, the action branch can be executed with significantly fewer denoising steps when video generation is not required, enabling a notably faster action prediction pipeline.

B.2. Hand-Centric Video Generation

Hand-Related Metrics. To evaluate the visual quality of hand-object interaction in generated videos, we compute a hand-region CLIP score ($\text{CLIP}_{\text{hand}}$) that measures semantic consistency between hand regions in the generated and ground-truth videos. We first apply the existing hand detector [25] to identify hand bounding boxes in the ground-truth videos. Each detected region is then expanded to ensure that both the hand and the interacted object fall within the crop. Let r_t^{gt} and r_t^{pred} denote the cropped regions from the ground-truth and generated frames at time t, respectively. The CLIP encoder $\phi(\cdot)$ is used to extract image embeddings, and the similarity for each frame is computed as the cosine similarity. The final hand-region CLIP score averages this similarity across all frames in the sequence:

$$\text{CLIP}_{\text{hand}} = \frac{1}{T}\sum_{t=1}^{T} \cos\left(\phi(r_t^{\text{gt}}), \phi(r_t^{\text{pred}})\right). \tag{6}$$

This metric reflects how well the generated sequence preserves the semantics of hand appearance and interaction at the action-critical regions. We also compute the IoU score between detected hand regions in the generated and ground-truth videos to assess the accuracy of the generated hand po-

sitions. We use the same hand detector to obtain bounding boxes b_t^{gt} and b_t^{pred} for each frame, and the sequence-level IoU score is obtained by averaging across all frames:

$$\text{IoU} = \frac{1}{T} \sum_{t=1}^{T} \frac{b_t^{\text{gt}} \cap b_t^{\text{pred}}}{b_t^{\text{gt}} \cup b_t^{\text{pred}}}. \tag{7}$$

A higher IoU indicates more effective and accurate control of hand actions, as well as better temporal consistency in the generated hand trajectories. Although some baseline methods do not use hand actions as explicit control signals, we still report their IoU scores for reference.