

RECOGNITION-BY-IMAGINATION: A GENERATIVE APPROACH TO FORM-FROM-MOTION

Anonymous authors

Paper under double-blind review

ABSTRACT

We study the problem of recognizing actions from a minimalist motion stimuli with no appearance cues. We frame this as an instantiation of Johansson’s famous experiments in 1973 on recognizing actions from a video showing a handful of moving keypoints against a black background. While this collection of keypoints seems to be spread arbitrarily in a single frame, their motion enables action perception. We establish that modern MLLMs are not able to represent such signals. The objective of this paper is then to enable video models to be sensitive to such subtle motion stimuli. To perceive this minimalist motion and recognize actions, we propose a novel two-step *Recognition by Imagination* (RBI) approach: (i) *imagination*: we fine-tune a video generation model to generate a realistic video solely from the keypoint stimuli, (ii) *recognition*: then, we employ a standard MLLM to classify the action. RBI significantly outperforms frontier MLLMs and shows superior generalization compared to supervised fine-tuning of open MLLMs. Interestingly, while RBI is fine-tuned on keypoint videos, it exhibits some surprising emergent capabilities. It works off-the-shelf with novel stimuli such as silhouettes, shape outlines, and even event-based camera inputs. Finally, we also show that the same model, fine-tuned only on keypoint videos, qualitatively generalizes to other psychophysical form-from-motion stimuli.

1 INTRODUCTION

The visual world around us is dynamic. Movement of other humans is one of the most predominant motion we observe through our lifespan. Naturally, human motion has been of keen interest to psychophysicists and neurologists. This interest was particularly accentuated by the pioneering experiments of Gunnar Johansson in 1973. He showed that humans can effortlessly recognize actions even when the kinematic motion patterns are represented by only a handful of light points attached to the major body joints (Johansson, 1973; Grossman et al., 2000)¹. Individual static frames may look like a meaningless group of points, yet when the sequence moves over time, we instantly perceive it as a human activity. This ability is regarded as *biological motion perception* since it goes beyond human motion to perceiving other animal motion too. Johansson’s keypoints represent a minimal *gist of motion*, disentangling it from other static cues like appearance, objects and background.

Multimodal LLMs (MLLMs) trained on web-scale video data are very strong at recognizing or localizing actions (Bai et al., 2025). But how much of this ability stems from static appearance cues (e.g., single-frame or an unordered set of frames) vs. temporal motion cues? Moreover, if we gradually strip off appearance cues, what is the minimal representation of motion perceptible to MLLMs? Evaluation analogous to Johansson’s original experiments for modern video models presents a unique opportunity to investigate these questions. Furthermore, as the human experiments show, there is sufficient signal to infer the action despite the sparsity in keypoints. The challenge, then, is: what computational approach is capable of extracting this signal in a generalizable manner?

To measure progress systematically, we introduce a benchmark, *JohanssonBench* to quantify action perception on Johansson’s keypoint (JKP) stimuli constructed from two existing human action datasets (Shahrourdy et al., 2016; Chen et al., 2015). We use a language-based evaluation since semantic actions are well defined in language, and, it allows us to test generalization robustly. The

¹We encourage the reader to view examples in the Supplementary video.

evaluation is divided into three splits to measure (i) cross-instance (train-test share the same actions), (ii) cross-action (actions in test are not in train), and (iii) cross-dataset generalization. We show that the best frontier MLLMs are around the halfway mark compared to a human baseline. While supervised fine-tuning open MLLMs fills this gap on the cross-instance split, it fails to generalize to unseen actions or a new dataset pointing to memorization rather than true grasp of motion.

To build a human-like, generalizable model for JKP perception, we take inspiration from neurology research on how human vision achieves this. Saygin et al. (2004) showed that the human motor system (which is responsible for controlling one’s own movement) gets activated while perceiving JKP. This suggests that observers may recognize an action, in part, by internally simulating how they would perform it. Inspired by this, we propose an analysis-by-synthesis approach. We fine-tune a video generation model to map a JKP stimulus to the RGB space (*Imagination*) and then employ a (frozen) MLLM to classify the action (*Recognition*). We call this approach as *Recognition by Imagination* (RBI). Specifically, we fine-tune LTX-2 19B on $\sim 14\text{K}$ JKP-RGB pairs from NTU-60. Once trained, we can attach any downstream MLLM for action inference. RBI with standard open MLLMs beats SFT across all splits demonstrating remarkable generalization. In fact, RBI with InternVL3-8B even matches the human baseline. We also conduct qualitative evaluation showing RBI’s generalization to: (i) challenging poses (e.g., sport actions), (ii) number of persons, (iii) occlusion, (iv) number of keypoints, (v) camera depth.

Along with strong generalization, we also discover some interesting emergent abilities of the video generation model fine-tuned on JKP stimuli. (i) The model works off-the-shelf even on stimuli slightly or considerably different from JKP, e.g., using keypoints + joints, or body outlines, or silhouettes, etc. This hints to JKP being a robust, minimal representation capturing the *gist of motion*. With this ability, we show applications to co-speech gesture generation without any additional training. (ii) The same model also works on Dynamic Vision Sensor (DVS) videos recorded on event-based cameras. This opens up privacy-preserving action recognition without the need for specialized architectures for DVS videos – instead one can use standard vision models on synthetic videos generated by RBI. (iii) Finally, RBI also shows surprising generalization to other challenging temporal stimuli, e.g., shape motion captcha examples from SpookyBench (Upadhyay et al., 2026), or the Hidden Bird illusion (Regan, 1986; Bach, 2006).

2 DATASET: JOHANSSONBENCH

In this Section, we introduce our benchmark based on Johansson’s biological motion perception (Johansson, 1973). We refer to this stimuli as Johansson’s Keypoints (JKP) videos for short. First, we describe the construction of JKP stimuli from RGB videos. Then, we describe the datasets used to create the `JohanssonBench` benchmark.

Constructing JKP stimuli. To replicate the setup of Johansson (1973), we use 2D pose detectors to construct the JKP stimuli. For a given RGB video, we first use DWPose (Yang et al., 2023) to extract pose per frame. We render the pose sequence but remove the stick-figure lines connecting joints, finger/face keypoints and re-color all remaining keypoints white on a black background. As shown in Fig. 1, this yields video stimuli that resemble the ones in Johansson (1973). For consistent evaluation across datasets, we preprocess the raw videos as follows: (i) we use a person detector and compute a maximal spatial crop such that the person occupies $\geq 60\%$ of each frame, (ii) we resize the shorter side to 480 px and pad the other to 832 px. More details are in Appendix Section A.1.

Formulating the benchmark. In line with the original Johansson’s experiments, we need datasets consisting of simple human actions (e.g., walking, throwing, etc.) with no confounding complexities (e.g., camera motion, background noise). Using this criteria, we find two datasets suitable for this task: **NTU** (Shahroudy et al., 2016) and **UTD-MHAD** (Chen et al., 2015). Videos from these datasets will be used in two ways: (i) as raw videos to establish upper bound performance for MLLMs trained on web videos, (ii) to construct JKP videos using the procedure described above. We construct the following splits based on these datasets.

- **NTU Cross-Instance (NTU-CI):** The NTU-60 (Shahroudy et al., 2016) dataset was originally proposed for 3D human action recognition. It consists of 60 actions involving one or two actors in a lab setting. Since the camera is static and the actions are well-defined, this serves as an appropriate base dataset to study Johansson’s experiments. We manually pick a subset of $C=21$

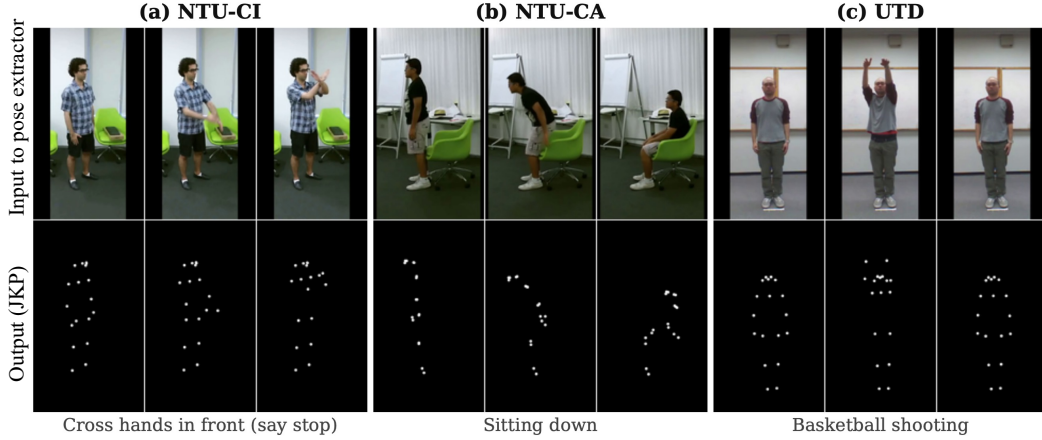


Figure 1: **Constructing Johansson’s keypoint videos.** We use DWPose to extract keypoints from RGB videos and construct corresponding videos that resemble Johansson’s experiments on humans in 1973. Here, we show an example video from each of the three evaluation datasets.

actions that (i) have clearly perceptible motion (e.g., we do not pick “sniff/smell” or “typing on a keyboard” as it is too subtle), (ii) human-centric and not object-biased (e.g., we do not pick “tear paper”, as it is ambiguous without the paper). As for the evaluation set, we randomly sample $N_c=25$ examples for each class c , amounting to a total of $N_{\text{test}}=525$ videos. In this cross-instance split, we use the same set of actions in the train-test sets while following the original data split across videos. The train set has a total of $N_{\text{train}}=14,063$ videos.

- **NTU Cross-Action (NTU-CA):** In order to test generalization to novel actions, we further split the action classes in NTU-CI such that $C_{\text{train}}=11$ and $C_{\text{test}}=10$. Thus, the test set consists of $N_{\text{test}}=250$ videos and the train set consists of $N_{\text{train}}=7,344$ videos with no overlap in the actions.
- **UTD:** is used as an unseen, hard test set. It consists of $C_{\text{test}}=27$ single-person actions performed by 8 actors amounting to a total of $N_{\text{test}}=861$ videos. It is substantially more challenging than NTU because it has temporally sensitive action pairs such as “right hand draw circle (clockwise)” vs. “right hand draw circle (counter clockwise)”.

The full list of actions in each of the datasets is provided in Appendix Section A.3. Some basic statistics for each of the three datasets are provided in Table 1.

Table 1: **Basic data statistics** of the three evaluation splits used in JohanssonBench. NTU-CI and NTU-CA denote the NTU cross-instance and cross-action splits, respectively; UTD denotes the cross-dataset UTD split. Durations are computed from the RGB benchmark videos.

Dataset split	Train videos	Test videos	Classes	Avg. Dur. (s)	FPS	Resolution
NTU-CI (ntu525)	14,063	525	21	2.39	30	480×832
NTU-CA (ntu250)	7,344	250	10	2.45	30	480×832
UTD (utd)	-	861	27	3.31	15	480×832

3 RECOGNITION BY IMAGINATION

To build a truly robust computational model capable of perceiving Johansson’s stimuli, we revisit how human vision solves this task. The predominant psychology research has modeled biological motion perception as a feedforward, discriminative process, e.g., through specialized visual mechanisms matching point-light kinematics against learned templates/features to directly classify the action (Johansson, 1973; Giese & Poggio, 2003; Lange & Lappe, 2006). Perhaps, this would be analogous to directly fine-tuning models that take in a JKP video and output a class label (e.g., MLLMs). As we shall show, this leads to memorization rather than generalization unlike human vision which generalizes remarkably well.

A second line of evidence shows that the observer’s motor system (premotor, primary motor, and parietal cortex that plan and control one’s own body movements) gets activated while perceiving JKP

videos (Saygin et al., 2004; Saygin, 2007; Calvo-Merino et al., 2005). This suggests that observers may recognize an action in part by internally simulating how they would perform it, rather than by classifying the visual input directly. Our approach takes inspiration from this view: analysis-by-synthesis is the computational analogue of such simulation.

Specifically, instead of fine-tuning the MLLM, we propose to train a video-to-video generative model to learn the mapping from sparse JKP inputs to the physical RGB instantiation of the action. The premise of this approach is that if indeed there is signal in sparse Johansson keypoints, a strong generative model should be able to pick it up. Once the RGB video is generated, we can use any downstream frozen model (e.g., MLLMs) to recognize the action.

Architecture. Given a pre-trained video generation model G_ϕ , our goal is to transform it into a video-to-video model G_ϕ^* fine-tuned on generating RGB video from a JKP video. We instantiate RBI with G_ϕ as LTX-2 (19B) (HaCohen et al., 2026), a text to video+audio generation model as our base model. LTX-2 broadly consists of three components: (i) modality-specific causal VAEs to compress inputs into latents, (ii) a text embedding pipeline based on Gemma3-12B followed by modality-specific connectors, and (iii) a dual-stream DiT that processes audio and video latents in parallel with separate text-to-modality cross attention followed by a shared audio-visual cross attention. We modify the architecture for LoRA fine-tuning for video-to-video generation task. We remove the audio pipeline, retain the text-to-video conditioning as is, and use a generic text prompt for the entire dataset (see Appendix Section B). For video conditioning (i.e., to ingest the input JKP video), we use LoRA in-context conditioning (Huang et al., 2024). Concretely, we project the JKP video to the same causal VAE’s latent space and sequence-concatenate the JKP latents with those of the noised latents for the ground truth RGB video. We insert LoRA adapters in each of the attention projection and FFN layers in DiT. This amounts to a total of 327M trainable parameters.

Training and inference. We fine-tune LTX-2 separately on the train set of each of the two splits: NTU-CI for cross-instance generalization and NTU-CA for cross-action generalization. For inference on unseen data (e.g., UTD), we use the model trained on NTU-CI. We LoRA fine-tune the base model (ltx-2-19b-dev) to generate RGB video conditioned on a JKP video. Note that we do not use any additional image conditioning for appearance information. We need inference to be fast since our task needs generating thousands of videos. Thus, for faster inference, we use the 8-step distilled model (ltx-2-19b-distilled) and attach the trained LoRA weights to it. We empirically confirm that the video quality is preserved even when using the distilled model during inference (see Table 8). For JohanssonBench, we keep the same constant text prompt from training.

Action recognition. Given the generated video, one can employ any downstream recognition model D_ψ (e.g., MLLM) to infer the action. D_ψ is ideally a video-text model so that we can test generalization to novel actions. Note that the same trained generative model G_ϕ^* can be attached to *any* discriminative MLLM D_ψ : their combination yields the complete RBI model $F_{\theta=\{\phi,\psi\}} = D_\psi(G_\phi^*(X_{JKP}))$. Note RBI is an unsupervised approach, as in, it is label-free with respect to categorical action classes.

4 EXPERIMENTAL RESULTS

Notation. Let $X_{\text{RGB}} \in \mathcal{X} = \mathbb{R}^{T \times H \times W \times C'}$ be an RGB video and $X_{\text{JKP}} \in \mathcal{X}$ be the corresponding JKP video. T is the number of frames, H, W is height, width and C' ($= 3$) is the channel dimension. X_{RGB} fed directly to an MLLM, $D_\psi(X_{\text{RGB}}) : \mathcal{X} \rightarrow \mathcal{C}$, serves as an “oracle” for our *imagination*-based model, $F_\theta(X_{\text{JKP}}) = D_\psi(G_\phi^*(X_{\text{JKP}}))$. In supervised fine-tuning, $D_\psi(\cdot)$ is fine-tuned on X_{JKP} .

Brief overview of experiments. In Section 4.1, we evaluate zero-shot open-weight and closed MLLMs on JohanssonBench on the NTU-CI split. The best MLLM (GPT 5.6 Sol; 43.4%) is just over half of the human baseline (81%). In Section 4.2, we also compare supervised fine-tuned (SFT) baselines with our generative (RBI) approach. While SFT performs strongly on the cross-instance split, it struggles on novel actions and the unseen dataset. In contrast, RBI beats SFT on all three datasets with remarkable generalizability. Not only that, RBI also generalizes with stimuli beyond JKP: be it segmentation masks, joints or silhouette contours as shown in the next Section.

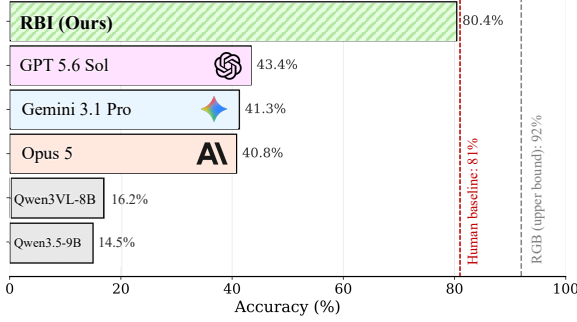


Figure 2: **JohanssonBench results.** We compare the best open/closed frontier models on the NTU-CI split of JohanssonBench. Given a video of Johansson’s keypoints with no appearance cues, the task is to recognize the action. Top frontier models (e.g., GPT 5.6 Sol) are about halfway compared to the human baseline. Our generative model (RBI) outperforms frontier models and matches human baseline.

4.1 ZERO-SHOT BASELINES

We conduct a thorough zero-shot evaluation with several open and closed models. For this experiment, we only evaluate on the NTU-CI test set consisting of $N=525$ samples.

Human baseline. As sanity check, we also conduct a systematic survey on Amazon Mechanical Turk (AMT) to get a human baseline on this evaluation set. Since this is a fundamental vision task, we recruit ordinary people with no prior expertise/knowledge of Johansson’s experiments. Each sample includes the JKP video along with a simple instruction prompt to choose one of $C=21$ action choices. Each sample is annotated by $k=3$ annotators. In total, our survey included $K=39$ annotators with each labeling ~ 40 examples on average. We found the Fleiss’s kappa to be $\kappa = 0.6$ (moderate-to-substantial)². We use a majority-vote to obtain the prediction for each sample. This results in baseline performance of 81%.

Evaluation setting for MLLMs. Given a JKP video, we construct a prompt for zero-shot action classification. We state all the action classes as options for the MLLM to select and force it to output the exact action class name to avoid bias to alphabetic/roman numeric options (Yang et al., 2025). We also randomize the option order to avoid order bias common in LLMs (Pezeshkpour & Hruschka, 2024). We automatically parse the output and check for correctness programatically. Later, we conduct ablations on these choices. In Appendix Section C.1, we also include results for few-shot recognition. In general, they do not perform much better than zero-shot baselines. Unless specified otherwise, we sample $F=16$ uniformly from the video. For models that need to take in videos based on FPS (e.g., Gemini), we use $FPS=4$ which is about 12-20 frames based on typical video lengths. To obtain the prediction, we use standard greedy decoding for closed-set MCQ with a short generation budget (max tokens 64 since action names are fairly short) following standard practice (Zhang et al., 2024a). Answer options are shuffled per sample with a fixed seed for reproducibility. We use the `ms-swift` (Zhao et al., 2024) codebase for evaluation (and training).

Models. We pick a host of open and closed MLLMs and evaluate it on our benchmark. Among open-weight models, we evaluate representatives from the following families: Qwen (2-VL, 2.5-VL, 3-VL, 3.5) (Bai et al., 2025), InternVL (3, 3.5) (Wang et al., 2025), Gemma 4 (Team et al., 2026), Molmo 2 (Clark et al., 2026), Video LLaMA 3 (Zhang et al., 2025), etc. We evaluate the closed models through their respective API: OpenAI (GPT 5, GPT 5.5, GPT 5.6), Google (Gemini 2.5, 3.0, 3.1, 3.5) and Anthropic (Claude Sonnet 4/5 and Opus 5). While the frontier continues to expand, we believe this set of models is representative of the frontier at the time of writing.

Results at a glance. The key highlight results are shown in Fig. 2. Chance is at 4.8% while the human baseline is at 81%. We show only the best open and closed models. The key takeaways are: (i) the best performing models (GPT 5.6 Sol, Gemini 3.1 Pro, Opus 5) are just around the halfway mark compared to humans, (ii) the best open model (InternVL3-VL 8B) is significantly behind almost all the closed models and enormously behind the *frontier* closed models on this task. Complete results are provided in Appendix Section C.1.

²Fleiss’s kappa is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items.

4.2 GENERATIVE (RBI) VS. DISCRIMINATIVE (SFT) APPROACHES

Given the relatively poor performance with zero-shot MLLMs, we next try supervised fine-tuning (SFT) baselines with four open models: Qwen3VL-8B, Qwen3.5 9B, InternVL3 8B, InternVL-3.5 8B. We attach LoRA adapters to the linear layers in each Transformer block while keeping the vision encoder frozen. We fine-tune with standard language modeling loss with the action class as the ground truth sequence. We use the same splits as were used to train RBI. We compare the performance of discriminative (MLLMs $\rightarrow$ SFT) vs. generative (RBI $\rightarrow$ MLLM) on all three splits in Table 2. On the cross-instance split, SFT is almost as strong as RBI: RBI beats SFT by a small margin of +1.8 points (averaged across models). But on cross-action and cross-dataset splits, RBI significantly outshines SFT by +33.2 and +11.0 points respectively. This firmly establishes the superior generalizability of generative approach and hints at memorization in discriminative SFT.

Table 2: We evaluate various MLLMs (in 8–9B parameter range) for action classification. RGB (zero-shot) denotes the upper bound evaluated for action classification on RGB videos. The columns under *JKP* are evaluated on Johansson’s keypoint videos. We include a zero-shot baseline. SFT denotes supervised fine-tuning on JKP. RBI denotes our approach (Recognition by Imagination). On the simpler split (a) (same actions across train and test), SFT matches RBI, but on the more challenging splits, RBI generalizes substantially better than SFT.

Split $\rightarrow$	(a) Cross-instance				(b) Cross-action				(c) Cross-dataset			
	NTU: $N=525, C=21$				NTU: $N=250, C=10$				UTD: $N=861, C=27$			
Input modality	RGB	JKP			RGB	JKP			RGB	JKP		
Protocol	ZS	ZS	SFT	RBI	ZS	ZS	SFT	RBI	ZS	ZS	SFT	RBI
Chance	4.8	—	—	—	10.0	—	—	—	3.7	—	—	—
Qwen3VL-8B-Inst.	76.6	16.2	61.3	67.8	82.0	27.2	42.0	75.2	25.6	3.1	8.0	24.2
Qwen3.5-9B	81.0	14.5	70.9	71.8	85.2	21.2	59.2	80.0	43.4	4.1	15.4	32.4
InternVL3.5-8B	84.4	12.2	77.1	76.6	90.1	20.8	44.8	80.8	29.0	2.3	14.8	20.9
InternVL3-8B	92.0	7.4	80.4	80.4	96.0	10.8	44.4	87.2	28.9	2.6	17.0	21.8
Average	83.5	12.6	72.4	74.2	88.3	20.0	47.6	80.8	31.7	3.0	13.8	24.8

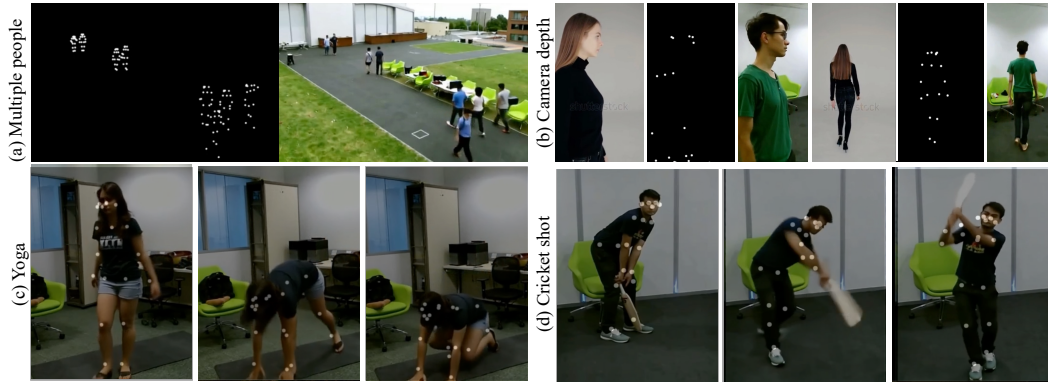


Figure 3: **Qualitative results with RBI generation.** In each case the input is a JKP video, and the output is a RGB video. We show RBI’s successful generalization to (a) multiple persons in a new scenario (e.g., University campus), (b) varying camera depth/ num. kps., (c, d) hard poses like Yoga asana and Cricket Batting.

Qualitative generalization. We also run several in-the-wild samples to qualitatively stress-test RBI (Fig. 3). It successfully generalizes to a multi-person JKP video even though it was fine-tuned on ≤ 2 persons, i.e., RBI leads to recognize each JKP group of points as a complete human and effectively composes information. We also test on challenging poses such as a Yoga asana or a Cricket shot and find very compelling results as shown in the Supplementary video. In Appendix Section D, we also test its robustness to occlusion, n.o. keypoints and even to animal motion.

5 GENERALIZATION TO FORM-FROM-MOTION STIMULI

We discover several interesting and somewhat surprising avenues where RBI generalizes. RBI works off-the-shelf with stimuli that differ slightly from JKP, e.g., keypoints + joints, body outlines, silhouettes, etc., beyond the JKP. It also generates plausible videos for inputs drastically different than JKP, e.g., event-camera videos (Dong et al., 2023) or motion captcha (Upadhyay et al., 2026). In a sense, this hints to JKP being a robust, minimal *gist of motion*, which when trained on, helps generation models generalize.

5.1 IMPACT OF PROGRESSIVE VISUAL ABSTRACTION

We want to study the impact of gradually decreasing contextual information from the video, or, in other terms, increasing the visual abstraction from RGB down to JKP videos. For this, we pick a subset of the NTU-CI evaluation set with $n_c=5$ samples per action totaling $n=105$ videos. We use SAM2 to obtain the intermediate segmentation masks and body outlines. As shown in Fig. 4, for both open/closed models, as we peel away appearance information, the performance drops quite significantly. While human vision even on JKP achieves remarkable results, this ability does not emerge in MLLMs trained on millions of videos. In contrast, RBI, trained only on JKP samples, maintains a consistently strong performance. In the Appendix Section A.4, we also analyze the temporal sensitivity of original RGB vs. RBI generated videos. We show that JKP perception is not solved simply by mapping the keypoints to a form (such as a person) in a single frame.

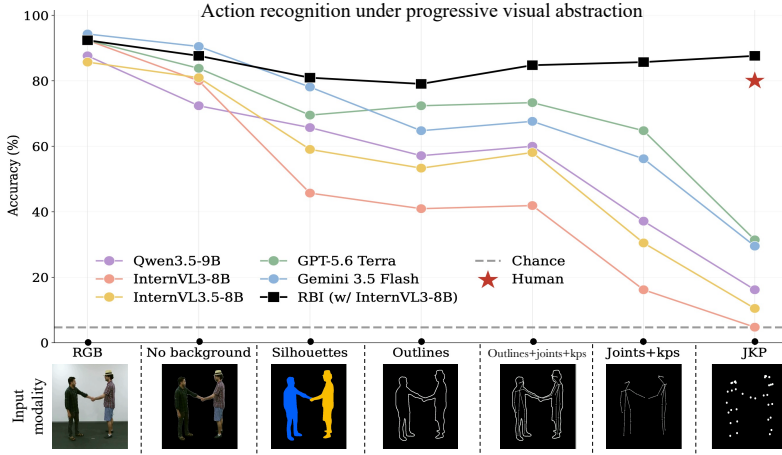


Figure 4: **Varying input stimuli.** As shown on the x -axis, we gradually vary the input stimuli to RBI (& other models) at test time even though RBI is only trained on JKP videos. On standard MLLMs, we observe a dramatic decline in accuracy as we strip off appearance information from RGB to JKP. In contrast, our proposed RBI retains the core visual abstraction irrespective of the background visual context.

5.2 PRIVACY-PRESERVING VIOLENCE RECOGNITION

Dynamic Vision Sensors (DVS) cameras are often used to detect violent incidents while preserving privacy. We discover that RBI, without any additional training, generates plausible RGB videos even from DVS videos instead of the JKP videos it is trained on. This opens up the possibility of privacy-preserving violence detection on RBI-generated videos with existing standard vision models. This also sidesteps the need to design specialized architectures for sparse DVS inputs. An example generation is shown in Fig. 5 (top-left). As a concrete proof-of-concept, we evaluate this idea on Bullying-10K (Dong et al., 2023): a dataset with 10 action classes with $N_{\text{train}} = 8,000$ and $N_{\text{test}} = 2,000$. Since video generation is expensive, we subsample 20% of the training data and generate the corresponding synthetic videos. Likewise, we generate the corresponding synthetic test set with RBI. Since the generated videos are plausible, standard video embedding models may be used with light-weight linear probes for action recognition. We use frozen embeddings from TARA (Bagad & Zisserman, 2026) and InternVideo-2 (Wang et al., 2024) and train a linear classifier on the subsampled synthetic train set and evaluate it on the synthetic test set. Results are shown in Table 3. Our simple baseline trained only on 20% of the data outperforms specialized models.

5.3 CO-SPEECH GESTURE GENERATION

Are Johansson’s keypoints also sufficient to encode human co-speech gestures? We test this on the Gesture Recognition in the Wild (GRW) benchmark (Hegde et al., 2025; 2026). Again, we use RBI off-the-shelf to generate a synthetic video of a gesturing person given the JKP inputs. We

Video model	Accuracy
<i>DVS; Full ft. w/ 100% data</i>	
SlowFast	74.0
X3D	76.9
STAFM-SlowFast	77.2
SNN-Chaos 2	80.8
EWB-ResNet18	86.5
DSC-SNN	87.4
<i>Synthetic (Ours); Linear w/ 20% data</i>	
VJEPa 2.1 (avgpool)	67.5
VJEPa 2 (avgpool)	78.3
InternVideo 2	81.9
TARA	88.2
TARA $\oplus$ InternVideo 2	89.5

Table 3: **Event-camera inputs.** RBI can be used off-the-shelf to generate realistic videos from event-camera signals enabling privacy-preserving action recognition.

Method	Recognize			Localize
	A@1	A@5	A@10	mIoU
Random	1.00	5.00	10.00	0.20
<i>Fine-tuned</i>				
GestSync	9.12	20.37	30.18	0.52
JEGAL	10.43	22.71	31.88	0.54
<i>MLLMs</i>				
Intern-VL	4.35	12.50	15.71	0.39
Qwen3-VL	5.92	13.19	19.27	0.21
Gemini-3.5	13.95	30.15	41.10	0.47
<i>GestureNet (Hegde et al. (2026))</i>				
RBI (KPs+Joints)	15.17	33.88	47.55	0.64
Original (RGB)	18.35	37.30	51.70	0.67

Table 4: **Gesture generation.** RBI can also be used for generating co-speech gestures conditioned on keypoints+joints. Synthetic gestures recover 90+% of performance of GestureNet compared to that with original videos.

find that it is beneficial to include the lines connecting keypoints and the finger keypoints instead of the raw JKP inputs since the gestures also involve hand shapes. An example generated gesture sequence is shown in Fig. 5. We evaluate the quality of generated gestures on (i) recognition, and (ii) localization in Hegde et al. (2026). We generate samples for the entire test set of 2000 videos and use the model from Hegde et al. (2026). Results are shown in Table 4. On average, using synthetic videos generated by RBI is able to recover over 90% of the performance compared to using original RGB videos. Note that our model was not explicitly trained on gestures.

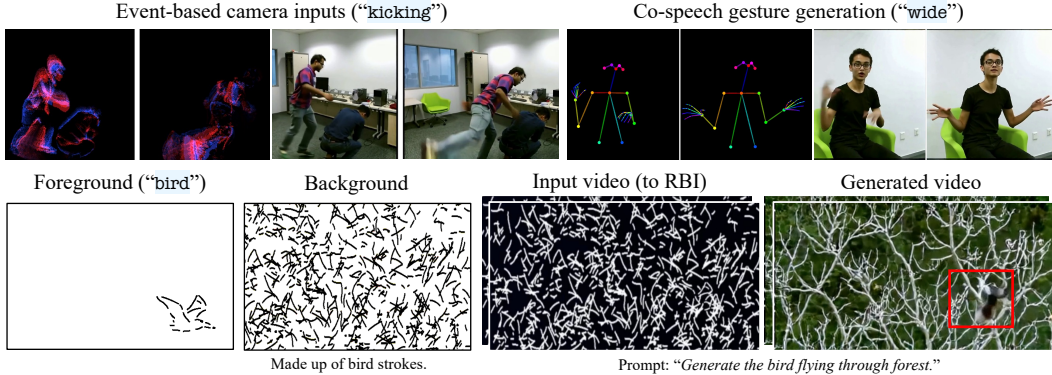


Figure 5: **Applications.** RBI works off-the-shelf on event-based camera inputs, or with complete pose map (for gesture generation). It also works on other psychophysical tests like the hidden bird illusion: it generates the bird inferring its location solely from its motion. More results in Appendix.

5.4 NOVEL FORM-FROM-MOTION STIMULI

We also qualitatively test on two new, challenging psychophysics stimuli: (i) we consider examples in SpookyBench (Upadhyay et al., 2026) where an object/shape/text is hidden in dynamic noise and becomes perceptible only through the motion of the noise. RBI is able to pick out the shape/object in several cases but cannot pick out the text letters from the noise. See Appendix Section D for more qualitative results. (ii) We test on the *hidden bird* illusion (Bach, 2006; Regan, 1986) which shows a flying bird made up of 21 penstrokes camouflaged against a background of the same strokes scrambled at random, so it is invisible when static but pops out instantly as it moves. With an appropriate prompt, RBI can locate the flying bird in the background noise (see Fig. 5 (bottom)).

6 RELATED WORK

Human (biological) motion perception. Johansson (1973)’s seminal paper showed that human vision is capable of recognizing *biological motion patterns* represented merely by a few bright spots on the main joints against a black background. This work was followed by a rich line of work

showing various interesting and somewhat surprising capabilities of human vision of JKP stimuli: recognizing gender (Kozlowski & Cutting, 1977), identity (Cutting & Kozlowski, 1977), emotion (Clarke et al., 2005), physical properties (Runeson & Frykholm, 1981; 1983), etc. Developmental psychologists have also found the core motion perception ability in newborn infants suggesting the presence of an inherent “life-detector” in human vision that perceives JKP moving dots as human or animal actions (Fox & McDaniel, 1982; Simion et al., 2008). Related psycho-physical phenomena similarly show that visual structure can emerge only through integration across views or time, as in random-dot stereograms, where depth-defined shapes become visible through binocular fusion (Julesz, 1960), and random-dot kinematograms, where coherent motion reveals structure across otherwise unstructured successive frames (Braddick, 1974). We include a more detailed discussion on *why* and *how* human vision achieves such motion perception in Appendix Section E.

Action recognition and video datasets. Early computer vision research had a prominent focus on human actions (Yamato et al., 1992; Yacoob & Black, 1999). Between 2000-2010, action recognition research focused on design of hand-crafted spatiotemporal features (Laptev, 2005) and standardized datasets (Schuldt et al., 2004; Soomro et al., 2012; Kuehne et al., 2011). As deep learning took off, action recognition became much easier thanks to massive datasets (Carreira & Zisserman, 2017) and very high-capacity models (He et al., 2016; Vaswani et al., 2017). As MLLMs started to dominate the video understanding discourse, benchmarks largely focus on QA or captioning. However, the community has repeatedly discovered that most of these datasets suffer from an *appearance bias* (Lei et al., 2023; Huang et al., 2018; Sevilla-Lara et al., 2021; Xue et al., 2026), i.e., the tasks can be solved by a single frame or an unordered set of frames without the need for genuine temporal/motion perception. There have been some notable examples of benchmarks to probe true temporal understanding (Li et al., 2026; Upadhyay et al., 2026; Cores et al., 2024). However, it is still difficult to completely disentangle appearance and motion without leaving artifacts (Choi et al., 2019; Fukuzawa et al., 2025; Li et al., 2023). Johansson’s experiments present a unique opportunity to strip off all appearance cues and focus on minimalist human motion representation.

Johansson’s experiments in computer vision. There have been two prior efforts that have explored this direction of Johansson’s biological motion perception for vision models. Han et al. (2024) first proposed to use Johansson’s keypoints to probe specialized vision backbones (e.g., ResNet, MViT) in a supervised manner. More recently, Kadambi et al. (2025) proposed a similar evaluation for MLLMs. While this work also hinges on the core Johansson’s experiments, we distinguish this work in the following ways: (i) systematic evaluation of generalization (e.g., cross-action and cross-dataset) while fine-tuning MLLMs, (ii) a novel computational approach, Recognition by Imagination, (iii) discovery of fascinating emergent properties of a model trained on Johansson’s keypoints.

7 CONCLUSION AND DISCUSSION

In this paper, we studied form-from-motion based on Johansson’s experiments: can semantic actions be recovered solely from the motion of a handful of keypoints? If so, the objective was to build a generalizable computation model to solve this task. In pursuit of these research questions, we developed a quantitative benchmark, *JohanssonBench*, and proposed a novel generative approach to this task. In the process, we made the following notable observations:

- Frontier MLLMs are around the halfway mark compared to human baseline. While SFT bridges this gap when train-test actions are same, it struggles to generalize to novel actions/datasets.
- We propose Recognition by Imagination (RBI) to first imagine the RGB video from the JKP input and then recognize the action. RBI beats SFT in all settings; generalizes remarkably well.
- RBI also works off-the-shelf with other stimuli such as event-based camera inputs, or other psychophysical tests like examples from *SpookyBench* or the *Hidden Bird* illusion.

Human perception of JKP stimulus indicates it is a minimal, compact representation of motion, i.e., a *gist of motion*. The robust generalization of this generative approach indicates that: (i) JKP is indeed the representational gist of motion, and (ii) RBI is able to meaningfully extract it. However, some broader questions still remain. Humans do this task seemingly without any training – so this ability perhaps emerged through evolution (Simion et al., 2008). In contrast, we explicitly train a model to map JKP to RGB. It is an open question on whether / how can we adapt these models with little-to-no training (e.g., by test-time training?). Moreover, while this ability emerges naturally in humans, as we show, that is not the case in MLLMs trained on web-scale data. RBI suggests that, perhaps, future MLLMs could benefit from an internal *imagination* mechanism that helps them simulate videos either explicitly or in the latent state before making a decision.

REFERENCES

- Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.
- AIGC-Apps. VideoX-Fun: A Video Generation Pipeline for Diffusion Transformer, 2026. URL <https://github.com/aigc-apps/VideoX-Fun>.
- Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.
- Michael Bach. Hidden Bird. Visual Phenomena & Optical Illusions, July 2006. URL <https://michaelbach.de/ot/cog-hiddenBird/>. Accessed: 2026-09-06.
- Piyush Bagad and Andrew Zisserman. Adapting MLLMs for Nuanced Video Retrieval. *European Conference on Computer Vision (ECCV)*, 2026.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.
- Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.
- Randolph Blake and Maggie Shiffrar. Perception of Human Motion. *Annual Review of Psychology*, 2007.
- Oliver Braddick. A Short-Range Process in Apparent Motion. *Vision research*, 14(7):519–527, 1974.
- Beatriz Calvo-Merino, Daniel E Glaser, Julie Grèzes, Richard E Passingham, and Patrick Haggard. Action Observation and Acquired Motor Skills: An fMRI Study with Expert Dancers. *Cerebral cortex*, 15(8):1243–1249, 2005.
- Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733. IEEE, 2017.
- Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor. In *International Conference on Image Processing (ICIP)*, pp. 168–172. IEEE, 2015.
- Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why Can’t I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.
- Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28652–28668, 2026.
- Tanya J Clarke, Mark F Bradshaw, David T Field, Sarah E Hampson, and David Rose. The Perception of Emotion from Body Movement in Point-Light Displays of Interpersonal Dialogue. *Perception*, 34(10):1171–1180, 2005.
- CMU Graphics Lab. CMU graphics lab motion capture database. <http://mocap.cs.cmu.edu>, 2003. The database was created with funding from NSF EIA-0196217.
- Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Lost in Time: A New Temporal Benchmark for Videollms. *arXiv preprint arXiv:2410.07752*, 2024.

- James E Cutting and Lynn T Kozlowski. Recognizing Friends by Their Walk: Gait Perception without Familiarity Cues. *Bulletin of the psychonomic society*, 9(5):353–356, 1977.
- James E Cutting, Cassandra Moore, and Roger Morrison. Masking the Motions of Human Gait. *Perception & psychophysics*, 44(4):339–347, 1988.
- Yiting Dong, Yang Li, Dongcheng Zhao, Guobin Shen, and Yi Zeng. Bullying10K: A Large-Scale Neuromorphic Dataset Towards Privacy-Preserving Bullying Recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:1923–1937, 2023.
- Robert Fox and Cynthia McDaniel. The Perception of Biological Motion by Human Infants. *Science*, 218(4571):486–487, 1982.
- Takumi Fukuzawa, Kensho Hara, Hirokatsu Kataoka, and Toru Tamaki. Can Masking Background and Object Reduce Static Bias for Zero-Shot Action Recognition? In *International Conference on Multimedia Modeling*, pp. 366–379. Springer, 2025.
- Prajwal Gatti, Simon Jenni, Fabian Caba Heilbron, and Dima Damen. Gen2balance: Generative balancing for long-tailed video action recognition. In *European Conference on Computer Vision (ECCV)*, 2026.
- Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–12. IEEE, 2025.
- Martin A. Giese. Biological and Body Motion Perception. In Johan Wagemans (ed.), *The Oxford Handbook of Perceptual Organization*, pp. 575–600. Oxford University Press, 2015. doi: 10.1093/oxfordhb/9780199686858.013.008.
- Martin A. Giese and Tomaso Poggio. Neural Mechanisms for the Recognition of Biological Movements. *Nature Reviews Neuroscience*, 2003.
- Markus Graf, Bianca Reitzner, Caroline Corves, Antonino Casile, Martin Giese, and Wolfgang Prinz. Predicting Point-Light Actions in Real-Time. *Neuroimage*, 36:T22–T32, 2007.
- ED Grossman and R Blake. Perception of Coherent Motion, Biological Motion and Form-from-Motion under Dim-Light Conditions. *Vision research*, 39(22):3721–3727, 1999.
- Emily D. Grossman and Randolph Blake. Brain Areas Active During Visual Perception of Biological Motion. *Neuron*, 35:1167–1175, 2002. doi: 10.1016/S0896-6273(02)00897-8.
- Emily D. Grossman, Maria Donnelly, Ray Price, David Pickens, Vincent Morgan, Gregory Neighbor, and Randolph Blake. Brain Areas Involved in Perception of Biological Motion. *Journal of Cognitive Neuroscience*, 12:711–720, 2000. doi: 10.1162/089892900562417.
- Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. LTX-2: Efficient Joint Audio-Visual Foundation Model. *arXiv preprint arXiv:2601.03233*, 2026.
- Shuangpeng Han, Ziyu Wang, and Mengmi Zhang. Flow Snapshot Neurons in Action: Deep Neural Networks Generalize to Biological Motion Perception. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:53732–53763, 2024.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- Eric Hedlin, Gopal Sharma, Shweta Mahajan, Xingzhe He, Hossam Isack, Abhishek Kar, Helge Rhodin, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised keypoints from pretrained diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22820–22830. IEEE, 2024.

- Sindhu B Hegde, KR Prajwal, Taein Kwon, and Andrew Zisserman. Understanding Co-Speech Gestures in-the-Wild. In *International Conference on Computer Vision (ICCV)*, pp. 9977–9987. IEEE, 2025.
- Sindhu B Hegde, KR Prajwal, and Andrew Zisserman. Recognizing Co-Speech Gestures in-the-Wild. *arXiv preprint arXiv:2605.31589*, 2026.
- Paul E. Hemeren. *Mind in Action: Action Representation and the Perception of Biological Motion*. PhD thesis, Lund University, 2008.
- Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.
- De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 7366–7375. IEEE, 2018.
- Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-Context Lora for Diffusion Transformers. *arXiv preprint arXiv:2410.23775*, 2024.
- Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-Time Multi-Person Pose Estimation Based on Mmpose. *arXiv preprint arXiv:2303.07399*, 2023.
- Gunnar Johansson. Visual Perception of Biological Motion and a Model for Its Analysis. *Perception & Psychophysics*, 14:201–211, 1973. doi: 10.3758/BF03212378.
- Mark H Johnson. Biological Motion: A Perceptual Life Detector? *Current biology*, 16(10):R376–R377, 2006.
- Bela Julesz. Binocular Depth Perception of Computer-Generated Patterns. *Bell System Technical Journal*, 39(5):1125–1162, 1960.
- Akila Kadambi, Marco Iacoboni, Lisa Aziz-Zadeh, and Srini Narayanan. Evaluating Point-Light Biological Motion in Multimodal Large Language Models. *arXiv preprint arXiv:2509.23517*, 2025.
- Lynn T Kozlowski and James E Cutting. Recognizing the Sex of a Walker from a Dynamic Point-Light Display. *Perception & psychophysics*, 21(6):575–580, 1977.
- Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A Large Video Database for Human Motion Recognition. In *International Conference on Computer Vision (ICCV)*, 2011.
- Joachim Lange and Markus Lappe. A Model of Biological Motion Perception from Configural Form Cues. *Journal of Neuroscience*, 2006.
- Ivan Laptev. On Space-Time Interest Points. *International Journal of Computer Vision (IJCV)*, 64(2):107–123, 2005.
- Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Association of Computational Linguistics (ACL)*, pp. 487–507, 2023.
- Baiqi Li, Kangyi Zhao, Ce Zhang, Chancharik Mitra, Jean de Dieu Nyandwi, and Gedas Bertasius. Timeblind: A Spatio-Temporal Compositionality Benchmark for Video LLMs. *arXiv preprint arXiv:2602.00288*, 2026.
- Haoxin Li, Yuan Liu, Hanwang Zhang, and Boyang Li. Mitigating and Evaluating Static Bias of Action Representations in the Background and the Foreground. In *International Conference on Computer Vision (ICCV)*, pp. 19854–19866. IEEE, 2023.

- Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. *arXiv preprint arXiv:2406.16862*, 2024.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, and Wenming Yang. RTMO: Towards High-Performance One-Stage Real-Time Multi-Person Pose Estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1491–1500. IEEE, 2024.
- George Mather and Sophie West. Recognition of Animal Locomotion from Dynamic Point-Light Displays. *Perception*, 22(7):759–766, 1993.
- Koichi Namekata, Yash Kant, Zhizheng Liu, Ryan Burgert, Yuancheng Xu, Kuan Heng Lin, Emmett Steven, Julien Philip, Li Ma, Andrea Vedaldi, et al. Go-with-the-track: Video compositing and motion control with point tracking. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026.
- Marina Pavlova and Alexander Sokolov. Orientation Specificity in Biological Motion Perception. *Perception & Psychophysics*, 62(5):889–899, 2000.
- Pouya Pezeshkpour and Estevam Hruschka. Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, 2024.
- Aina Puce and David Perrett. Electrophysiology and Brain Imaging of Biological Motion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358:435–445, 2003. doi: 10.1098/rstb.2002.1221.
- D Regan. Form from Motion Parallax and Form from Luminance Contrast: Vernier Discrimination. *Spatial vision*, 1(4):305–318, 1986.
- Sverker Runeson and Gunilla Frykholm. Visual Perception of Lifted Weight. *Journal of experimental psychology: Human Perception and Performance*, 7(4):733, 1981.
- Sverker Runeson and Gunilla Frykholm. Kinematic Specification of Dynamics as an Informational Basis for Person-and-Action Perception: Expectation, Gender Recognition, and Deceptive Intention. *Journal of experimental psychology: general*, 112(4):585, 1983.
- Ayse Pinar Saygin. Superior Temporal and Premotor Brain Areas Necessary for Biological Motion Perception. *Brain*, 130(9):2452–2461, 2007.
- Ayse Pinar Saygin, Stephen M Wilson, Donald J Hagler Jr, Elizabeth Bates, and Martin I Sereno. Point-Light Biological Motion Perception Activates Human Premotor Cortex. *The Journal of neuroscience*, 24(27):6181–6188, 2004.
- Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing Human Actions: A Local SVM Approach. In *International Conference on Pattern Recognition (ICPR)*, volume 3, pp. 32–36. IEEE, 2004.
- Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Winter Conference on Applications of Computer Vision (WACV)*, pp. 535–544. IEEE, 2021.
- Amir Shahrudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu Rgb+ D: A Large Scale Dataset for 3D Human Activity Analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1010–1019, 2016.
- Ziyao Shanguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing Visual Temporal Reasoning Capabilities in Multimodal Foundation Models. In *International Conference on Learning Representations*, volume 2025, pp. 7593–7734, 2025.

- Ayush Shrivastava, Sanyam Mehta, Daniel Geng, and Andrew Owens. Point prompting: Counterfactual tracking with video diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 152967–152983, 2026.
- Francesca Simion, Lucia Regolin, and Hermann Bulf. A Predisposition for Biological Motion in the Newborn Baby. *Proceedings of the National Academy of Sciences*, 105(2):809–813, 2008.
- Khurram Soomro, Amir Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. *arXiv preprint arXiv:1212.0402*, 2012.
- Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 Technical Report. *arXiv preprint arXiv:2607.02770*, 2026.
- Nikolaus F. Troje and Cord Westhoff. The Inversion Effect in Biological Motion Perception: Evidence for a “Life Detector”? *Current Biology*, 2006.
- Ujjwal Upadhyay, Mukul Ranjan, Zhiqiang Shen, and Mohamed Elhoseiny. Time Blindness: Why Video-Language Models Can’t See What Humans Can? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30906–30918, 2026.
- Jan Vanrie and Karl Verfaillie. Perception of Biological Motion: A stimulus set of human point-light actions. *Behavior Research Methods, Instruments, & Computers*, 36(4):625–629, 2004.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.
- Guangyuan Wang, Li Hu, Dechao Meng, Zhongyi Zhang, Peng Zhang, Mingyang Huang, Ruoshi Zhang, Ke Sun, Zhe Zhang, Xingjun Wang, et al. Wan-animate-2: Pushing the application boundaries of character animation. *arXiv preprint arXiv:2608.06009*, 2026.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internv13. 5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025.
- Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling Foundation Models for Multimodal Video Understanding. In *European Conference on Computer Vision (ECCV)*, pp. 396–416. Springer, 2024.
- Zihui Sherry Xue, Romy Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:90925–90955, 2026.
- Yaser Yacoob and Michael J Black. Parameterized Modeling and Recognition of Activities. *Computer Vision and Image Understanding (CVIU)*, 73(2):232–247, 1999.
- Junji Yamato, Jun Ohya, Kenichiro Ishii, et al. Recognizing Human Action in Time-Sequential Images Using Hidden Markov Model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 92, pp. 379–385, 1992.
- Zhen Yang, Ping Jian, and Chengzhi Li. Option Symbol Matters: Investigating and Mitigating Multiple-Choice Option Symbol Bias of Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1902–1917, 2025.
- Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective Whole-Body Pose Estimation with Two-Stages Distillation. In *International Conference on Computer Vision Workshops (ICCVW)*, pp. 4212–4222. IEEE, 2023.
- Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-Token Vit: Training Vision Transformers from Scratch on Imagenet. In *International Conference on Computer Vision (ICCV)*, pp. 558–567, 2021.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models. *arXiv preprint arXiv:2407.12772*, 2024a.

Yuang Zhang, Jiayi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024b.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A Scalable lightweight Infrastructure for Fine-Tuning. *arXiv preprint arXiv:2408.05517*, 2024.

STATEMENTS

AI USE STATEMENT

In this work, we used generative AI tools to (i) clean and preprocess dataset, (ii) implement methods (e.g., fine-tuning LTX-2 for video-to-video generation, or pose detection inference), (iii) provide feedback on research methodology (e.g., in refining the broad idea and approach). Additionally, we used generative AI tools to (i) create or modify scientific figures or images, (ii) summarize or analyze existing literature, (iii) brainstorming. We have reviewed all AI-assisted work. For example, we manually verified each stage of the data processing pipeline through visual and quantitative sanity checks. Since the fine-tuning code for LTX-2 is publicly available, the AI tools only re-organized the code and wrote meta-scripts to run fine-tuning. We verified the pipeline visually through generated samples as well as by overfitting on a small subset of the dataset. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our approach is based on fine-tuning an open-source video generation model for a video-to-video task. Since the base model is pre-trained on massive amounts of video data, it inherits the biases of the dataset it was trained on and can be potentially misused. Furthermore, since it is fine-tuned on human action data, it also biased towards generating one or multiple humans in an indoor lab setting making the model narrower in its scope. We have verified that our fine-tuning data does not include any NSFW examples. In summary, our fine-tuning recipe does not exacerbate the risks associated with the publicly available base open model.

REPRODUCIBILITY STATEMENT

In terms of data, NTU-60 (Shahroudy et al., 2016) dataset is publicly available. In terms of model, our approach is based on fine-tuning an open-source video model, LTX-2 (HaCohen et al., 2026), which has open-sourced the code for fine-tuning for a video-to-video task. Thus, it should be feasible to directly use their code for our task. Our preprocessing steps are clearly described in Section A.1 and should be easy to implement. Once preprocessed, the LTX-2 codebase also has code to compute the latents.

APPENDIX

Table of Contents

A	Data	16
A.1	Construction and pre-processing	16
A.2	Quality of poses from DWPose	17
A.3	Split details	17
A.4	Temporal sensitivity analysis	17
A.5	Evaluation with 3D Motion Capture datasets	19
B	Model	19
B.1	Architecture and training details	19
B.2	Prompts	20
C	Quantitative Experiments	21
C.1	Full results on JohanssonBench	21
C.2	Comparing model predictions with human annotations	22
C.3	Data experimental choices	23
C.4	Scaling properties: model and data	25
D	Qualitative Experiments	26
D.1	Novel form-from-motion stimuli	26
D.2	Psychology-inspired robustness checks	28
D.3	Generalization to animal motion by text prompting	28
E	Additional Related Work	28

A DATA

In this Section, we first describe how we construct JKP stimuli from raw RGB videos in NTU-60. Then, we compare various pose detection methods in this construction. Then, we describe details of the three splits. Finally, we conduct temporal sensitivity analysis of the dataset to measure how much of it can be solved by a single frame or shuffled frames.

A.1 CONSTRUCTION AND PRE-PROCESSING

From RGB video to Johansson keypoint video. Figure 1 illustrates the result of constructing a Johansson-like stimuli from RGB video: each frame of the video is reduced to a sparse set of keypoints on a black background. The process follows these steps.

- 1. Person-detector guided spacetime crop.** Each clip is cropped tightly around the actor using a *fixed* bounding box that is held constant across the entire clip (a spacetime crop). The procedure is:
 - Per-frame box.* We use standard YOLO person detector to get a bounding box in each frame. Joint locations are read from the accompanying skeleton
 - Temporal union.* The per-frame boxes are merged across all frames so that the actor never exits the crop at any point in the clip.
 - Margin expansion.* The union box is expanded so that the actor occupies approximately at least 60% of the crop area (the *target occupancy*), leaving a small margin around the extremities. The final maximal box is then used to crop the video.

2. **Letterbox to 832×480 portrait.** Before keypoint extraction, all cropped clips are padded to $480 \text{ px} \times 832 \text{ px}$ (width $\times$ height) portrait orientation adding black bars to fill the fixed canvas. This normalizes the spatial resolution seen by the pose estimator.
3. **Pose extraction and point-light rendering (DWPose).** Johansson keypoint videos are produced by **DWPose**, a two-stage skeleton estimator.
 - (a) A **YOLOX-L** person detector (`yolo_x_l.onnx`) localizes the actor in each frame.
 - (b) A **DW-LL** whole-body pose estimator (`dw-ll_ucoco_384.onnx`) predicts OpenPose-style body keypoints from the detected bounding box.

Keypoints are then rendered using the *Johansson preset*: face keypoints and hand keypoints are suppressed entirely. The keypoints are drawn as white filled circles (radius 8 px). Limb connections are removed. The result is a appearance-free point-light video in the spirit of the original Johansson displays (Johansson, 1973), containing only the structural and kinematic information of the moving body.

A.2 QUALITY OF POSES FROM DWPOSE

To get a sense of how robust the JKP videos are, we compare four different 2D pose extractors: DWPose (Yang et al., 2023), RTMPose-L, RTMPose-X (Jiang et al., 2023), and RTMO-L (Lu et al., 2024) on the a subset of NTU-CI consisting of $N=105$ videos ($N_c=5$ per class). Each method produced an OpenPose-18 keypoint sequence per video. Since no 2D pose ground truth is available, we measured pairwise agreement bidirectionally: treat method A as pseudo-ground truth for method B, repeat with B as pseudo-ground truth, then average both scores. We report torso-normalized AUC-PCK and mean OKS (higher is better), plus normalized MPJPE (lower is better). The methods show strong overall agreement as shown in Fig. 6. DWPose and RTMO-L agree most strongly: AUC-PCK 0.885, mean OKS 0.899, and normalized MPJPE 0.167. This indicates that the existing DWPose JKP controls are consistent with an independently designed, bottom-up 2D pose estimator. Thus, using DWPose is a reasonable choice as also verified by manual review.

AUC-PCK $\uparrow$ · avg. 0.841						Mean OKS $\uparrow$ · avg. 0.841						N-MPJPE $\downarrow$ · avg. 0.234					
DWPose	—	0.843	0.801	0.885	0.854	DWPose	—	0.846	0.798	0.899	0.848	DWPose	—	0.226	0.293	0.167	0.201
RTMPose-L	0.843	—	0.849	0.835	0.845	RTMPose-L	0.846	—	0.845	0.845	0.830	RTMPose-L	0.226	—	0.234	0.244	0.227
RTMPose-X	0.801	0.849	—	0.793	0.815	RTMPose-X	0.798	0.845	—	0.803	0.798	RTMPose-X	0.293	0.234	—	0.320	0.292
RTMO-L	0.885	0.835	0.793	—	0.885	RTMO-L	0.899	0.845	0.803	—	0.899	RTMO-L	0.167	0.244	0.320	—	0.140
Sapiens2	0.854	0.845	0.815	0.885	—	Sapiens2	0.848	0.830	0.798	0.899	—	Sapiens2	0.201	0.227	0.292	0.140	—
	DWPose	RTMPose-L	RTMPose-X	RTMO-L	Sapiens2		DWPose	RTMPose-L	RTMPose-X	RTMO-L	Sapiens2		DWPose	RTMPose-L	RTMPose-X	RTMO-L	Sapiens2

Figure 6: **Comparing pose extraction methods.** Pairwise agreement of 2D pose extractors on 105 NTU videos. Each score averages both pseudo-ground-truth directions; DWPose and RTMO-L show the strongest agreement.

A.3 SPLIT DETAILS

The detailed list of actions in each of the three evaluation splits is provided in Table 5. **NTU-CI** is a cross-identity split of NTU RGB+D (525 videos, 21 classes). **NTU-CA** is the held-out cross-action subset (250 videos, 10 classes). **UTD** is a cross-dataset split of UTD-MHAD (861 videos, 27 classes).

A.4 TEMPORAL SENSITIVITY ANALYSIS

Quantifying static bias in NTU-60. While NTU-60 (Shahroudy et al., 2016) is collected in a lab setting with fixed camera (similar backgrounds), it is still worth investigating how temporally sensitive the actions are. We experiment with $k \in \{1, 2, 4, 8, 16\}$ (number of frames uniformly

Table 5: Full list of action classes in each evaluation split. **NTU-CI** is a cross-identity split of NTU RGB+D (525 videos, 21 classes). **NTU-CA** is the held-out cross-action subset (250 videos, 10 classes). **UTD** is a cross-dataset split of UTD-MHAD (861 videos, 27 classes).

Split	Videos	Classes	Actions
NTU-CI	525	21	Pickup; throw; sitting down; standing up (from sitting position); clapping; hand waving; kicking something; hopping (one foot jumping); jump up; pointing to something with finger; cross hands in front (say stop); falling; punching/slapping other person; kicking other person; pushing other person; point finger at the other person; hugging other person; giving something to other person; handshaking; walking towards each other; walking apart from each other.
NTU-CA	250	10	Sitting down; standing up (from sitting position); clapping; kicking something; hopping (one foot jumping); jump up; falling; punching/slapping other person; kicking other person; walking towards each other.
UTD	861	27	Right arm swipe to the left; right arm swipe to the right; right hand wave; two hand front clap; right arm throw; cross arms in the chest; basketball shooting; draw x; draw circle (clockwise); draw circle (counter clockwise); draw triangle; bowling (right hand); front boxing; baseball swing from right; tennis forehand swing; arm curl (two arms); tennis serve; two hand push; knock on door; hand catch; pick up and throw; jogging; walking; sit to stand; stand to sit; forward lunge (left foot forward); squat.

sampled) evaluated with two models: Qwen3.5-9B and InternVL3-8B (Shangguan et al., 2025). We also experiment with and without random shuffling of the frames. Results are in Fig. 7 (a). Single-frame baseline is very poor indicating that actions need at least ≥ 2 frames. Different actions are more/less temporally sensitive as shown in Fig. 7 (c). `clapping` / `handwaving` can be recognized by a single frame, but something like `punch/slapping` improves with increased temporal context.

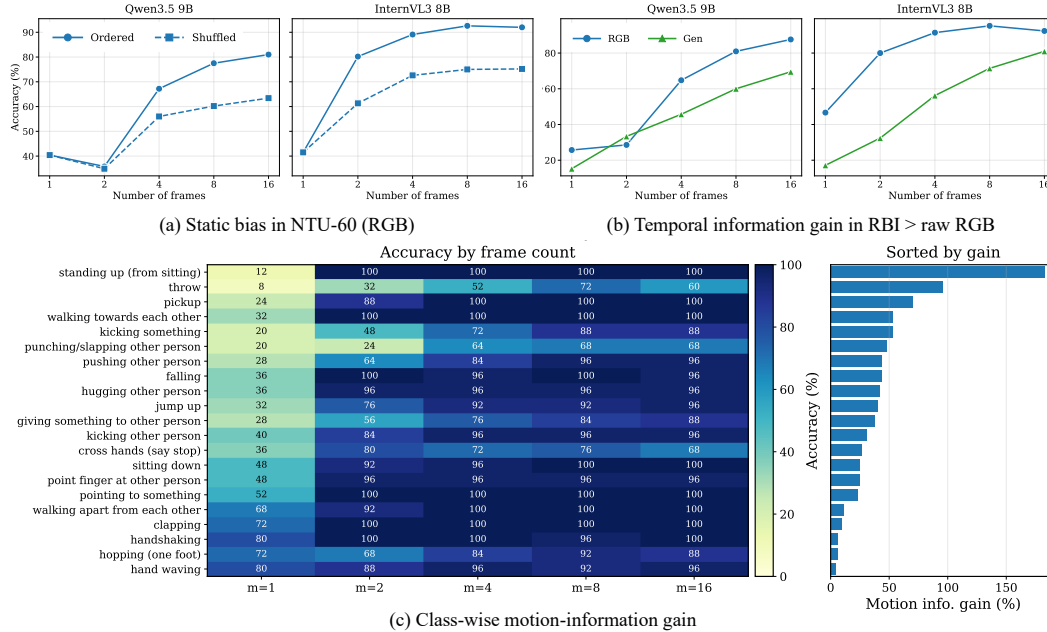


Figure 7: **Temporal sensitivity analysis.** We conduct temporal sensitivity analysis of the dataset to measure how much of it can be solved by a single frame or shuffled frames. (a) shows the raw RGB videos evaluated with varying number of frames with/without shuffling with two MLLMs; the task benefits with increasing temporal context and deteriorates under shuffling. (b) we repeat the same analyses but for RBI where we subsample fewer frames to be passed to the generation model. (c) We analyze class-wise performances to highlight classes that can be recognized by a single frame (e.g., hand waving) vs. those that improve with increasing temporal context (punching/slapping).

JKP is not solved simply by mapping the keypoints to a form in a single frame. To show this, we sample k frames from the input JKP video (say with n) frames and repeat the k frames (in order) to make a dummy new JKP video with n frames. We pass this as input to the RBI model. Then, from the generated video, we sample fixed $m = 16$ frames and pass to the MLLM for recognition. If RBI was solving the generation task frame-by-frame, we should see high performance even when $k = 1$. To quantify this, we define a new metric: *motion information gain* (MIG).

Let $A(k)$ denote model accuracy when using k uniformly sampled frames, evaluated at $k \in \{1, 2, 4, 8, 16\}$. For a doubling of the frame count from k to $2k$, we define the *relative gain* as

$$G(k \rightarrow 2k) = \frac{A(2k) - A(k)}{A(k)} \times 100\%. \quad (1)$$

We then define the *Motion Information Gain* (MIG) as the average relative gain across all consecutive doublings $k \in \{1, 2, 4, 8\}$:

$$\text{MIG} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} G(k \rightarrow 2k) = \frac{1}{4} \sum_{k \in \{1, 2, 4, 8\}} \frac{A(2k) - A(k)}{A(k)}. \quad (2)$$

Intuitively, MIG captures how much a model’s accuracy continues to benefit from additional temporal context (more frames), averaged uniformly across scales. A high MIG indicates that performance keeps improving with denser sampling (e.g., contact/interaction actions), whereas a low MIG indicates early saturation (e.g., pose-transition actions readable from a single frame). As shown in Table 6 and Fig. 7(b), RBI shows higher MIG than original RGB videos, i.e., it benefits more from motion information since it has to generate a complete video.

Table 6: **Motion Information Gain (MIG)** for RGB vs. RBI input, computed per Eq. (2) over frame-count doublings $m=1 \rightarrow 16$ (ordered, ground-truth frame order).

Model	Input	$G(1 \rightarrow 2)$	$G(2 \rightarrow 4)$	$G(4 \rightarrow 8)$	$G(8 \rightarrow 16)$	MIG
Qwen3.5-9B	RGB	11.1	126.7	25.0	8.2	42.8
Qwen3.5-9B	RBI	118.7	37.1	31.3	15.9	50.7
InternVL3-8B	RGB	71.4	14.3	4.2	-3.0	21.7
InternVL3-8B	RBI	88.9	73.5	27.1	13.3	50.7

A.5 EVALUATION WITH 3D MOTION CAPTURE DATASETS

We construct Johansson’s point-light stimuli from monocular 2D pose estimation (DWPose). In some complex actions, this may introduce detector jitter, self-occlusion artifacts, and depth ambiguities that could confound perceptual failures with detector noise. We augment our evaluation with that on two additional datasets based on 3D motion capture. Since these are directly captured in 3D, we simply render the keypoints on a 2D plane with an appropriate camera pose. Specifically, we use (i) a classical Johansson dataset by Vanrie & Verfaillie (2004), (ii) the CMU Motion Capture database (CMU Graphics Lab, 2003). Results are shown in Fig. 8.

B MODEL

In this Section, we provide remaining details for the model, training, inference and its hyperparameters. We also provide the text prompt used for the generation model as well as the downstream MLLM.

B.1 ARCHITECTURE AND TRAINING DETAILS

Following the notation introduced in the main paper, let $X_{\text{RGB}} \in \mathcal{X} = \mathbb{R}^{F \times H \times W \times C'}$ be an RGB video and $X_{\text{JKP}} \in \mathcal{X}$ be the corresponding JKP video. First, they are both encoded with the Causal VAE into latents $Z_{\text{RGB}, \text{JKP}} \in \mathbb{R}^{B \times S \times D}$, B is batch size, $S = F' \times H' \times W'$ sequence length, $D = 128$. Here,

$$F' = 1 + (F - 1)/8, H' = H/32, W' = W/32.$$

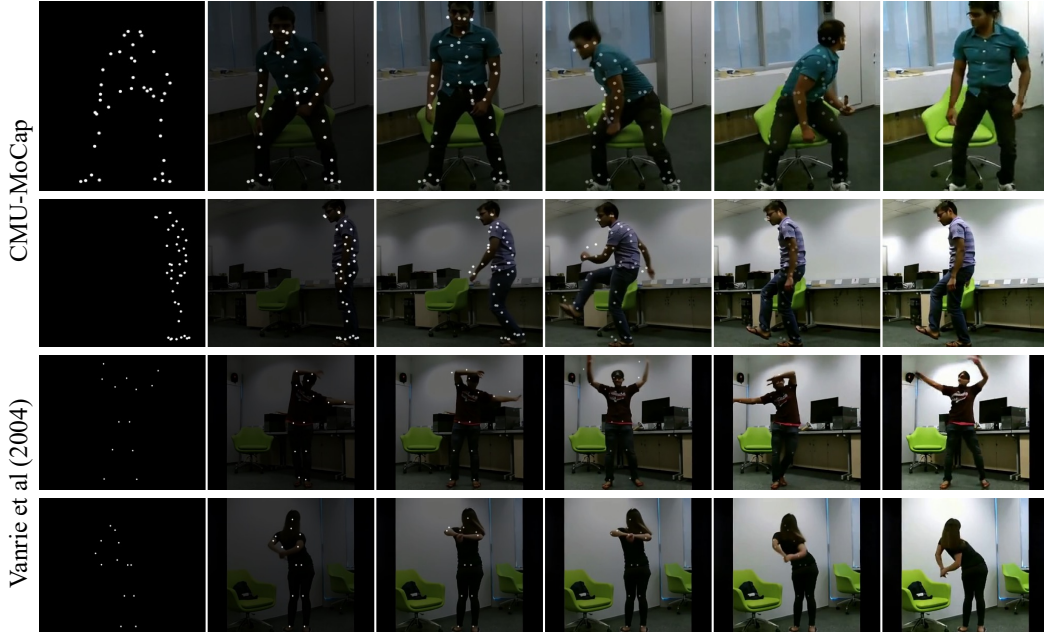


Figure 8: **Evaluating RBI on 3D motion capture samples.** In these examples, the input JKP is constructed by directly rendering the 3D points onto a 2D image. RBI works reasonably well off-the-shelf on such samples. On the left we show the input JKP image and on the right is RGB image. In between, we blend the images to demonstrate if the generated RGB matches the JKP.

For a 512×512 resolution video with 121 frames (4s at 30 FPS), we get $16 \times 16 \times 16 = 4096$ video tokens. The latents for the reference video and the target video are concatenated to obtain a sequence of shape $[B, 2S, D]$. The corresponding space-time positions for each of them are tensors of shape $[B, 3, S, 2]$ where the first dimension has components for (t, h, w) ; each video token has $(s_t, e_t), (s_h, e_h), (s_w, e_w)$ where s denotes the start and e denotes the end, i.e., each token corresponds to a space-time tube in some sense. The position tensors are also concatenated into $[B, 3, 2S, 2]$. Note that we do not have any image conditioning (e.g., first frame) apart from the JKP video. Text latents are constructed from all the layers of Gemma3-12B; see HaCohen et al. (2026) for more details on the text processing pipeline.

Training. The model is trained to match the velocity fields of the latents using a standard flow-matching loss (Lipman et al., 2022). Reference tokens are kept clean while target tokens are noised. Likewise, the loss is only applied to the target tokens.

Inference. The distilled 8-step sampler uses the default σ schedule from LTX-2. We define the size of the video to be generated based on the given reference video dimensions. Note that the distilled LTX-2 model halves the resolution in Stage 1. We can always upscale the reference video to get upscaled generated video. We do not use Stage 2 ($2\times$ spatial upsampler). Our default pipeline uses the original 832×480 resolution for reference videos. Usually, we observe that higher spatial resolution results in higher quality videos and also yield improved downstream performance.

Hyperparameters. The key LoRA hyperparameters are $r = 64$, $\alpha = 1$, and learning rate of $2e^{-4}$. We use batch size of 1 (to manage variable number of frames) with gradient accumulation steps 4. We attach LoRA matrices to all the linear layers (Q, K, V, O projections, and the two FFN layers in each of the 48 DiT blocks).

B.2 PROMPTS

Text prompt for generative model. We use a constant text prompt during training and inference (unless otherwise specified). The full prompt is provided below.

Observe the provided body keypoints and imagine a person(s) naturally performing the depicted movement. Full body visible, bright lightning, static camera.

Text prompt for MLLMs. Here, we provide the detailed prompt template used for action recognition with an MLLM. The action choices are filled based on the split used (Table 5). We use the same template for SFT.

System

You are a helpful and harmless assistant expert in analyzing videos.

User

<video> Watch the video carefully and identify the human action being performed. Select the single best matching action from the following {num_choices} choices:

{choices}

Assistant

{action_label}

Video

{video_path}

C QUANTITATIVE EXPERIMENTS

In this Section, we provide additional experimental results and analyses. First, we include the full results on all three splits. We also compare a different generative model (Wan 2.1) with LTX-2. Then, we analyze how the best model performs against human baseline through analyzing the confusion matrices. Then, we provide ablations on various choices in the data pipeline. We then analyze the scaling properties along both data and model dimensions.

C.1 FULL RESULTS ON JOHANSSONBENCH

Full zero-shot results. We include zero-shot results for all open and closed models evaluated at the time of writing in Fig. 9. Interestingly, advancing generations of a given model family (e.g., Qwen or Gemini) seem to regularly improve performance on the benchmark.

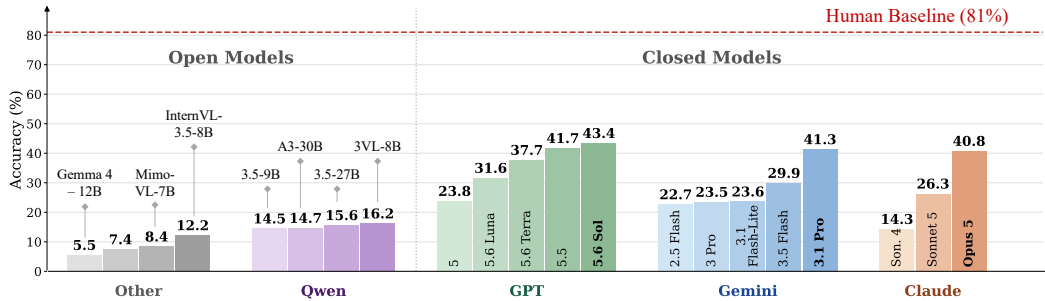


Figure 9: **JohanssonBench Results.** We evaluate several open/closed models on NTU-CI split of JohanssonBench. Top frontier models (e.g., GPT 5.6 Sol) significantly outperform open models but are about halfway mark compared to the human baseline.

SFT vs. RBI. The complete results w.r.t. the Table 2 in the main paper are provided in Table 7. We include few-shot ($k = 2$) baselines along with zero-shot numbers. In this case, we provide two randomly sampled JKP videos from the train set along with their labels in the prompt. While 2-shot baseline shows a tiny benefit over 0-shot with Qwen3.5-9B/InternVL3-8B, it hurts in case of Qwen3VL-8B/InternVL3.5-8B.

We also include evaluation with using a different generative model. We experiment with another open-source model, Wan 2.1. Since it is not straightforward to do in-context LoRA with this model, we use first frame conditioning with a random frame from the train set and fine-tune an image-video-to-video model building on AIGC-Apps (2026). Across all splits, On average, RBI (LTX-2) clearly outperforms RBI (Wan 2.1).

(a) **Cross-instance** (NTU: $N=525$, $C=21$)

<i>Input modality</i>	<i>RGB</i>	<i>Johansson's keypoints</i>					
Protocol	Zero-shot	Zero-shot	Two-shot	SFT	RBI (Wan 2.1)	RBI (LTX2)	Δ
Chance	4.8	—	—	—	—	—	—
Qwen3VL-8B-Inst.	76.6	16.2	14.1	61.3	64.6	67.8	+6.5
Qwen3.5-9B	81.0	14.5	18.1	70.9	70.1	71.8	+0.9
InternVL3.5-8B	84.4	12.2	7.8	77.1	73.9	76.6	-0.5
InternVL3-8B	92.0	7.4	8.4	80.4	80.2	80.4	+0.0
Average	83.5	12.6	12.1	72.4	72.2	74.2	+1.8

(b) **Cross-action** (NTU: $N=250$, $C=10$)

<i>Input modality</i>	<i>RGB</i>	<i>Johansson's keypoints</i>					
Protocol	Zero-shot	Zero-shot	Two-shot	SFT	RBI (Wan 2.1)	RBI (LTX2)	Δ
Chance	10.0	—	—	—	—	—	—
Qwen3VL-8B-Inst.	82.0	27.2	20.0	42.0	70.4	75.2	+33.2
Qwen3.5-9B	85.2	21.2	21.6	59.2	74.0	80.0	+20.8
InternVL3.5-8B	90.1	20.8	21.6	44.8	82.8	80.8	+36.0
InternVL3-8B	96.0	10.8	12.8	44.4	83.6	87.2	+42.8
Average	88.3	20.0	19.0	47.6	77.7	80.8	+33.2

(c) **Cross-dataset** (UTD: $N=861$, $C=27$)

<i>Input modality</i>	<i>RGB</i>	<i>Johansson's keypoints</i>					
Protocol	Zero-shot	Zero-shot	Two-shot	SFT	RBI (Wan 2.1)	RBI (LTX2)	Δ
Chance	3.7	—	—	—	—	—	—
Qwen3VL-8B-Inst.	25.6	3.1	5.1	8.0	23.8	24.2	+16.2
InternVL3-8B	28.9	2.6	2.7	17.0	20.9	21.8	+4.8
InternVL3.5-8B	29.0	2.3	2.1	14.8	20.8	20.9	+6.1
Qwen3.5-9B	43.4	4.1	5.8	15.4	29.8	32.4	+17.0
Average	31.7	3.0	3.9	13.8	23.8	24.8	+11.0

Table 7: We evaluate various MLLMs (in 8–9B parameter range) for action classification. RGB (zero-shot) denotes the upper bound evaluated for action classification on RGB videos. The columns under *Johansson's keypoints* are evaluated on JKP videos. We include zero-shot and few-shot baselines. SFT denotes supervised fine-tuning on JKP videos. RBI (Wan 2.1) and RBI (LTX2) denote our approach with different Diffusion backbones to first generate a RGB video from a JKP video; the RGB video is then passed to the MLLM (zero-shot). Δ is RBI (LTX2) – SFT. On the simpler cross-instance split (same actions across train and test), RBI outperforms SFT only slightly, but on the more challenging cross-action and cross-dataset splits, RBI generalizes substantially better than SFT.

C.2 COMPARING MODEL PREDICTIONS WITH HUMAN ANNOTATIONS

How much do the predictions from RBI or GPT 5.6 align with human judgement of actions from JKP stimuli? To answer this, we compare the confusion matrices on predictions with GPT 5.6 Sol and RBI (with Qwen3.5 9B) with that from human annotations. For this, we use the NTU-CI

split with $N=525$ videos. Human majority-vote annotations achieve 81.0% accuracy, compared with 43.4% for GPT 5.6 Sol and 71.8% for RBI. RBI also agrees more strongly with human predictions than GPT (62.7% versus 40.2%), with Cohen’s κ increasing from 0.37 to 0.61. Similarly, the correlation between the row-normalized confusion matrices is substantially higher for RBI than GPT (Pearson $r = 0.89$ versus $r = 0.68$). These results indicate that RBI not only improves classification accuracy but also produces a pattern of class confusions that more closely resembles human judgments.

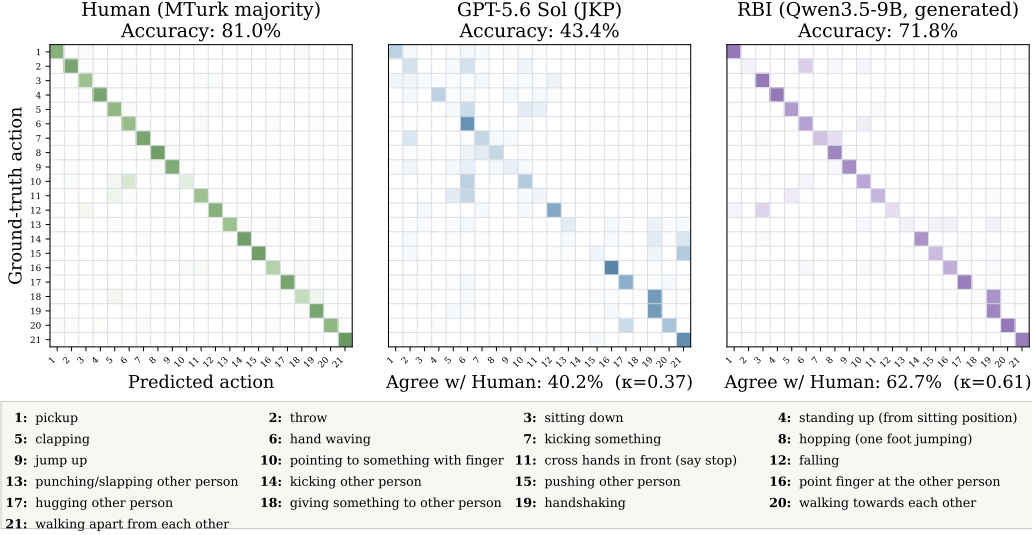


Figure 10: **Comparison of model predictions with human judgement.** RBI produces more accurate predictions and a confusion pattern more closely aligned with human annotations than GPT 5.6 Sol. Values are row-normalized by ground-truth class.

C.3 DATA EXPERIMENTAL CHOICES

We conduct a variety of ablations on various design choices in JKP evaluation. Given the JKP stimulus has a bunch of sparse keypoints, it is worth checking how visual patchification in MLLMs interacts with these keypoints (Section C.3.1). We find little-to-no effect on the final performance when varying the keypoint size or shapes. Next, in Section C.3.2, we vary the frame rate and the number of frames (with uniform sampling). Increasing either improves performance but saturates at a given point.

C.3.1 KEYPOINTS AND TOKENIZATION

Given the vital role that tokenization plays in Vision Transformers (Yuan et al., 2021), we conduct a systematic experiment where we modify the keypoint size/shape to accommodate it cleanly into standard tokenization routines. As shown in Fig. 11, we visualize different keypoint scales to fit to the standard ViT tokenization scheme. Results on subset of NTU-CI evaluation set are shown in brackets. For an input image with Johansson’s keypoints, default tokenization (a) can split up each keypoint across different patches. In (b), we reduce the size of the keypoints such that one keypoint lies inside one patch. In (c), we enlarge the keypoint to a square such that one keypoint is equivalent to one patch. While (a) and (b) perform similarly, using (c) leads to a drop in performance. Thus, we retain the default tokenization (a) for simplicity.

C.3.2 FRAME RATE AND NUMBER OF FRAMES

As shown in Fig. 12, temporal sampling and number of frames both matter on JohanssonBench. Accuracy rises steadily as more frames are drawn uniformly from each clip, from 2.9% at 2 frames to a peak of 15.6% at 32, with a small drop at 64 frames that suggests diminishing returns rather than a need for very long contexts. FPS sampling shows a sharper threshold: 1–2 FPS remains at

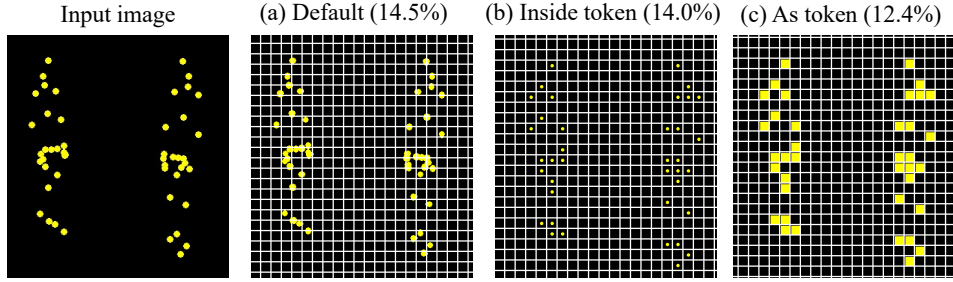


Figure 11: **Keypoints and tokenization.** We visualize different keypoint scales to fit to the standard ViT tokenization scheme. For an input image with Johansson’s keypoints, default tokenization (a) can split up each keypoint across different patches. In (b), we reduce the size of the keypoints such that one keypoint lies inside one patch. In (c), we enlarge the keypoint to a square such that one keypoint is equivalent to one patch. We show accuracy on each tokenization scheme. While (a) and (b) perform similarly, using (c) leads to a drop in performance. Thus, we retain the default tokenization (a).

7.2%, then jumps to 14.7% at 4 FPS and only modestly improves thereafter (15.6% at 16 FPS), indicating that a minimum temporal resolution is required to read the motion. In case of uniform frame sampling, increasing the frames smoothly improves performance before saturating around 32 frames. We use $F = 16$ frames as default for our experiments as using 2x frames improves performance slightly but doubles the compute.

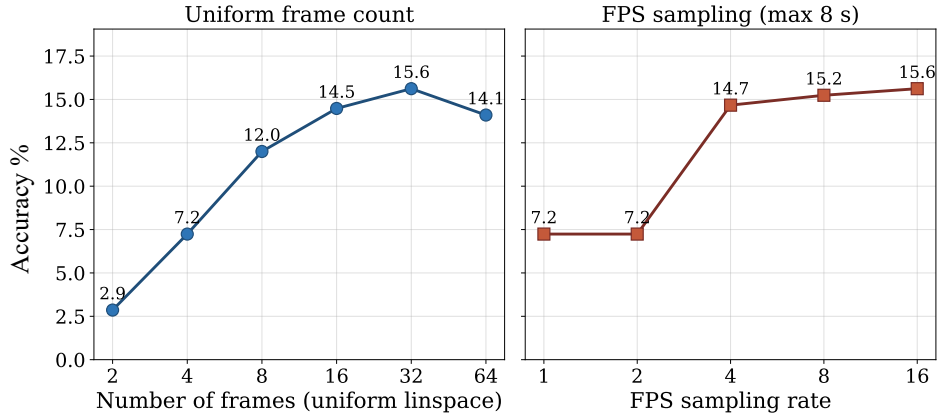


Figure 12: **Ablation on number of frames and FPS.** Performance improves smoothly with increased number of frames (uniform sampling) saturating around 32 frames. In case of fixed FPS sampling, performance only starts to improve with a high-enough temporal resolution (FPS = 4).

C.3.3 ABLATION ON KEYPOINT LOCATIONS

We evaluate RBI for robustness to the keypoint locations at test-time. Fig. 13 shows an example frame with three variants of the keypoint locations. In the third frame (from left), we place the keypoints in the middle of the line connecting to its parent keypoint. In the final image, we randomly vary the position of the keypoint along the line connecting to its parent keypoint. We randomize the positions in the first frame and then keep them in that position throughout the video.

C.3.4 VARYING NUMBER OF KEYPOINTS IN JKP

We also qualitatively analyze the number of keypoints in the input JKP videos. For that, we use samples from the CMU-Motion Capture dataset which has $k = 41$ keypoints by default. We lower k down to 6 and find that around $k = 14$ seems the minimum number of points for the model to infer accurate human pose sequence from the JKP input.

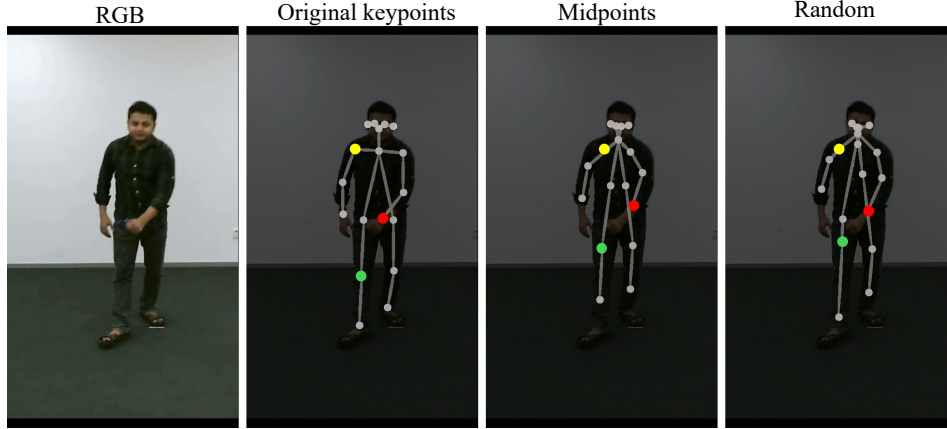


Figure 13: **Varying the keypoint locations.** We evaluate RBI for robustness to the keypoint locations. This figure shows a sample frame with three variants of the keypoint locations. The first image shows original RGB while the second one shows the original JKP frame with connecting lines shown only for reference and three keypoints highlighted. In the third image, we place the keypoints in the middle of the line connecting to its parent keypoint. In the final image, we randomly vary the position of the keypoint along the line connecting to its parent keypoint. We randomize the positions in the first frame and then keep them in that position throughout the video (i.e., we do not randomly change the relative position across frames).

Table 8: **Comparing base vs. distilled models at inference.** While we LoRA fine-tune the base model, at inference, we have the option of picking either. While base model results in slightly better performance than distilled (8-step) model, it is $\sim 17\times$ slower - thus, we use the distilled model. While LTX-2 offers an upsampling stage on top of the distilled model, we find that it often hallucinates and degrades performance quantitatively. LTX-2.3 does not result in an improvement over LTX-2 on our benchmark.

Model family	Checkpoint	Accuracy	Wall clock time (avg. per sample) (s)
LTX-2 (19B)	Base	72.2	102.2
	Distilled	71.8	6.7
	Distilled + LTX-upsampled	57.9	12.4
LTX 2.3 (22B)	Distilled	69.0	7.4
LTX 2.5 (22B)	Distilled	71.6	6.8

C.4 SCALING PROPERTIES: MODEL AND DATA

There are various axes to scale: dataset size, model size, model generations (of the same family). By model, here, we mean the downstream MLLM D_ψ and not the generative model G_ϕ . As for the generative model, LTX-2 only offers a single variant of 19B parameters.

First, in Section C.4.1, we look at scaling model size across

C.4.1 QWEN FAMILY SCALING

As shown in Fig. 16 (left), across the Qwen family, generation outweighs size: Qwen2-VL and Qwen2.5-VL stay near 5–7% even at 7B, whereas Qwen3-VL and Qwen3.5 leap into the mid-teens, with Qwen3-VL 8B at 16.2% and Qwen3.5 27B at 15.7%. Larger models generally help within a generation, but newer architectures give the biggest gains. Note that this evaluation is restricted to the NTU-CI split with $N = 525$ samples in the test set.

A similar model size scaling analysis is also conducted for Qwen3.5 across all three splits in Fig. 17. We also compare with SFT baseline with the same downstream model. On NTU-CI split (same train-

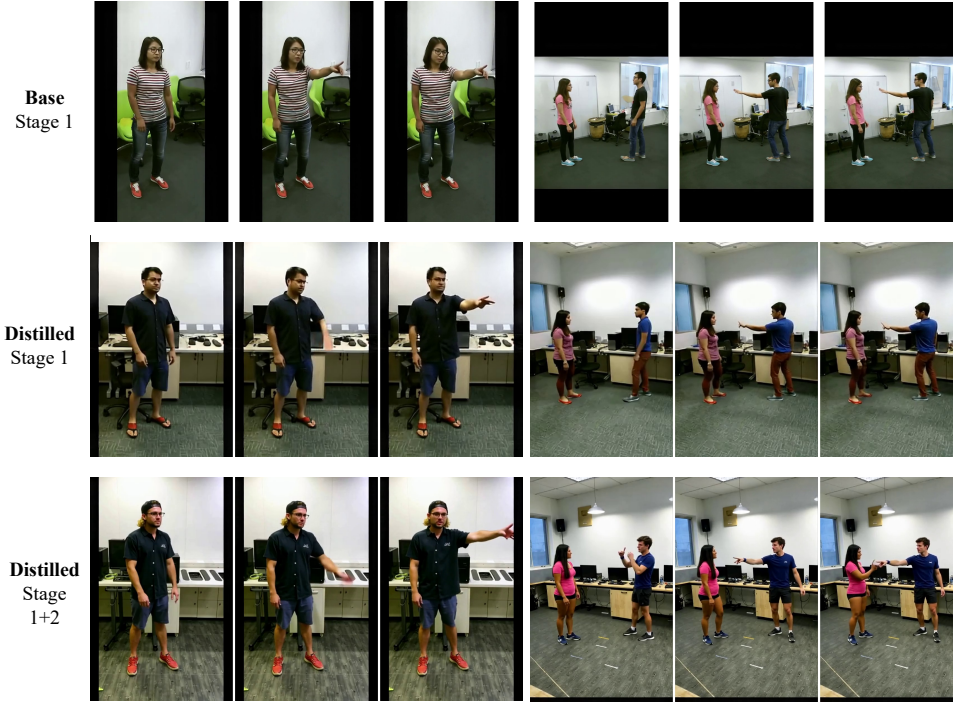


Figure 14: **Comparing base vs. distilled checkpoint at inference.** Using base (30-step generation) vs distilled (8-step generation) is similar in quality although the distilled checkpoint is 17x faster - so we use for inference. LTX-2 offers a upsampling stage for the distilled checkpoint but it often results in hallucinations and unnecessary details.

test actions), while SFT beats RBI in the small model range, RBI scales up faster to beat SFT in the large model range. On the harder NTU-CA and UTD splits, RBI clearly beats SFT across all model sizes and shows steeper scaling curves.

C.4.2 AMOUNT OF FINE-TUNING DATA.

In Fig. 16 (right), we vary the amount of fine-tuning data (JKP-RGB pairs) from the NTU-CI split, i.e., we vary the amount of data in training G_{ϕ}^* and evaluate with $D_{\psi} \in \{\text{Qwen3.5 9B, InternVL3-8B}\}$. We find a meaningful bump in performance even with just 5% of data (~ 700 examples) but it continues to improve with more data.

D QUALITATIVE EXPERIMENTS

In this Section, we present several qualitative evaluations with RBI. First, we show qualitative results on the Hidden bird illusion and the SpookyBench dataset. Then, we show results on several simple robustness test, e.g., JKP video that occludes part of the motion trajectory. Finally, we also show that RBI also works on animal motion with suitable text prompts.

D.1 NOVEL FORM-FROM-MOTION STIMULI

We consider entirely novel psychophysical stimuli where the central element is only perceptible by motion alone. First, we consider the hidden bird illusion (Bach, 2006). Second, we consider examples from the SpookyBench dataset (Upadhyay et al., 2026). For both of these examples, we observed that using the default LoRA strength of 1.0 does not yield satisfactory result. Thus, we increase the LoRA strength to 1.7 to emphasize more the fine-tuned knowledge over the base model prior. This resulted in more accurate generations.

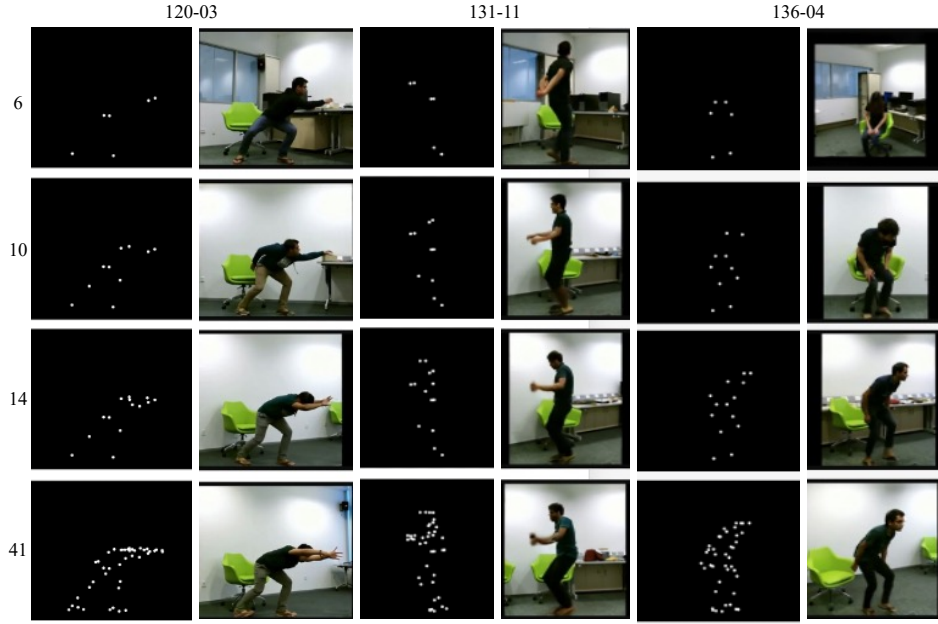


Figure 15: **Varying number of input keypoints.** We vary the number of keypoints in the input JKP on samples from CMU-MoCap. We find $k = 14$ to a (qualitatively) reasonable threshold which preserves motion information. For $k < 14$, we find various hallucinations in the pose (e.g., a standing pose imagined as a sitting pose).

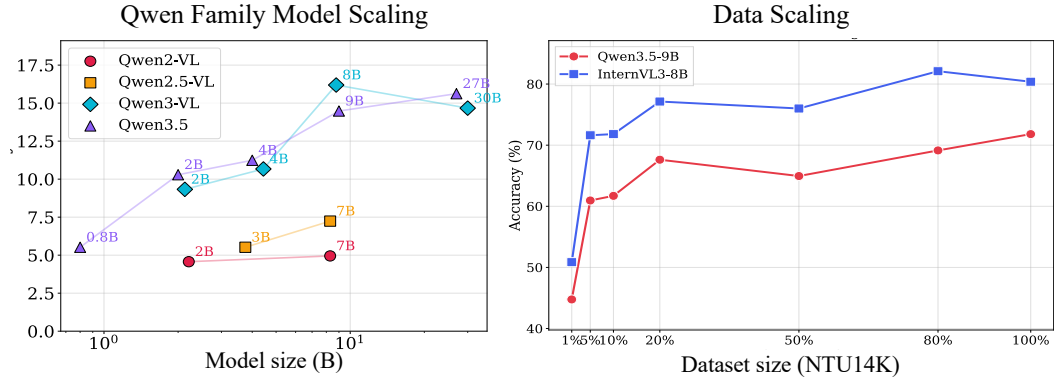


Figure 16: **Ablation on model size and training data %.** (Left) Varying model size across different generations of the Qwen family evaluated zero-shot. (Right) Varying the amount of fine-tuning data for RBI and evaluation across two downstream MLLMs (again, zero-shot).

Hidden bird illusion. As shown in Fig. 18, in this case, we consider a sketch of a bird with 21 strokes as foreground. Then, the same set of strokes are randomly arranged to create a background. We invert the colors to prepare a white-on-black video where the bird flies through the random-strokes background. We prompt RBI to detect and generate the flying bird. Without any training on such stimuli, RBI is able to render the flying bird in the right locations. This example is best viewed as an mp4 video (attached with Supplementary Material).

SpookyBench. Next, we also consider examples of motion noise captcha from SpookyBench (Upadhyay et al., 2026). It consists of shape/object/word rendered as noise against a background of dynamic noise. It is this motion of noise that lets the central element pop up. We input the raw noise video to RBI and prompt it to detect the hidden object/shape. To get a sense of the difficulty of this task, it is best to view examples as mp4 videos (attached with Supplementary Material). Results are shown in Fig. 19 and Fig. 20. In case of fixed objects (e.g., deer) or shape (e.g.,

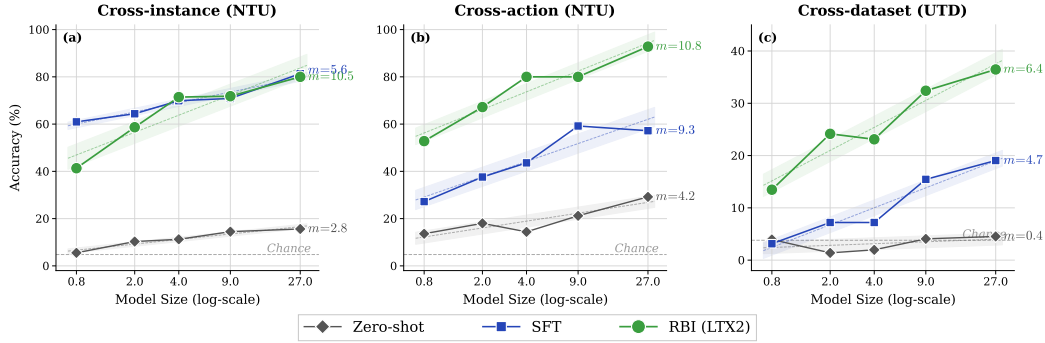


Figure 17: **Qwen3.5 scaling across evaluation splits.** Accuracy (%) vs. model size for zero-shot JKP (gray), JKP + SFT (blue), and RBI on generated RGB video (green) on (a) cross-instance NTU ($N=525$), (b) cross-action NTU ($N=250$, novel classes), and (c) cross-dataset UTD ($N=861$). Dashed lines: log-linear fits; shaded bands: $\pm$ RMSE.

heart), RBI is able to figure out its contour. Although, because it is fine-tuned on humans, it tries to fill in the shape with human-like elements (e.g., feet of the deer seem human feet, the horns seem like human arms). Funnily, in some cases (see Fig. 20 (right)), it incorrectly tries to approximate more complex shapes (e.g., penguin) with a human in a similar pose. Finally, we report that the model is unable to render text accurately although it does get the broad area of where the text is located but cannot render the letters.

D.2 PSYCHOLOGY-INSPIRED ROBUSTNESS CHECKS

We also conduct several simple robustness tests inspired by analogous experiments with humans. For instance, experiments with turning the JKP input upside-down leads to degradation in human ability to recognise the action (Troje & Westhoff, 2006; Pavlova & Sokolov, 2000). RBI also struggles with such an input - it incorrectly generates an upright human with unusual limb lengths. Likewise, we also test with an occlusion which hides some keypoints in the middle of the video (Graf et al., 2007). RBI is quite robust to this occlusion and generates the walking person as if there was no occlusion. Interestingly, this also matches human observation. We also test with temporally shuffled frames which looks like a randomly sped up video. RBI generates multiple persons in this case. Finally, we also check the case of a person walking towards or away from the camera. The original input is of a person walking towards the camera - RBI works perfectly in this case. To simulate person walking away from camera, we temporally reverse the original keypoint video. As a human, we could sense the discrepancy in the time-reversed video. RBI interestingly also seems confused - it generates a person walking behind but with facing the camera.

D.3 GENERALIZATION TO ANIMAL MOTION BY TEXT PROMPTING

While RBI is trained on human keypoint videos, can it also generalize to animal keypoint motion? This has also been an interest of study for psychologists (Mather & West, 1993) with human experiments. Much like humans, we find that RBI too generalizes quite well to animal keypoints, although it requires careful text prompt to generate realistic animal motion. Results are shown in Fig. 21.

E ADDITIONAL RELATED WORK

Owing to the limited space in the main paper, we discuss some additional related work in human vision below.

Why does the human brain perceive JKP coherently? One may ask, why do we perceive such a sparse stimuli so effortlessly? We argue that JKP is a minimal, extremely compact representation of human (and animal) motion (Johansson, 1973; Blake & Shiffrar, 2007). This is important in a predatory-prey situation where one may want to quickly observe and infer the action of the other

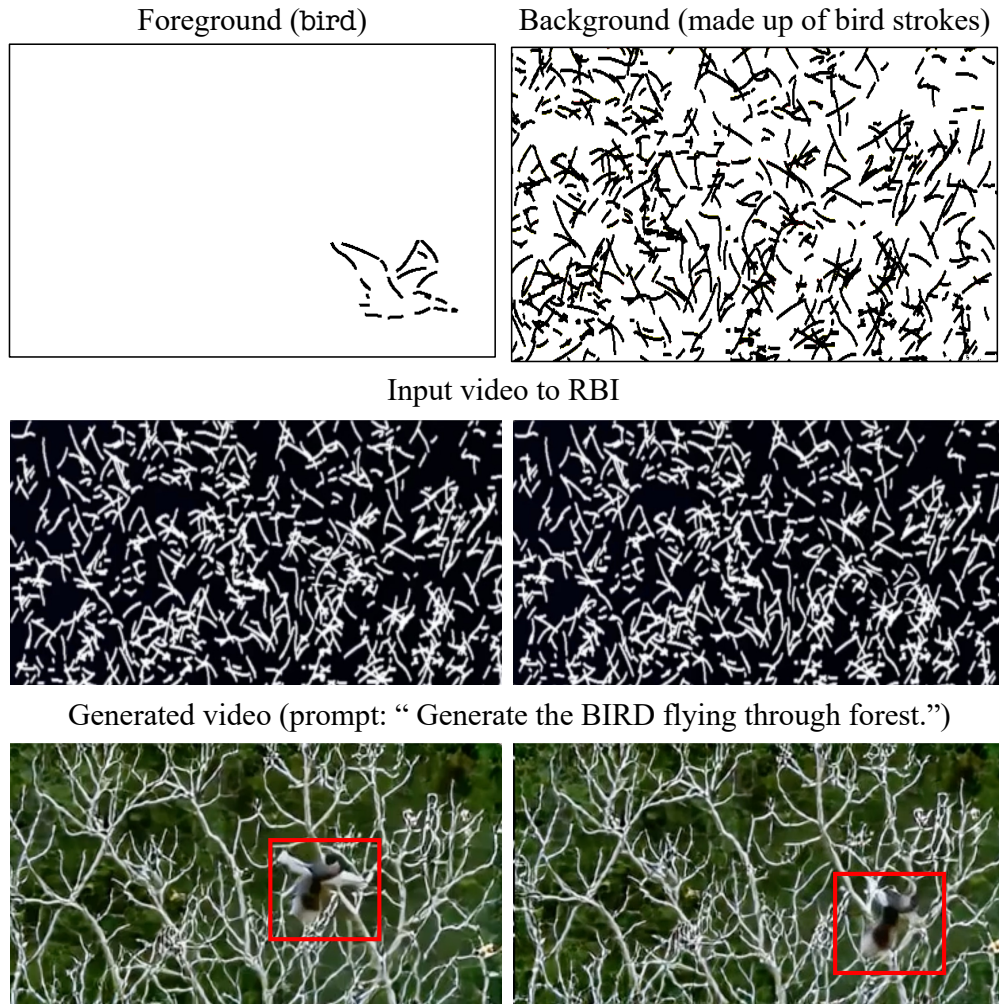


Figure 18: **Bird’s illusion.** See <https://michaelbach.de/ot/cog-hiddenBird/index.html> for an intuitive example. We consider a sketch of a bird with 21 strokes as foreground. Then, the same set of strokes are randomly arranged to create a background. We invert the colors to prepare a white-on-black video where the bird flies through the random-strokes background. We prompt RBI to detect and generated the flying bird. Without any training on such stimuli, RBI is able to render the flying bird in the right locations.

(Johnson, 2006). Perhaps, this compact representation may also help us perceive actions in low light conditions, or at a distance, or when the visual signal is poor or occluded (Grossman & Blake, 1999; Cutting et al., 1988). Under these conditions, an appearance-free, purely motion-centric representation is likely to be robust.

How does the human brain perceive JKP? Neuroscience work suggests this recognition relies on a dedicated brain region (posterior STS) that acts as a detector specialized for spatiotemporal point patterns, responding to organized joint motion but not to scrambled versions of the same motion vectors (Grossman et al., 2000; Grossman & Blake, 2002). This is separate from generic optical-flow/motion areas (analogous to MT/MT+ in vision models), indicating the brain treats “structured keypoint motion” as its own category rather than folding it into general motion estimation (Grossman & Blake, 2002). Furthermore, the recognition pipeline is tuned to sparse, skeleton-like inputs rather than dense pixel-level appearance (Puce & Perrett, 2003). Behavioral studies show humans build action categories directly from these sparse spatiotemporal keypoint trajectories (Hemerik, 2008), and computational models formalize this as a hierarchy: local motion/keypoint features are

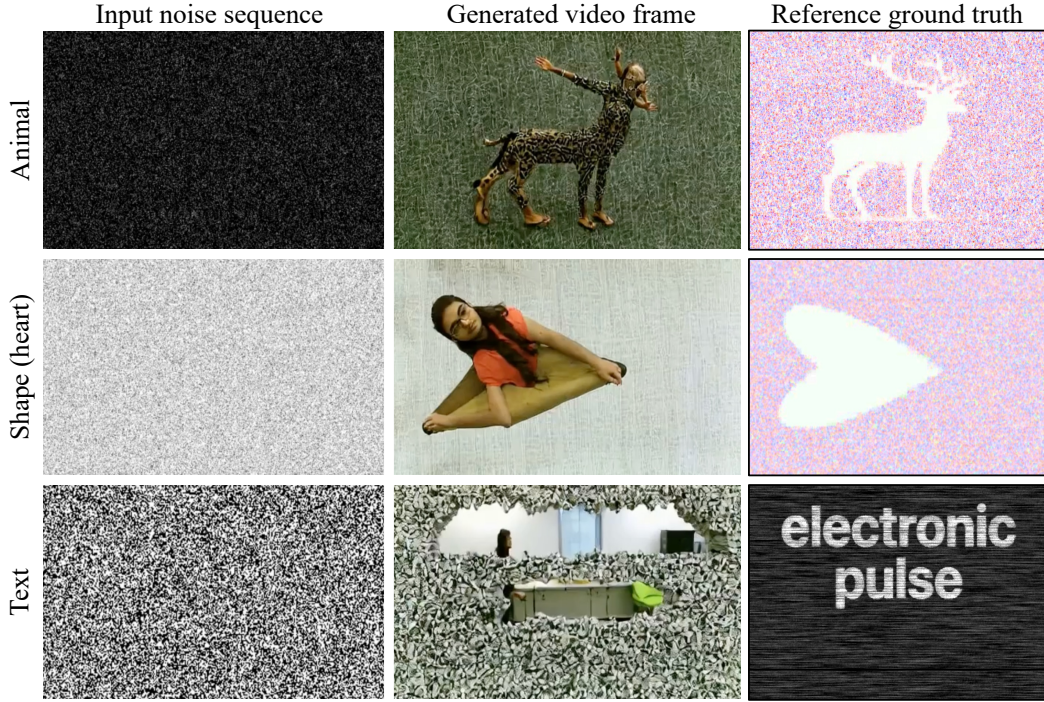


Figure 19: **Examples from SpookyBench.** Ground truth reference is obtained by simple operations (e.g., normalized temporal difference) to highlight the signal from the noise. RBI accurately identifies the shape (animal, heart) through its motion perception. It fills in with a bizarre-looking object - but the highlight is that it identifies and separates the signal (shape) from noise. For the case of text rendering, it is unable to get the letters right but only gets the broad area where the text is located.

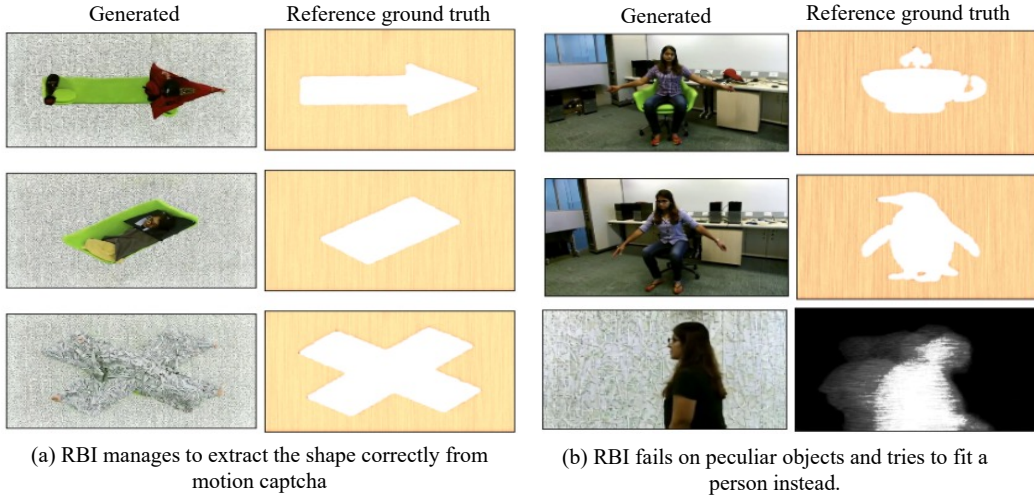


Figure 20: **Some more successful and failed examples from SpookyBench.** RBI is able to detect simpler shapes such as an arrow, parallelogram or cross. In case of more complex objects like a penguin, it tries to approximate it by a person in a weird pose. It also does not work as well when the foreground object/person is moving very fast or when the camera moves rapidly.

grouped bottom-up into body structure, refined by learned, top-down priors about plausible human movement (Giese, 2015). Another line of work suggests that the observer’s motor system (premotor, primary motor, and parietal cortex that plan and control one’s own body movements) gets activated

Prompt: “Observe the point-tracks of a **person and a dog**”.



Prompt: “Observe the keypoints of a **dog** in this video.”



Prompt: “Observe the keypoints of a **horse in a green field**.”



Figure 21: **More qualitative results with point-tracking and varying text prompts.** We investigate if RBI (trained on human keypoint videos) also generalize to animal keypoint motion. With the right text prompt, we find the model does generate accurate realistic motion of animals.

while perceiving JKP videos (Saygin et al., 2004; Saygin, 2007; Calvo-Merino et al., 2005). This suggests that observers may recognize an action in part by internally simulating how they would perform it, rather than by classifying the visual input directly.

Visual generation for perception tasks. As generative models have improved enormously, they are increasingly being used as data synthesis engines for training perception models. Diffusion models have already shown promising performance on several visual recognition tasks (Azizi et al., 2023; Gatti et al., 2026). The imagine-then-act theme is also widely popular in robotics (Bharadhwaj et al., 2024; Hu et al., 2024; Liang et al., 2024). Similarly, world models continue to be widely used (Agarwal et al., 2026). In this work, while we use a video generation model in a similar spirit, it is not to generate synthetic data, or simulate the future, but rather to solve a purely visual motion perception task grounded in psychophysics.

Controlled video generation. There is also a lot of related work on generating raw video from skeleton pose maps for animation (Wang et al., 2026; Zhang et al., 2024b). A key difference with our work is that skeleton pose maps reveal the *form* (person) in a single frame whereas in case of JKP, one needs to figure out the shape from motion. Going from point-tracks to RGB (Geng et al., 2025; Namekata et al., 2026) or vice-versa (Hedlin et al., 2024; Shrivastava et al., 2026) are related tasks. We do not use dense point tracks but instead consider a sparse, minimal set of keypoints sufficient to perceive biological motion. In the future, we will explore more generalized keypoint-to-video generation setup where keypoints are obtained with point trackers.

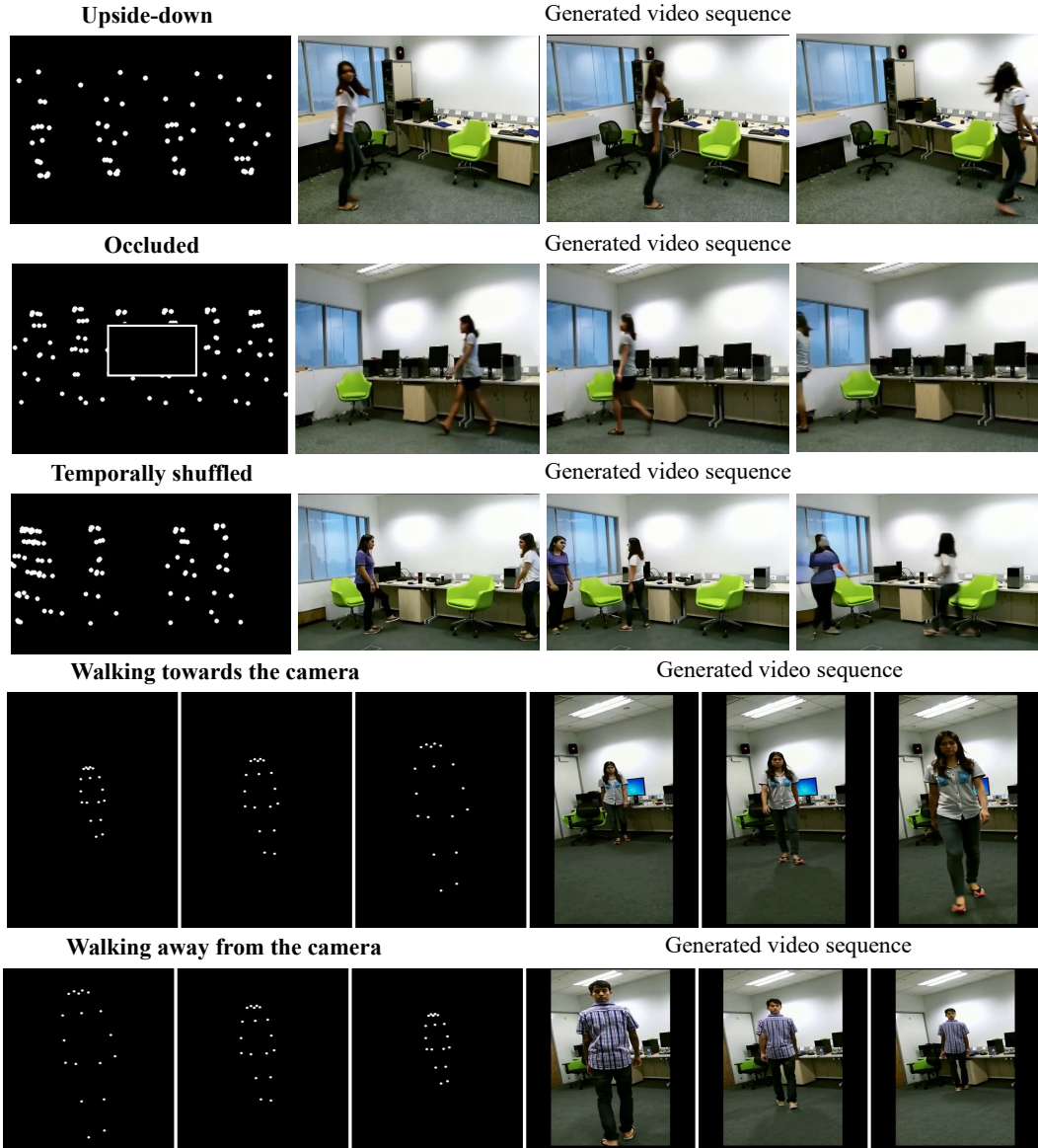


Figure 22: **Robustness checks.** We conduct several simple robustness tests inspired by analogous experiments with humans, e.g., upside-down or occluded input or change in the scale of the actor (by simulating walking towards / away from camera). By and large, RBI’s generations tend to match human judgement in these experiments (Pavlova & Sokolov (2000); Graf et al. (2007)).