InterPet4D: A Multimodal 4D Human-Pet Interaction Dataset for Pet Motion Generation

Yichen Peng*[1], Jyun-Ting Song*[1,2], Chen-Chieh Liao*[1],
Kris Kitani[2], Hideki Koike[1], and Erwin Wu[1]

[1] Institute of Science Tokyo
[2] Carnegie Mellon University

Fig. 1: The **InterPet4D** multimodal Human-Pet Interaction Dataset.

Abstract. Human-pet interaction estimation and generation remain underexplored due to the absence of high-quality large-scale dataset. We present **InterPet4D**, the first multimodal dataset capturing natural interactions between humans and dogs. Using a synchronized multi-view capture system, we record human–dog obedience tasks and provide annotations for both humans and dogs, including multiview and egocentric videos, segmentations, 2D/3D keypoints, meshes, and audio tracks. **Interpet4D** consists of 6.8 million frames collected from 13 dogs of 11 breeds interacting with 23 human participants. We further introduce the **InterPetMoGen** framework for human-pet interaction motion generation. Our proposed model achieves an FID score of 11.21, substantially outperforms the Seq2Seq or DiT baselines, demonstrating the effectiveness of Interpet4D for modeling realistic human–pet interactions.

* Equal contribution.

1 Introduction

Human–animal interactions exhibit temporally coordinated and mutually responsive movement patterns, where the motion of one agent directly influences and adapts to the other. Modeling such structured relationships is fundamental for understanding cross-species behavior and has important applications in socially aware robotics, virtual agents, animation, and behavioral analysis.

However, most existing research on interactive behavior has focused on human–human or human–object interactions, largely due to the absence of large-scale and realistic human-animal datasets. Capturing such interaction data presents unique challenges, as it requires structured and repeatable coordination between human and animal participants, which is difficult to achieve without controlled training, making large-scale data collection particularly challenging. Furthermore, close-range human–animal interactions can result in severe cross-occlusion, which degrades the reconstruction of fine-grained details on the participants.

To address these challenges, we introduce **InterPet4D**, the first large-scale multimodal 4D dataset of naturalistic human-dog interactions. Our dataset captures synchronized multiview, egocentric, audio, and 3D motion data from 23 participants and 13 dogs. We design a systematic interaction protocol covering 4 categories of common interactions: *petting*, *commanding*, *calling*, and *free-form*, enabling structured analysis of dog behavior. Figure 1 provides an overview of the InterPet4D dataset, including multiview video, egocentric video, audio commands, and reconstructed human and dog motion. To facilitate future research, we release InterPet4D on Hugging Face: `https://huggingface.co/datasets/ohicarip/interpet4d`.

Beyond the dataset, we propose **InterPetMoGen (IPMG)**, a Motion GPT-based framework for gesture-to-pet motion generation. Given a sequence of human hand/body gestures and accompanying audio, **IPMG** generates plausible 3D dog motion responses conditioned on human gestures and audio. The model adopts a MotionGPT-style autoregressive transformer that predicts discrete pet motion tokens learned by a PetVAE tokenizer. We further introduce modality-aware attention (MAA) masks that enable coarse-to-fine motion generation by combining bidirectional conditioning with causal autoregressive decoding.

Our contributions are summarized as follows:

- We introduce **InterPet4D**, the first large-scale multimodal 4D dataset of human–pet interactions, containing 6.8M synchronized frames with multiview RGB video, egocentric video, audio, and reconstructed 3D motion of both humans and dogs across 23 participants and 13 dogs of 11 breeds.
- We design a systematic interaction protocol covering 4 categories of human–dog interactions with standardized annotation pipelines, enabling structured analysis of cross-species behavior.
- We propose **IPMG**, a MotionGPT-based framework including a PetVAE tokenizer and MAA module that generates diverse and realistic dog motion responses conditioned on human gestures and audio signals.

2 Related Work

2.1 Human-centered Interaction Datasets

The study of human interactions with the physical world has progressed rapidly. GRAB [41] captures whole-body grasping of objects with detailed hand motion. BEHAVE [2] provides multi-view recordings of humans interacting with rigid objects. ARCTIC [8] focuses on articulated object manipulation with dexterous hand motion. For human–human interaction, InterHuman [17] proposes a large-scale dataset and a diffusion-based model for two-person motion generation. BUDDI [24] reconstructs 3D human–human close interactions from monocular images. These works demonstrate that modeling interactive dynamics requires capturing the joint distribution of all interacting agents. Our work extends this principle to the human-pet domain, where the morphological asymmetry between human and animal introduces additional challenges.

2.2 Animal Pose and Shape Estimation

Estimating the 3D pose and shape of animals has attracted growing attention. SMAL [48] introduces a parametric 3D body model for animals learned from toy figurines. Subsequent works focus on dogs, including WLDO [3], BARC [32], and BITE [33], which improve monocular 3D reconstruction using breed priors and contact constraints, as well as DogMo [43], which reconstructs dog motion from monocular videos, and AnimalAvatar [34], which reconstructs animatable 3D animals and motion from videos. RigAnything [20] further enables automatic skeletal rigging for arbitrary animal meshes. For pose estimation, DeepLabCut [23] provides a widely-used markerless framework across species, while RatBodyFormer [11] applies transformer models to estimate rodent body pose. On the data side, Animal Kingdom [25] and COP3D [38] provide large-scale animal video datasets for recognition and 3D reconstruction. However, existing datasets mainly focus on single-animal settings and rarely capture human–animal interactions or multimodal signals such as audio and human motion.

2.3 Conditional Motion Generation

Generating human motion conditioned on various signals has seen remarkable progress. Action-conditioned methods such as ACTOR [29] use VAE-based architectures for class-conditional generation. Text-conditioned approaches have flourished with MDM [42], which applies diffusion models to motion generation from text descriptions, and T2M-GPT [46], Duolando [39], MDLS [6], which uses VQ-VAE with GPT-based autoregressive generation. MotionDiffuse [47], FloodDiffusion [5] enable fine-grained text-driven motion synthesis. MoMask [9] introduces masked modeling for efficient motion generation. In the audio domain, co-speech gesture generation methods [18, 19, 28, 45] synthesize body and hand gestures from speech audio. MotionGPT [13] treats motion as a language and unifies multiple motion tasks in a single framework. However, all existing

methods operate within a single species while our work pioneers the cross-species setting, generating animal motion conditioned on human multi-modal signals.

3 InterPet4D Dataset

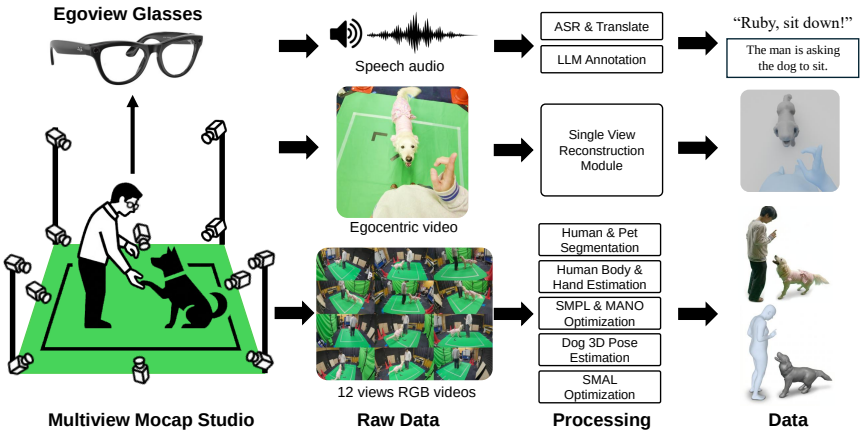

Fig. 2: Data collection setup. Our capture environment consists of 12 synchronized third-person RGB cameras arranged in a square, complemented by Rayban Meta Glasses worn by the participant providing an egocentric view and audio.

3.1 Data Collection Setup

We establish a multi-sensor capture environment to record synchronized multi-modal data of human-pet interactions. The capture space is a 6m×6m area.

Multi-view cameras studio. We deploy 12 wire-synchronized GoPro Hero13 cameras (1920×1080, 60 fps) arranged in a square configuration around the capture area, providing comprehensive multi-view coverage of both the human and the dog, as shown in Figure 2. Different from traditional human-target mocap studio, 8 out of 12 cameras are placed in the lower-ground area to achieve a better view of the pet and human's hand.

Egocentric view and audio. In addition to the third-person cameras, each participant wears a pair of Ray-Ban 2 Meta glasses, which provide an egocentric RGB video and audio stream capturing the first-person perspective of the interaction. The videos are synchronized with the GoPro cameras through software-based alignment using a hand-clapping signal, resulting in a maximum synchronization error of one frame. This egocentric view is particularly valuable for capturing close-range hand-pet interactions that can't be recorded from the 3rd-person-views.

Participants. To ensure stable interactions, we collaborated with a local pet training agency and recruited 11 experienced trainers together with different

Table 1: InterPet4D dataset statistics. Our dataset provides synchronized multimodal 4D data of human–dog interactions across diverse participants, breeds, and interaction types.

Data Attribute	Value
No. of dogs	13
No. of dog breeds	11 (+2 puppies)
No. of human	23 (incl. 11 trainers)
Third-person cameras	12 (1920×1080, 60 fps)
Egocentric camera	1 (1200×1600, 60 fps)
Total recording	161 sessions
Total frames	6.83M
Human representation	Pose, SMPL-X, MANO
Dog representation	Pose, SMAL
Audio	Aligned voice command
Text Caption	Annotated each clip

Table 2: Comparison with other Interaction or Pet datasets. InterPet4D is the first dataset providing synchronized human and pet 3D interactive motion with audio.

Dataset	Subjects	No.Human	No.Pet	Audio	Interactive	Views	Frames	Public
InterHuman [17]	Human	12 pairs	–	–	✓	76	107M	✓
Seamless Inter. [1]	Human	4000+	–	✓	Speech-only	1	400M+	✓
CoP3D [38]	Dog & Cat	–	4200	–	–	1	0.6M	✓
DogMo [43]	Dog	–	10	–	–	5	1M	–
InterPet4D	Human+Dog	23	13	✓	✓	12+Ego	6.8M	✓*

*The full dataset including annotations will be released upon publication.

breeds of trained dogs. In addition, 12 volunteers participate in the recordings, resulting in a total of 23 human subjects. Two untrained puppies were also involved for diversity, which results in 13 dogs spanning 11 breeds, covering a wide range of body sizes, as shown in Table 1. All adult dogs were trained in basic obedience tasks such as sitting, turning, and calling gestures, enabling consistent execution of the interaction protocol and still allowing natural behavior.

3.2 Interaction Taxonomy

We design a structured interaction protocol covering four categories of common human-dog interactions, each emphasizing different communication modalities:

1. **Petting** – The participant physically touches or strokes the dog, involving close-range hand motion and the dog's postural response (*e.g.*, leaning in, tail wagging).
2. **Commanding** – The participant issues verbal and gestural commands (*e.g.*, "sit", "stay", "shake", "turn around"), and the dog responds with trained behaviors.
3. **Calling** – The participant calls the dog from a distance using voice and hand beckoning repetitively, capturing the dog's locomotion and attention shifts.

4. **Free-form** – Unscripted interactions allowing participants to interact naturally, capturing the full diversity of everyday human–dog dynamics, including fetch, tug-of-war, and chase.

Each participant–dog pair performs 3-4 sessions per category, with each session lasting about 60 seconds.

3.3 Dataset Statistics

InterPet4D is the first dataset to provide multimodal recordings of human–pet interactions with synchronized 3D motion for both agents. Table 1 summarizes the key statistics of InterPet4D. To enable fair and reproducible benchmarking, we provide a fixed 80:20 train/validation split. All dogs appear in both sets, resulting in 200 training clips and 40 validation clips. Table 2 compares InterPet4D with other human-interaction and pet datasets. For more details, such as dataset analysis and examples, please check the supplementary documents.

4 Data Processing Pipeline

4.1 Human Reconstruction

We represent humans using the SMPL [21], SMPL-X [27], and MANO [31] parametric models. Human reconstruction is decomposed into body and hand reconstruction.

Body Reconstruction The human body pose is first estimated from multi-view third-person cameras and fitted to the SMPL [21] model. We first localize the human in each view using Mask R-CNN [10] with DeepSORT tracking [44], followed by manual filtering to remove failure cases. We then estimate 2D body keypoints using HRNet-WholeBody [14] for each detected bounding box and triangulate the multi-view 2D keypoints to obtain 3D joint estimates. To further refine the 3D body pose, we minimize an energy function that incorporates body symmetry, temporal smoothness, and temporal bone-length constraints [15]. Finally, an SMPL [21] body model is fitted to the refined 3D joint estimates to recover the full-body pose and mesh for each frame.

Hand Reconstruction Hand pose is estimated using a pipeline similar to body reconstruction. We first localize the hands in each view by running YOLO [12] and comparing the detections with the reprojected SMPL hand mesh. Using the resulting close-up hand crops, we apply WiLoR [30] to estimate 2D hand keypoints in each view. Following the same scheme as body reconstruction, the multi-view 2D hand keypoints are triangulated to obtain initial 3D estimates and then refined using the same energy formulation as in body reconstruction, augmented with a visibility-aware optimization [37, 40]. Finally, we fit MANO [31] to the refined 3D hand keypoints to recover hand pose, and fit SMPL-X [27] to the combined body and hand keypoints to recover the full-body.

4.2 Dog Reconstruction

We represent dogs using sparse 3D keypoints and the SMAL [48] body model. For each view, we first detect the dog using RTMDet [22] and estimate 2D dog keypoints using HRNet-W32 [7]. We then run Mask R-CNN [10] to segment humans, and use the resulting human masks to filter out 2D dog keypoints that fall in human-occluded regions. The filtered 2D keypoints from multiple views are triangulated to obtain per-frame 3D joint estimates, augmented with an RTS smoother [35] to improve temporal consistency. Finally, we fit the SMAL [48] model to the reconstructed 3D keypoints as SMALify [4]. Since mesh fitting from sparse keypoints is highly underconstrained, we predefined the shape parameters by manually fitting SMAL [48] to a single representative frame in a canonical pose to stabilize the reconstruction.

4.3 Audio and Text

Audio Extraction & Translation Raw audio is captured by the built-in microphones of the Ray-Ban Meta glasses. Among the 13 dogs, 11 are trained with Japanese commands, one with English commands, and one with Mandarin Chinese commands. To unify the language modality, we use an automatic speech recognition (ASR) system [36] to transcribe the spoken commands and translate them into English. In addition, the temporal boundaries of each command are automatically annotated using Qwen3-ForcedAligner [36].
Interaction labels. Each sequence is annotated with an interaction category (Section 3.2) and a textual description of the human–pet behavior (e.g., "The man first calls the dog, then commands the dog to sit and turn around."). The textual annotations are automatically generated from the English transcripts of the recorded audio using a Qwen3-based large language model [36] and only include descriptions of the verbal actions.

5 InterPetMoGen: Human-to-Pet Motion Generation

5.1 Problem Formulation

Given a human motion sequence $\mathbf{H} = \{h_t\}_{t=1}^{T}$, where h_t contains the 3D joint positions and rotations of J_b body joints and J_h hand joints at frame t, and a corresponding audio feature sequence $\mathbf{A} = \{a_t\}_{t=1}^{T}$, where $a_t \in \mathbb{R}^{d_a}$ denotes the audio feature at frame t, Our goal is to generate a pet motion sequence $\mathbf{P} = \{p_t\}_{t=1}^{T}$, where $p_t \in \mathbb{R}^{J_p \times 3}$ represents the 3D positions of J_p pet joints.
To better capture the distinct characteristics of human body and hand movements and learn their relationship towards pet reaction, we propose **InterPetMoGen**. We represent body and hand as two separate streams and tokenize them independently in our model. The mapping from human motion and audio to pet motion is inherently one-to-many: the same human input may correspond to multiple plausible pet responses. Therefore, we aim to learn the conditional distribution $p(\mathbf{P} \mid \mathbf{H}, \mathbf{A})$ rather than a deterministic mapping.

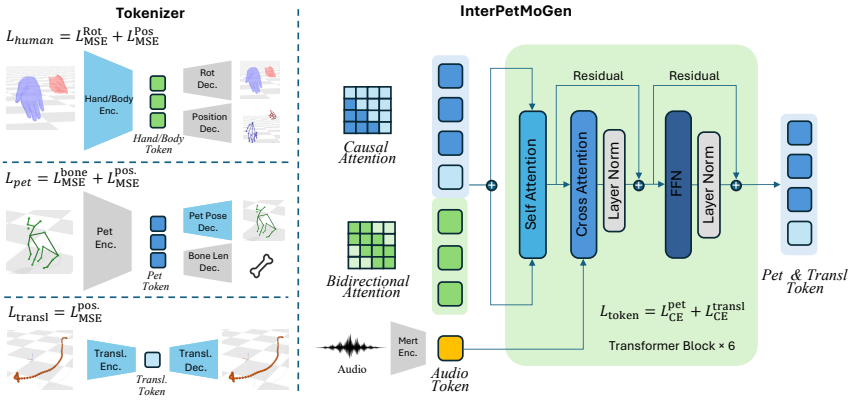

Fig. 3: InterPetMoGen architecture. Human gesture tokens and audio features are encoded by modality-specific encoders and fused through an autoregressive transformer. The model generates global translation and pet motion tokens conditioned on human body, hand, and audio tokens. The generated tokens are decoded into continuous dog motion using the PetVAE decoder.

5.2 Model Overview

As illustrated in Figure 3, **IPMG** consists of two major components. **Multi-modality tokenization**, which converts continuous signals from different modalities into discrete tokens. **Autoregressive transformer generation**, which models the conditional distribution of pet motion tokens given human motion and audio tokens.

Specifically, human body and hand gestures are first encoded into motion tokens using modality-specific encoders. Pet motion is represented using a discrete latent space learned by a PetVAE. In addition, global translation and audio features are converted into dedicated tokens. All tokens are then concatenated and fed into a transformer model, which predicts the next pet motion token conditioned on the multi-modal context. The generated pet tokens are finally decoded back to continuous motion sequences using the PetVAE decoder.

5.3 Multimodal Condition Encoder

Pet Motion Tokenizer (PetVAE). Pet motion exhibits different kinematic structures from human motion. To obtain a compact yet expressive representation, we learn a discrete latent space using a PetVAE.

Given a pet motion sequence $\mathbf{P}$, the encoder maps continuous joint positions into latent embeddings, which are then quantized using a learnable codebook to obtain discrete pet tokens. The decoder reconstructs the full pose sequence from the quantized embeddings. In addition to reconstructing joint positions, we utilize a **Bone Length Regressor** to enforce skeletal consistency. Specifically, alongside

the pose decoder that predicts 3D joint positions, an auxiliary regressor predicts the bone length of each skeletal segment. This auxiliary task encourages the model to preserve the underlying skeletal structure when reconstructing poses. To obtain a compact discrete representation of pet motion, we train a PetVAE to encode pose sequences into a codebook of motion tokens. The model is optimized with the reconstruction objective $\mathcal{L}_{pet} = \mathcal{L}^{pos}_{MSE} + \mathcal{L}^{bone}_{MSE} + \mathcal{L}_{VQ} + \beta\mathcal{L}_{commit}$, where $\mathcal{L}^{pos}_{MSE}$ supervises joint position reconstruction and $\mathcal{L}^{bone}_{MSE}$ enforces bone-length consistency. $\mathcal{L}_{VQ}$ and $\mathcal{L}_{commit}$ follow the standard VQ-VAE formulation [26]. This objective encourages the learned motion tokens to preserve both accurate pose geometry and realistic skeletal proportions.

Human Motion Tokenizer (Hand + Body). Human motion contains both body movements and detailed hand gestures. To better capture their distinct motion patterns, we tokenize body and hand joints as two separate streams. Given the human motion sequence $\mathbf{H} = \{\theta^{body}, \theta^{hand}, \mathbf{J}^{body}, \mathbf{J}^{hand}\}$ derived from SMPL body parameters and MANO hand parameters, an encoder maps the continuous joint rotations and positions into latent embeddings, which are then quantized using the codebook to obtain human motion tokens. Two decoders reconstruct joint rotations and positions from the quantized embeddings. It is trained with the following objective $\mathcal{L}_{human} = \mathcal{L}^{pos}_{MSE} + \mathcal{L}^{rot}_{MSE} + \mathcal{L}_{VQ} + \beta\mathcal{L}_{commit}$, where $\mathcal{L}^{rot}_{MSE}$ and $\mathcal{L}^{pos}_{MSE}$ supervise the reconstruction of joint rotations and positions, respectively.

Translation Tokenizer. Global root translations are modeled separately using a lightweight translation encoder–decoder module. The encoder maps the root translation sequence $\mathbf{T}$ into latent embeddings, which are then quantized into translation tokens. By decoupling global trajectory from local joint motion, the translation tokens capture coarse motion dynamics and serve as an intermediate representation between human conditioning tokens and fine-grained pet pose tokens in the autoregressive generation process.

Audio Tokenizer. We extract audio features using the pretrained MERT [16] and project them into token embeddings. These tokens provide temporal cues for modeling human–pet interactions.

5.4 Autoregressive Transformer

Given the multi-modal token sequence, we train a transformer to model the conditional distribution of pet motion tokens.

Human motion tokens serve as conditioning signals, while translation and pet motion tokens are generated autoregressively. Each transformer block contains self-attention, cross-attention, and feed-forward layers.

Modality-Aware Attention (MAA). Human–pet interaction exhibits a natural hierarchical structure, where human motion provides global context, translation captures coarse motion, and pet pose represents fine-grained movements. To reflect this dependency, we design a modality-aware attention mechanism that controls information flow between tokens.

Let the multi-modal token sequence be $\mathbf{T} = [\mathbf{T}_{human}, \mathbf{T}_{transl}, \mathbf{T}_{pet}]$, where $\mathbf{T}_{human} = [\mathbf{T}_{hand}, \mathbf{T}_{body}]$ denotes human motion tokens, and $\mathbf{T}_{transl}$ and $\mathbf{T}_{pet}$

denote translation and pet motion tokens. Human tokens use bidirectional attention as global conditioning, while translation and pet tokens attend to preceding tokens only, forming a coarse-to-fine generation hierarchy.

Audio Cross-Attention. Audio tokens are incorporated through a cross-attention module, providing temporal cues for interaction timing. The model predicts the next token with the objective $\mathcal{L}_{token} = \mathcal{L}_{CE}^{pet} + \mathcal{L}_{CE}^{transl}$.

During inference, tokens are generated autoregressively. Pet motion tokens are decoded by the PetVAE decoder, while translation tokens are decoded to recover global motion. The final pet motion is obtained by combining decoded poses with the predicted translations.

6 Experiments

6.1 Implementation Details

All models are implemented in PyTorch and trained on a single NVIDIA H100 GPU. We downsample the raw dataset to 30 fps and split per-dog into 80%/20% train/val sets. All motion sequences are segmented into clips of length $T = 300$ frames (approximately 10 seconds), with zero-padding applied to shorter sequences. During training, we apply a sliding window with a 50-frame stride to increase sample diversity. For IPMG training, these sequences are further chunked into 10-second windows, resulting in approximately 87.5 minutes of motion data. All sequences are normalized by subtracting the waist position in the first frame and aligning them to a canonical coordinate system based on the initial facing direction.

IPMG is trained in two stages. We first train four VQ-VAEs to tokenize each modality using a shared codebook size of 512 and temporal downsampling. We then train a 28M-parameter prefix-LM GPT to autoregressively generate dog motion tokens conditioned on motion tokens, with MERT audio features injected via cross-attention. The VQ-VAEs and GPT are trained for 500 and 300 epochs, respectively. More details are provided in the appendix.

6.2 Evaluation Metrics

We evaluate the generated motions using commonly used metrics in motion generation, including Fréchet Inception Distance (FID), Retrieval Precision (R-Precision), and Diversity (Div).

Fréchet Inception Distance (FID)↓ FID measures the distributional similarity between generated motions and ground-truth motions. Lower values indicate that the generated motions are closer to real motion data. We report FID in both kinetic (FID_k) and static (FID_s) feature spaces.

Retrieval Precision (R-Precision)↑ R-Precision evaluates how well the generated pet motions align with the conditioning signals. We report R-Precision with respect to hand ($R_{Prec.}^{hand}$) and body ($R_{Prec.}^{body}$) conditioning signals.

Diversity (Div)↑ Diversity measures the variability of generated motion samples. Higher values indicate that the model produces more diverse motion patterns. We report diversity in both kinetic (Div_k) and static (Div_s) feature spaces.

Table 3: Quantitative comparison and ablation study of baseline architectures and design variants.

Architecture / Variant	$FID_k \downarrow$	$FID_s \downarrow$	$R^{Hand}_{Prec.} \uparrow$		$R^{Body}_{Prec.} \uparrow$		$Div_k \uparrow$	$Div_s \uparrow$
			Top-1	Top-3	Top-1	Top-3		
seq2seq-Transf. (plain MHA)	21.22	24.18	0.25	0.41	0.22	0.38	5.01	5.07
Diffusion-Transf.	64.37	67.94	–	0.26	–	0.23	4.42	4.50
IPMG w/ causal MHA	13.83	15.44	0.38	0.56	0.36	0.54	**6.20**	5.85
IPMG w/o PetVAE	14.15	16.21	–	0.56	–	0.52	5.62	5.70
IPMG w/ MAA (Ours)	**11.21**	**12.96**	**0.46**	**0.63**	**0.43**	**0.59**	5.93	**6.01**

6.3 Baselines

Since no prior methods exist for human-to-pet gesture-to-motion generation, we compare our method with two representative architectures for motion generation: **Seq2Seq-Transformer** adopts a standard encoder-decoder transformer with plain multi-head attention (plain MHA), which autoregressively predicts motion tokens conditioned on the input signal. We further compare it with a **causal MHA** variant to examine the effect of attention masking. Diffusion-Transformer (**DiT**) formulates motion generation as a denoising diffusion process with a transformer backbone. We also report **IPMG (w/o PetVAE)** to evaluate the contribution of the PetVAE module. The full model, **IPMG w/ MAA**, combines PetVAE with the proposed multi-modal motion attention mechanism. These variants allow us to isolate the effects of autoregressive modeling, PetVAE tokenization, and the proposed MAA mechanism.

6.4 Quantitative Comparison

Table 3 presents the quantitative comparison with different baseline architectures. The proposed **IPMG** achieves the best performance across all evaluation metrics. Compared to the autoregressive **Seq2Seq-Transformer**, IPMG reduces FID_k by 47.2% (from 21.22 to 11.21) while significantly improving the alignment of motion conditions, increasing $R^{hand}_{Prec.}$ from 0.41 to 0.63, and $R^{body}_{Prec.}$ from 0.38 to 0.59. The diversity of generated motions improves from 5.01 to 5.93, indicating that IPMG produces more varied, yet realistic, motion sequences. Compared with the **causal MHA** variant, IPMG further improves FID_k/FID_s from 13.83/15.44 to 11.21/12.96 and increases hand/body R-Precision from 0.56/0.54 to 0.63/0.59. This indicates that allowing the model to access broader contextual information across the interaction sequence is beneficial for generating more realistic and condition-consistent pet motions. The improvement is especially important for human-pet interaction modeling, where the pet's response often depends not only on preceding human actions but also on the overall interaction context.

Compared to the **DiT** baseline, which directly applies a generic diffusion backbone to motion generation, IPMG reduces FID_k by 82.6% and substantially

Table 4: Ablation study on input modalities. Using both body and hand inputs improves motion quality, alignment, and diversity compared to single-modality inputs.

Input	$FID_k \downarrow$	$FID_s \downarrow$	$R^{hand}_{Prec.}@3 \uparrow$	$R^{body}_{Prec.}@3 \uparrow$	$Div_k \uparrow$	$Div_s \uparrow$
Body	13.48	15.72	0.44	0.57	5.36	5.41
Hand	17.92	19.83	0.58	0.36	5.88	5.94
Body+Hand	**11.21**	**12.96**	**0.63**	**0.59**	**5.93**	**6.01**

Table 5: Ablation on the PetVAE bone constraint. The full PetVAE improves motion quality and stability compared to the version without bone constraints.

	$FID \downarrow$	MPJPE	Vel. Err.
PetVAE (w/o bone)	12.53	7.26	1.79
PetVAE	9.20	6.42	1.03

improves alignment scores for hand and body motions (0.26 to 0.63, and 0.23 to 0.59). This result suggests that generic diffusion architectures struggle to model structured human–pet interactions without appropriate motion priors.

Finally, removing the PetVAE module leads to consistent performance degradation across all metrics. Adding PetVAE further reduces FID_k by 20.8% (14.15 to 11.21) while improving both alignment and diversity scores, demonstrating that the learned discrete motion representation helps stabilize training and improve motion realism.

6.5 Ablation Study

Input modality ablation. Table 4 presents an ablation study on different input modalities. Using both body and hand signals consistently yields the best performance across all metrics. When only body information is used, the model achieves better body-level alignment ($R^{body}_{Prec.} = 0.57$) but performs worse on hand-level alignment ($R^{hand}_{Prec.} = 0.44$), indicating that body signals alone provide global interaction context but lack the fine-grained cues required for modeling detailed human–pet interactions.

In contrast, using only hand input improves hand-level alignment ($R^{hand}_{Prec.} = 0.58$) but significantly degrades body-level alignment ($R^{body}_{Prec.} = 0.36$), indicating that local hand signals alone lack sufficient global motion context.

PetVAE ablation. Table 5 evaluates the effect of the bone constraint in PetVAE. Removing the bone constraint leads to degraded motion quality across all metrics, increasing FID from 9.20 to 12.53 and MPJPE from 6.42 to 7.26. In addition, the motion velocity error increases from 1.03 to 1.79, indicating less stable motion dynamics. These results show that incorporating bone constraints helps preserve kinematic consistency and improves the realism of the learned motion representation.

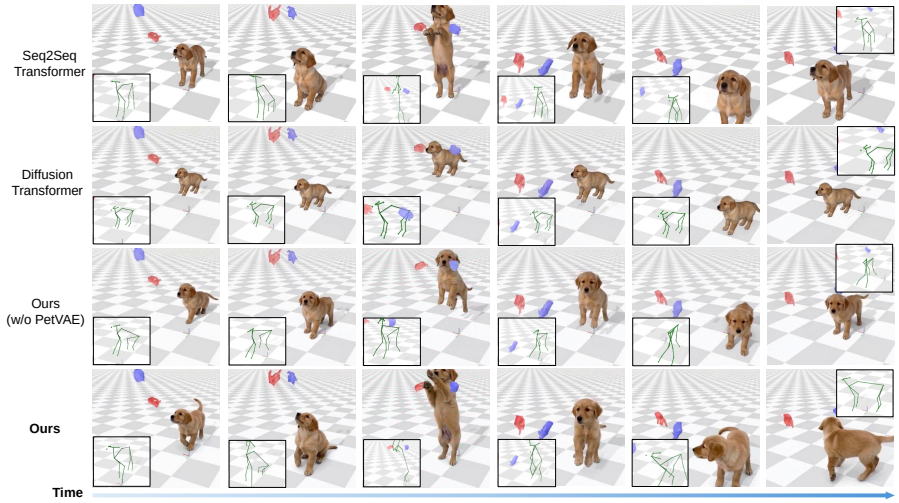

Fig. 4: Qualitative results. We visualize InterPetMoGen outputs in a sequential command of "come->jump->turn" compared to other baselines. For each example, only the trainer's hands are rendered for better visibility of the dog.

6.6 Qualitative Results

Figure 4 presents qualitative comparisons of generated dog motion sequences under a sequential command scenario ("come → jump → turn"). Our method produces the most natural and responsive behaviors among all compared approaches. The generated motions not only remain temporally smooth, but also clearly react to the trainer's hand cues, including turning toward the command direction and executing the instructed actions.

In contrast, the Seq2Seq baseline can produce partially plausible motions, but shows limited responsiveness to input gestures. The dog's orientation remains unchanged across the sequence, and actions such as turning are rarely observed. Our model w/o PetVAE module is still able to somehow react to the commands, but the resulting motions lack realism and appear less physically natural. Finally, the DiT baseline performs the worst among the compared methods, producing unstable and less meaningful motions, likely due to the limited amount of training data available for diffusion-based modeling in this setting.

6.7 User Study

Because pet behavior does not deterministically follow human gestures, evaluating the perceptual quality of generated motions is essential. We therefore conducted a user study with 12 participants to assess the naturalness and appropriateness of the generated dog responses.

Each participant evaluated three randomly selected human-interaction input sequences and the corresponding dog motions generated by different methods. To

Table 6: User study results. Average participant ratings on a 7-point Likert scale evaluating the naturality, responsiveness, and overall quality of generated dog motions. Higher scores indicate better perceptual quality.

Method	Naturality ↑	Responsiveness ↑	Overall ↑
Seq2Seq-Transf.	4.04	3.67	3.67
DiT	2.48	2.39	2.12
IPMG (w/o PetVAE)	3.82	3.64	3.70
IPMG (Ours)	**6.58**	**6.55**	**6.63**

make the evaluation more intuitive, we rendered the skeletal motions into realistic dog videos using a finetuned video DiT model (Figure 4). Participants rated each result on a 7-point Likert scale for *naturality*, *responsiveness*, and *overall motion quality*.

The results are summarized in Table 6. Our full model achieves the highest scores across all criteria, with 6.58 in naturality, 6.55 in responsiveness, and 6.63 in overall quality, clearly outperforming the strongest baseline, Seq2Seq-Transformer (4.04, 3.67, and 3.67). This suggests that explicitly modeling human-pet interactions with our proposed components leads to more natural and responsive pet motions than directly applying generic seq-to-seq motion generation models. Removing PetVAE also degrades performance across all criteria, confirming its importance for capturing realistic and diverse pet motion patterns during human-pet interaction. A repeated-measures ANOVA further shows significant condition effects for naturality ($F(3,6) = 5.78$, $p = .033$), responsiveness ($F(3,6) = 24.91$, $p < .001$), and overall quality ($F(3,6) = 9.39$, $p = .011$), indicating that the observed differences among methods are statistically significant.

Importantly, these subjective results are consistent with our quantitative evaluation (Section 6) and the qualitative comparisons in Figure 4. Together, these results demonstrate that IPMG not only improves objective metrics but also produces perceptually more natural and responsive dog motions in human–pet interaction scenarios.

7 Limitations and Future Work

While InterPet4D is the first dataset of its kind, it has several limitations.

First, the dataset currently covers only dogs; extending to other pet species (*e.g.*, cats) would increase generality but requires species-specific pose models. **Second**, the current framework does not model physical contact forces between human and dog, which are important in petting and playing interactions. Incorporating physics-based constraints [33] could improve physical plausibility. **Third**, our model generates motion clips of fixed length (10 seconds), extending to variable-length or autoregressive generation would enable modeling of longer interactions.

8 Conclusion

We presented InterPet4D, the first large-scale multimodal 4D dataset for human–pet interaction, featuring synchronized 12-view third-person and egocentric RGB video from Ray-Ban 2 glasses, 3D human body and hand motion, 3D dog pose, and audio across 23 participants and 13 dogs. We established a systematic interaction taxonomy and annotation pipeline, providing a standardized benchmark for human-pet interaction research. We further introduced InterPetMoGen, an autoregressive model that leverages discrete motion representations to synthesize plausible pet responses conditioned on human motion and audio. We hope InterPet4D will facilitate future research on multimodal human-pet interaction modeling.

9 Acknowledgment

This work was supported in part by JST ASPIRE JPMJAP2404, JST CRONOS JPMJCS24N8, Crescent Inc., and Shanda Interactive Intelligence Collaborative Research Cluster.

References

1. Agrawal, V., et al.: Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset (2025), https://arxiv.org/abs/2506.22554
2. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: BEHAVE: Dataset and method for tracking human object interactions. In: CVPR. pp. 15935–15946 (2022)
3. Biggs, B., Bober, O., Sherrill, J.M., Koepke, A.S., Mayol-Cuevas, W.W., Sherrill, D.M., Sherrill, S.M., Sherrill, J.M.: Who left the dogs out? 3D animal reconstruction with expectation maximization in the loop. In: ECCV. pp. 195–211 (2020)
4. Biggs, B., Roddick, T., Fitzgibbon, A., Cipolla, R.: Creatures great and smal: Recovering the shape and motion of animals from video. In: Asian Conference on Computer Vision. pp. 3–19. Springer (2018)
5. Cai, Y., Wu, Y., Li, K., Zhou, Y., Zheng, B., Liu, H.: Flooddiffusion: Tailored diffusion forcing for streaming motion generation (2026), https://arxiv.org/abs/2512.03520
6. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023)
7. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5386–5395 (2020)
8. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: CVPR. pp. 12943–12954 (2023)
9. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions (2023), https://arxiv.org/abs/2312.00063

10. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
11. Higami, A., Oshima, K., Shiramatsu, T.I., Takahashi, H., Nobuhara, S., Nishino, K.: Ratbodyformer: Rat body surface from keypoints (2025), https://arxiv.org/abs/2412.09599
12. Hussain, M.: Yolo-v1 to yolo-v8, the rise of yolo and its complementary nature toward digital manufacturing and industrial defect detection. Machines **11**(7), 677 (2023)
13. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36** (2024)
14. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: European Conference on Computer Vision. pp. 196–214. Springer (2020)
15. Khirodkar, R., Song, J.T., Cao, J., Luo, Z., Kitani, K.: Harmony4d: A video dataset for in-the-wild close human interactions. Advances in Neural Information Processing Systems **37**, 107270–107285 (2024)
16. Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Lin, C., Ragni, A., Benetos, E., Gyenge, N., Dannenberg, R., Liu, R., Chen, W., Xia, G., Shi, Y., Huang, W., Guo, Y., Fu, J.: Mert: Acoustic music understanding model with large-scale self-supervised training (2023)
17. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: InterGen: Diffusion-based multi-human motion generation under complex interactions. In: IJCV (2024)
18. Liu, H., Zhu, Z., Becherini, G., Peng, Y., Su, M., Zhou, Y., Zhe, X., Iwamoto, N., Zheng, B., Black, M.J.: Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1144–1154 (June 2024)
19. Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., Zheng, B.: Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII. p. 612–630. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20071-7_36, https://doi.org/10.1007/978-3-031-20071-7_36
20. Liu, I., Xu, Z., Wang, Y., Tan, H., Xu, Z., Wang, X., Su, H., Shi, Z.: Riganything: Template-free autoregressive rigging for diverse 3d assets. ACM Transactions on Graphics (TOG) **44**(4), 1–12 (2025)
21. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM TOG **34**(6), 1–16 (2015)
22. Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., Chen, K.: Rtmdet: An empirical study of designing real-time object detectors (2022)
23. Mathis, A., Mamidanna, P., Cull, K.M., Hainmueller, B., Hosp, S., Liao, C.L., Bethge, M.: DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. Nature Neuroscience **21**(9), 1281–1289 (2018)
24. Müller, L., Ye, V., Pavlakos, G., Black, M.J., Kanazawa, A.: BUDDI: Building dynamic human-human interaction. In: CVPR. pp. 978–988 (2024)
25. Ng, X.L., Ong, K.E., Zheng, Q., Ni, Y., Yeo, S.Y., Liu, J.: Animal kingdom: A large and diverse dataset for animal behavior understanding. In: CVPR. pp. 19023–19034 (2022)
26. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning (2018), https://arxiv.org/abs/1711.00937

27. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019)
28. Peng, Y., Song, J.T., Jung, S., Liu, R., Liu, H., Chu, X., Liu, R., Wu, E., Koike, H., Kitani, K.: Dyadit: A multi-modal diffusion transformer for socially favorable dyadic gesture generation (2026), https://arxiv.org/abs/2602.23165
29. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: ICCV. pp. 10985–10995 (2021)
30. Potamias, R.A., Zhang, J., Deng, J., Zafeiriou, S.: Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12242–12254 (2025)
31. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)
32. Rüegg, N., Zuffi, S., Schindler, K., Black, M.J.: BARC: Learning to regress 3D dog shape from images with reinforcement of breed-specific 3D shape priors. In: CVPR. pp. 3886–3896 (2022)
33. Rüegg, N., Zuffi, S., Schindler, K., Black, M.J.: BITE: Beyond priors for improved three-D dog pose estimation. In: CVPR. pp. 8867–8876 (2023)
34. Sabathier, R., Mitra, N.J., Novotny, D.: Animal avatars: Reconstructing animatable 3d animals from casual videos. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXIX. p. 270–287 (2024). https://doi.org/10.1007/978-3-031-72986-7_16
35. Särkkä, S.: Unscented rauch–tung–striebel smoother. IEEE transactions on automatic control **53**(3), 845–849 (2008)
36. Shi, X., Wang, X., Guo, Z., Wang, Y., Zhang, P., Zhang, X., Guo, Z., Hao, H., Xi, Y., Yang, B., Xu, J., Zhou, J., Lin, J.: Qwen3-asr technical report. arXiv preprint arXiv:2601.21337 (2026)
37. Shin, S., Lee, C., Chen, H., Song, J.T., Halilaj, E., Kitani, K.: Bodycontact4d: A multi-view video dataset for understanding human and environment interactions. In: Thirteenth International Conference on 3D Vision
38. Sinha, S., Shapovalov, R., Reizenstein, J., Rocco, I., Neverova, N., Vedaldi, A., Novotny, D.: Common pets in 3d: Dynamic new-view synthesis of real-life deformable categories. CVPR (2023)
39. Siyao, L., Gu, T., Yang, Z., Lin, Z., Liu, Z., Ding, H., Yang, L., Loy, C.C.: Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment (2024), https://arxiv.org/abs/2403.18811
40. Song, J.T., Kim, J., Cao, J., Lei, Y., Yagi, T., Kitani, K.: Contact4d: A video dataset for whole-body human motion and finger contact in dexterous operations. In: Thirteenth International Conference on 3D Vision
41. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: ECCV. pp. 581–600 (2020)
42. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)
43. Wang, Z., Chen, S., Mo, L., Gao, X., Shen, Y., Ding, L., Liang, W.: Dogmo: A large-scale multi-view rgb-d dataset for 4d canine motion recovery (2025), https://arxiv.org/abs/2510.24117
44. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017)

45. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3D human motion from speech. In: CVPR. pp. 469–480 (2023)
46. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2M-GPT: Generating human motion from textual descriptions with discrete representations. In: CVPR. pp. 14141–14150 (2023)
47. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: MotionDiffuse: Text-driven human motion generation with diffusion model. IEEE TPAMI **46**(4), 2514–2527 (2024)
48. Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J.: 3d menagerie: Modeling the 3d shape and pose of animals. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6365–6373 (2017)