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[2], and Erwin Wu[2]

[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, greatly 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 multi-view and egocentric RGB video, 3D human body and hand motion, 3D dog body motion, and audio across a diverse set of 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.

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 (MMA) 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 multi-view 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 MMA 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 [43] captures whole-body grasping of objects with detailed hand motion. BEHAVE [2] provides multi-view recordings of humans interacting with rigid objects. ARCTIC [9] focuses on articulated object manipulation with dexterous hand motion. For human–human interaction, InterHuman [19] proposes a large-scale dataset and a diffusion-based model for two-person motion generation. BUDDI [26] 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 Esimation

Estimating the 3D pose and shape of animals has attracted growing attention. SMAL [50] introduces a parametric 3D body model for animals learned from toy figurines. Subsequent works focus on dogs, including WLDO [3], BARC [34], and BITE [35], which improve monocular 3D reconstruction using breed priors and contact constraints, as well as DogMo [45], which reconstructs dog motion from monocular videos, and AnimalAvatar [36], which reconstructs animatable 3D animals and motion from videos. RigAnything [22] further enables automatic skeletal rigging for arbitrary animal meshes. For pose estimation, DeepLabCut [25] provides a widely-used markerless framework across species, while RatBodyFormer [13] applies transformer models to estimate rodent body pose. On the data side, Animal Kingdom [27] and COP3D [40] 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 [31] use VAE-based architectures for class-conditional generation. Text-conditioned approaches have flourished with MDM [44], which applies diffusion models to motion generation from text descriptions, and T2M-GPT [48], Duolando [41], MDLS [7], which uses VQ-VAE with GPT-based autoregressive generation. MotionDiffuse [49], FloodDiffusion [5] enable fine-grained text-driven motion synthesis. MoMask [10] introduces masked modeling for efficient motion generation. In the audio domain, co-speech gesture generation methods [20, 21, 30, 47] synthesize body and hand gestures from speech audio. MotionGPT [15] 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

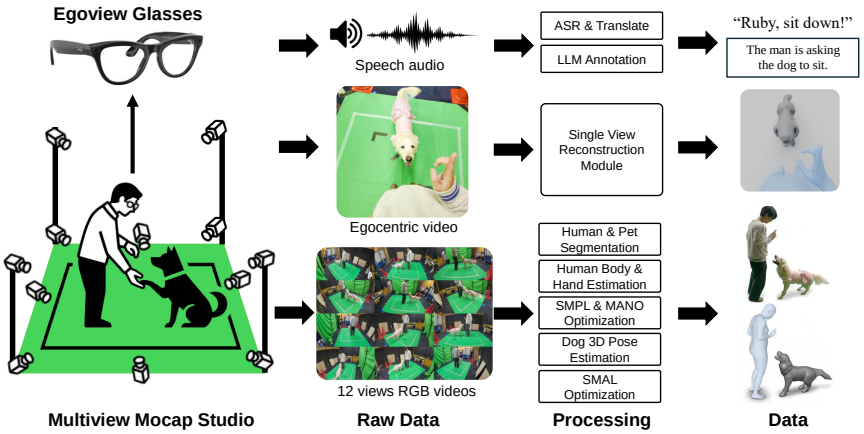


Fig. 2: Data collection setup. Our capture environment consists of 12 synchronized third-person RGB cameras arranged in a square, complemented by Rayban Meta Glasses worn by the participant providing an egocentric view and audio.

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.

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 [19]	Human	12 pairs	–	–	✓	76	107M	✓
Seamless Inter. [1]	Human	4000+	–	✓	Speech-only	1	400M+	✓
CoP3D [40]	Dog & Cat	–	4200	–	–	1	0.6M	✓
DogMo [45]	Dog	–	10	–	–	5	1M	–
InterPet4D	Human+Dog	23	13	✓	✓	12+Ego	6.8M	✓*

*The full dataset including annotations will be publicly available after acceptance.

Participants. To ensure stable interactions, we collaborated with a local pet training agency and recruited 11 experienced trainers together with different 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 [23], SMPL-X [29], and MANO [33] 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 [23] model. We first localize the human in each view using Mask R-CNN [11] with DeepSORT tracking [46], followed by manual filtering to remove failure cases. We then estimate 2D body keypoints using HRNet-WholeBody [16] 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 [17]. Finally, an SMPL [23] 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 [14] and comparing the detections with the reprojected SMPL hand mesh. Using the resulting close-up hand crops, we apply WiLoR [32] 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 [39, 42]. Finally, we fit MANO [33] to the refined 3D hand keypoints to recover hand pose, and fit SMPL-X [29] 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 [50] body model. For each view, we first detect the dog using RTMDet [24] and estimate 2D dog keypoints using HRNet-W32 [8]. We then run Mask R-CNN [11] 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 [37] to improve temporal consistency. Finally, we fit the SMAL [50] 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 [50] 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 [38] to transcribe the spoken commands and translate them into English. In addition, the temporal boundaries of each command are automatically annotated using Qwen3-ForcedAligner [38].
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 [38] 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.

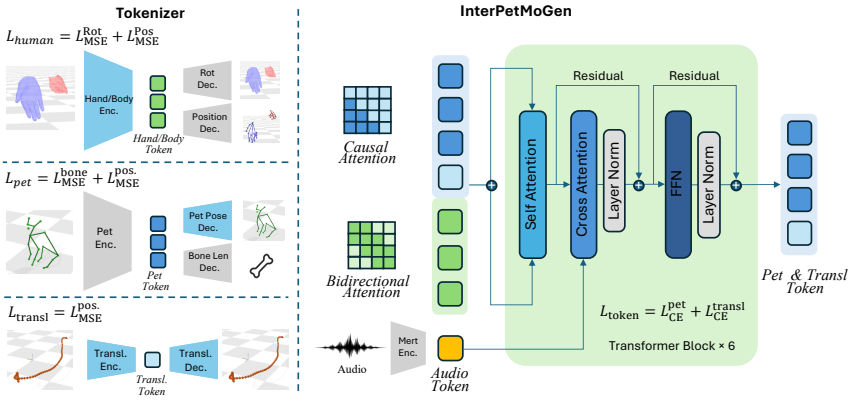


Fig. 3: InterPetMoGen architecture. Human gesture and audio are encoded by modality-specific Transformer encoders and fused via cross-modal attention. The fused condition **c** guides a denoising diffusion decoder that generates dog motion sequences. During inference, classifier-free guidance balances the fidelity and motion diversity.

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 gesture 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 **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 [28]. 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 [18] 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. 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.

In the first stage, we train four VQ-VAEs to tokenize each modality, using a shared codebook size of 512 and 2× temporal downsampling (4× for relative motion). Each VQ-VAE is trained for 500 epochs (∼30 minutes). In the second stage, we train a prefix-LM GPT (28M parameters) to autoregressively generate dog motion tokens conditioned on human motion and audio. The input sequence contains 525 tokens from human motion, relative motion, and dog motion, while audio features from MERT are injected through cross-attention. The GPT model is trained for 300 epochs (∼65 minutes). Additional 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.

Table 3: Quantitative comparison with baseline architectures. Our method achieves the best performance in motion quality (FID), alignment (R-Prec.), and diversity.

Architecture	$FID_k \downarrow$	$FID_s \downarrow$	$R^{Hand}_{Prec.} \uparrow$	$R^{Body}_{Prec.} \uparrow$	$Div_k \uparrow$	$Div_s \uparrow$
seq2seq-Transf.	21.22	24.18	0.41	0.38	5.01	5.07
Diffusion-Transf.	64.37	67.94	0.26	0.23	4.42	4.50
IPMG (w/o PetVAE)	14.15	16.21	0.56	0.52	5.62	5.70
IPMG (Ours)	**11.21**	**12.96**	**0.63**	**0.59**	**5.93**	**6.01**

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.

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 that autoregressively predicts motion tokens conditioned on the input signal. Diffusion-Transformer (**DiT**) models motion generation as a denoising diffusion process using a transformer backbone. In addition, we include an ablated version of our model, **IPMG (w/o PetVAE)**, where the PetVAE module is removed to evaluate its contribution. The full model, **IPMG**, integrates PetVAE with the proposed motion generation framework.

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

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	$FID_k \downarrow$	$FID_s \downarrow$	$R^{hand}_{Prec.} \uparrow$	$R^{body}_{Prec.} \uparrow$	$Div_k \uparrow$	$Div_s \uparrow$
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

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.

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.

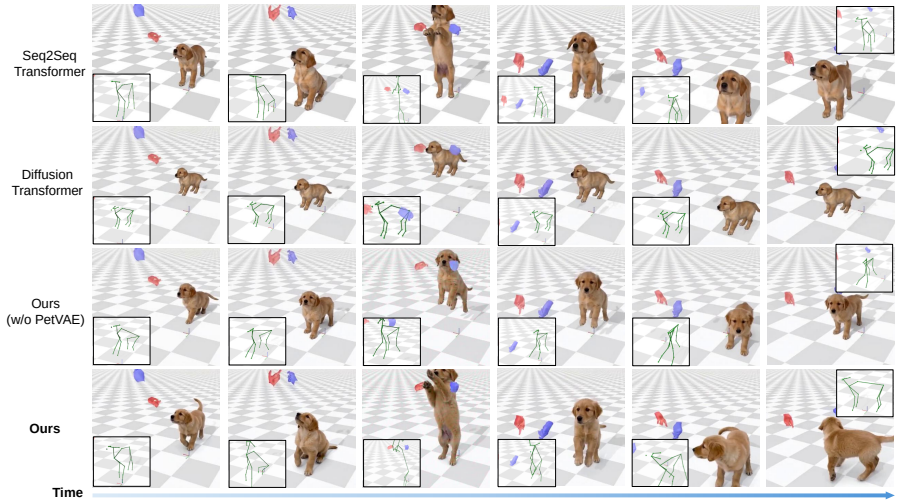


Fig. 4: Qualitative results. We visualize InterPetMoGen outputs in a sequential command of "come->jump->turn" compared to other baselines. For each example, only the trainer's hands are rendered for better visibility of the dog.

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**

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 was shown three randomly selected human–interaction input sequences together with the corresponding generated dog motions from different

methods. For each sequence, participants evaluated the results produced by our model and the other baselines. Since most participants are not familiar with skeleton representations, we rendered the generated skeletal motions into realistic dog videos using a finetuned video DiT model to facilitate intuitive evaluation (see Figure 4). Participants rated the generated motions using a 7-point Likert scale based on three criteria: *naturality* of the dog motion, *responsiveness* to the human input, and *overall motion quality*.

The results are summarized in Table 6. Our full model achieves the highest scores across all criteria, obtaining 6.58 in naturality, 6.55 in responsiveness, and 6.63 in overall quality, substantially outperforming the baseline methods. In comparison, the strongest baseline (Seq2Seq-Transformer) achieves scores of 4.04, 3.67, and 3.67, respectively, while DiT performs significantly worse. Removing the Pet-VAE component also leads to noticeable performance degradation, confirming its importance for modeling realistic human–pet interactions.

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**, our markerless dog pose reconstruction, while practical, introduces noise compared to marker-based capture, particularly for fast motions and heavy occlusion. **Third**, our model generates dog motion as a reactive response to human input, but real interactions are bidirectional, the human also adapts to the pet. Modeling this mutual adaptation is an important direction for future work. **Fourth**, 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 [35] could improve physical plausibility. **Finally**, 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

auto regressive 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.

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. Cao, J., Tang, H., Fang, H.S., Shen, X., Lu, C., Tai, Y.W.: Cross-domain adaptation for animal pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9498–9507 (2019)
7. 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)
8. 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)
9. 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)
10. 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
11. 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)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018), https://arxiv.org/abs/1706.08500
13. 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
14. 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)
15. 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)

16. 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)
17. 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)
18. 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)
19. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: InterGen: Diffusion-based multi-human motion generation under complex interactions. In: IJCV (2024)
20. 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)
21. 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
22. 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)
23. 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)
24. 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)
25. 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)
26. Müller, L., Ye, V., Pavlakos, G., Black, M.J., Kanazawa, A.: BUDDI: Building dynamic human-human interaction. In: CVPR. pp. 978–988 (2024)
27. 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)
28. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning (2018), https://arxiv.org/abs/1711.00937
29. 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)
30. 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
31. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3D human motion synthesis with transformer VAE. In: ICCV. pp. 10985–10995 (2021)

32. 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)
33. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)
34. 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)
35. 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)
36. 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
37. Särkkä, S.: Unscented rauch–tung–striebel smoother. IEEE transactions on automatic control **53**(3), 845–849 (2008)
38. 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)
39. 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
40. 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)
41. 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
42. 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
43. 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)
44. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)
45. 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
46. 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)
47. 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)
48. 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)
49. 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)

50. 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)

This supplementary material contains the following sections:

- **A.** Network Architecture & Training Details
- **B.** Dog Skeleton Definition
- **C.** Multi-GoPro Capture & Synchronization Details
- **D.** Dog Shape & Motion Analysis
- **E.** Audio Modality Ablation
- **F.** Dog Motion FID Model
- **G.** Ethical Statement
- **H.** Additional Qualitative Results (Supplementary Video)

A Network Architecture & Training Details

We provide detailed architecture specifications and training hyperparameters for each component of InterPetMoGen.

PetVAE. The PetVAE encoder takes a pet motion sequence of shape $(T, 60)$ (*i.e.*, 20 keypoints $\times$ 3 coordinates) as input. We adopt an encoder-decoder architecture backbone with a single-level EMA vector quantization bottleneck. The encoder and decoder each consist of 4 stages with channel multipliers $[1, 2, 4, 4]$ and a base dimension of 128, yielding a latent dimension of 128 with a codebook size of 1024. Temporal downsampling (factor $4\times$) is applied in the first two encoder stages via strided convolutions, and symmetrically upsampled in the decoder. Each stage contains 2 residual blocks with RMSNorm and SiLU activations.

Human Motion Tokenizer. Human body and hand motions are tokenized independently using two separate VQ-VAE modules sharing the same architecture but with different input dimensions. The body tokenizer takes SMPL joint rotations and positions $(\theta^{body}, \mathbf{J}^{body})$ as input, where $\theta^{body} \in \mathbb{R}^{T\times D_b}$ and $\mathbf{J}^{body} \in \mathbb{R}^{T\times J_b\times 3}$. The hand tokenizer takes MANO joint rotations and positions, denoted as $\mathbf{x}^{hand} = (\theta^{hand}, \mathbf{J}^{hand})$, where $\theta^{hand} \in \mathbb{R}^{T\times D_h}$ and $\mathbf{J}^{hand} \in \mathbb{R}^{T\times J_h\times 3}$. Both tokenizers use $4\times$ temporal downsampling and a codebook of size 512. Each module includes two decoder heads: one for joint rotations and one for joint positions.

Translation Tokenizer. A lightweight VQ-VAE encodes the global root translation sequence $\mathbf{T} \in \mathbb{R}^{T\times 3}$ into discrete tokens. This module uses $4\times$ temporal downsampling to capture coarse motion dynamics and a codebook of size 512.

Audio Tokenizer. Audio features are extracted using the pretrained MERT [18] model, yielding frame-level features of dimension $d_a = 1024$. These features are projected into token embeddings of dimension 512 via a linear layer and fed to the transformer through cross-attention.

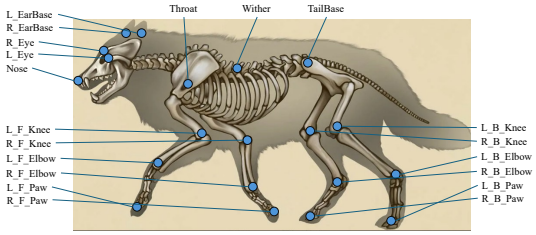


Fig. 5: Dog skeleton topology. Our 20-keypoint skeleton covers the facial region, torso, limbs, and tail base.

Autoregressive Transformer (GPT). The prefix-LM GPT model has 30.3M parameters and consists of 6 transformer blocks. Each block contains a self-attention layer with 8 heads, a cross-attention layer for audio conditioning, and a feed-forward network with hidden dimension 2048. The input sequence consists of 375 tokens: 150 hand tokens (75 left + 75 right), 75 SMPL body tokens, 75 pet-human relative translation tokens, and 75 pet motion tokens. Human motion tokens (hand + body) use bidirectional attention as global conditioning (prefix), while relative translation and pet tokens are generated autoregressively with causal masking. During inference, we apply top-k sampling with $k = 50$ and temperature $\tau = 1.0$.

Training Details. All models are implemented in PyTorch and trained on a single NVIDIA H100 GPU.

- **VQ-VAEs (Stage 1):** Trained for 500-1000 epochs using the Adam optimizer with a learning rate of 3×10^{-5}, $\beta_1 = 0.5$, $\beta_2 = 0.999$, weight decay of 0.01, and batch size 128. A MultiStepLR schedule reduces the learning rate by half at epochs 100, 200, and 300. The commitment loss weight $\beta = 0.02$ is linearly warmed up over the first 50 epochs.
- **GPT (Stage 2):** Trained for 300 epochs using the AdamW optimizer with a learning rate of 3×10^{-4}, batch size 32, weight decay 0.01, and gradient clipping with max norm 1.0. A MultiStepLR schedule reduces the learning rate by half at epochs 100 and 200.

B Dog Skeleton Definition

We adopt a 20-keypoint dog skeleton based on the AnimalPose definition, as illustrated in Figure 5. The keypoints correspond to major anatomical landmarks of the dog, including the facial region, torso, limbs, and tail base. Table 7 lists all keypoints together with their parent keypoints, defining the kinematic hierarchy. We treat the *Withers* joint as the root of the skeleton, and construct the skeletal graph by connecting each keypoint to its parent according to the kinematic tree. We also provide a `data_sample` file in the supplementary material for reference.

Table 7: Dog skeleton definition. The 20 AnimalPose [6] keypoints and their parent connections following the AnimalPose skeleton. (We follow the it's definition, which does not include tail tip keypoints.)

Idx	Keypoint	Parent	Idx	Keypoint	Parent
0	L_Eye	-	10	L_B_Knee	6
1	R_Eye	0	11	R_B_Knee	6
2	L_EarBase	0	12	L_F_Elbow	8
3	R_EarBase	1	13	R_F_Elbow	9
4	Nose	2	14	L_B_Elbow	10
5	Throat	0	15	R_B_Elbow	11
6	TailBase	7	16	L_F_Paw	12
7	Withers	5	17	R_F_Paw	13
8	L_F_Knee	5	18	L_B_Paw	14
9	R_F_Knee	5	19	R_B_Paw	15

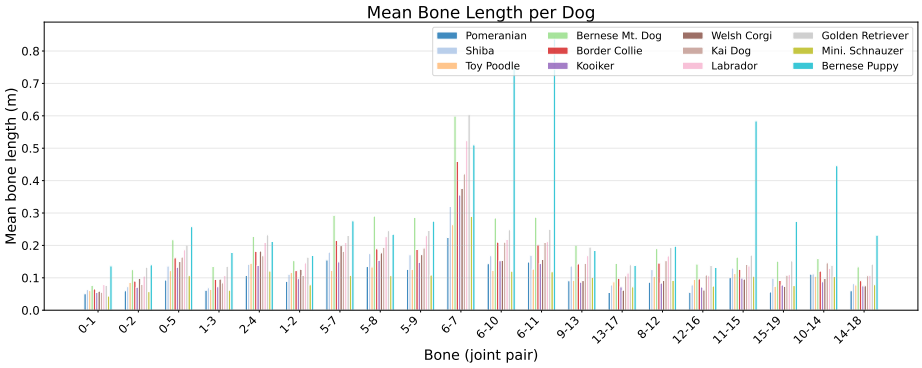
C Multi-GoPro Capture & Synchronization Details

Camera Calibration. The 12 GoPro cameras were calibrated using a checkerboard pattern with a (7×4) grid and a cell size of 0.1. We first detected chessboard corners in the synchronized calibration images, then estimated the intrinsic parameters of each camera, including lens distortion, and solved for the extrinsic parameters in a common coordinate system. The average calibration error reported across the 12 cameras was 0.245 pixels (approximately 0.25 pixels) in 1080p resolution.

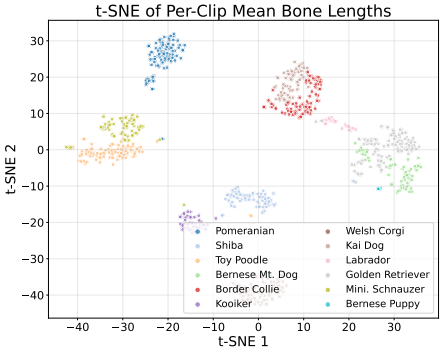
Camera Synchronization Our capture system consists of multiple GoPro cameras controlled via the OpenGoPro HTTP API. Prior to each recording session, we synchronize the internal clocks of all cameras to the host PC's system time through the API's `set_date_time` endpoint, ensuring a consistent time reference across devices. Recording is then triggered simultaneously on all cameras using parallel HTTP requests, so that each camera begins capturing at approximately the same moment. Each GoPro embeds a timecode in the recorded video stream metadata, reflecting its synchronized internal clock. In post-processing, we extract these embedded timecodes using `ffprobe` and compute per camera temporal offsets relative to the latest-starting camera. All video streams are then front-trimmed accordingly and cropped to their common overlapping duration, yielding frame-level synchronized multi-view footage. For the egocentric camera, synchronization with the multi-view system is achieved manually using a clapping gesture recorded at the beginning of each session as a visual cue.

D Dog Shape & Motion Analysis

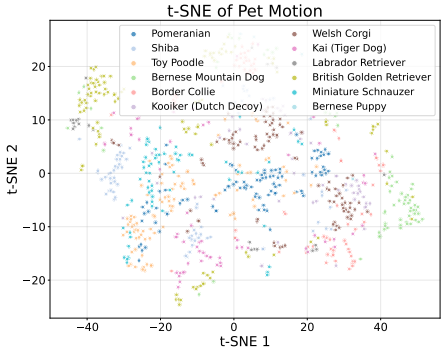
Our dataset captures 12 dogs spanning a wide range of breeds, from small breeds such as Pomeranian and Toy Poodle to large breeds such as Bernese Mountain



(a) Mean bone length per breed.



(b) t-SNE of per clip bone lengths.



(c) t-SNE of per clip motion features.

Fig. 6: Shape and motion analysis. (a) Mean bone lengths differ markedly across the 12 breeds, capturing diverse body proportions. (b) per clip bone length vectors form tight per dog clusters in t-SNE space, demonstrating consistent and accurate shape estimation. (c) Motion features are broadly distributed and overlapping across breeds, reflecting the rich motion diversity in our dataset.

Dog and Labrador Retriever. Note that we exclude one puppy from the original 13 dogs due to its limited number of recorded sequences, and conduct the following analysis on the remaining 12. We analyze the shape and motion characteristics of the captured data to validate the quality and diversity of our dataset.

Shape consistency. For each dog, we compute the mean bone length of all 20 skeleton bones across every clip. Figure 6a reports the per bone mean lengths grouped by breed. The bone-length profiles vary substantially across breeds, reflecting genuine differences in body proportions, for example, limb bones (6-7, 6-10, 6-11) of the Bernese Mountain Dog are roughly three times longer than those of the Pomeranian. To further assess shape estimation consistency, we apply t-SNE to the 20-dimensional per clip mean bone-length vectors (Figure 6b). Clips from the same dog form tight, well separated clusters, confirming that

Table 8: Audio modality ablation. Removing audio conditioning leads to consistent degradation in motion quality and alignment, suggesting that audio provides complementary temporal cues for modeling human-pet interactions.

Input	$FID_k \downarrow$	$FID_s \downarrow$	$R^{hand}_{Prec.} \uparrow$	$R^{body}_{Prec.} \uparrow$	$Div_k \uparrow$	$Div_s \uparrow$
IPMG (w/o Audio)	13.69	14.24	0.60	0.58	5.73	5.81
IPMG (Full)	**11.21**	**12.96**	**0.63**	**0.59**	**5.93**	**6.01**

our skeleton estimation pipeline produces consistent body shapes within each individual and clearly distinguishes the 12 distinct body types.

Motion diversity. We extract motion features for each clip, including the mean pose, pose variability, and mean velocity, and project them using t-SNE (Figure 6c). In contrast to the shape embeddings, the motion embeddings exhibit broadly distributed and partially overlapping clusters across breeds. This suggests that while individual dogs display characteristic motion tendencies that form loosely grouped regions in the embedding space, the overall motion distribution remains diverse, spanning a wide range of behaviors across the dataset.

E Audio Modality Ablation

To evaluate the contribution of audio conditioning, we compare the full IPMG model (body + hand + audio) with a variant that removes the audio input. As shown in Table 8, the absence of audio results in consistent degradation across motion quality, alignment, and diversity metrics. The increase in FID_k and the decrease in $R^{hand}_{Prec.}$ indicate that audio provides complementary temporal cues that help align human gestures with corresponding pet responses. This is particularly relevant in voice-driven interactions such as calling or commanding.

F Dog Motion FID Model

Following the standard practice of using a pretrained classifier as a feature extractor for Fréchet Inception Distance (FID) [12], we train a dog identity classifier on our motion dataset to serve as the backbone for motion FID computation. The classifier takes normalized motion sequences of shape $(T, 60)$ (20 keypoints $\times$ 3 coordinates) as input, and processes them through a 1D convolutional ResNet encoder (4 residual blocks, 512 channels) with temporal downsampling. Global average pooling is applied to the encoded features to obtain a 512-dimensional representation, which is then mapped to 12 dog identity classes via a linear head, trained with cross entropy loss. After training, we discard the classification head and use the penultimate 512-dimensional features to compute FID between the generated and ground-truth motion distributions, measuring both the quality and diversity of the generated motions.

G Ethical Statement

Human participants. All human participants provided written informed consent prior to data collection. The study protocol was reviewed and approved by the local Institutional Review Board. Participants were informed of the data capture procedure, the intended use of the data for research purposes, and their right to withdraw at any time.

Animal welfare. All dog participants were accompanied by their owners or trained handlers throughout the recording sessions. No aversive training methods or restraints were used. Dogs were free to move voluntarily within the capture space and were given breaks between sessions. The interaction protocol was designed in consultation with professional dog trainers to ensure that all tasks fall within the range of typical obedience exercises.

Data release. The dataset will be released under a license for non-commercial research purposes. Users must agree to a data use agreement prohibiting redistribution and use for surveillance or biometric identification.

H Additional Qualitative Results (Supplementary Video)

Additional qualitative results are provided in the supplementary video (`interpet4d_supp_video_final.mp4`).