

AniMo: Species-Aware Model for Text-Driven Animal Motion Generation

Supplementary Material

The supplementary material is structured as follows:

- Section A: Presents the joint definitions in the AniMo4D dataset.
- Section B: Displays the prompt used for generating descriptions.
- Section C: Investigates the influence of RVQ parameters on the experimental results.
- Section D: Compares AniMo and the baseline models in case studies to further validate the model’s effectiveness.
- Section E: Summarizes user preferences for motions generated by AniMo compared to other baseline models.
- videos folder: Includes Figure_4.mp4, which is the video corresponding to Figure 4 in the main text, and Case_Study.mp4, which is the video for the case study.

A. Joint Definitions

We provide the joint definitions used in the AniMo4D dataset, which contains 30 joints. As illustrated in Figure A1, each joint is indexed and labeled with its corresponding name.

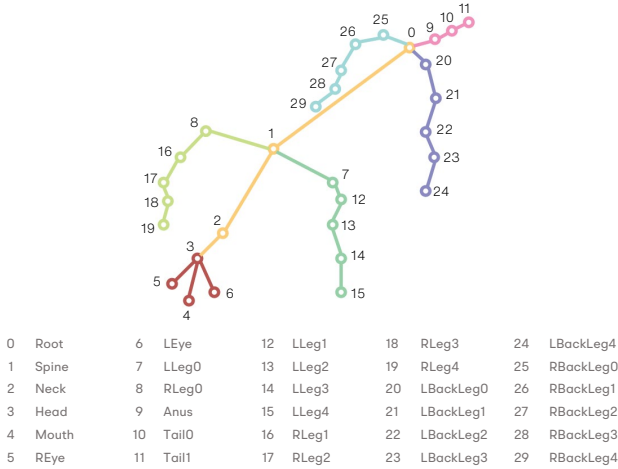


Figure A1. Joint definitions in the AniMo4D dataset.

B. Prompt

To generate textual descriptions, we utilized Llama [8], Qwen [1], and Deepseek [24] to expand motion labels, animal species, and attributes into detailed sentences describing animal motions. The prompt used is as follows:

```
messages=[
  {
    "role": "user",
    "content":
      f"Please generate a sentence about an-
      imal motion based on the following format:
      [{species}], [{attribute}], [{motion label}]
      Rules:
      1. Each motion should be a complete and logical
      movement.
      2. If multiple motions are present, they must be se-
      quential and logically connected.
      3. Use proper prepositions for motion directions.
      Examples:
      Input: [lion, male, walk stop]
      Output: The male lion walks and then stops.
      Input: [tiger, female, climb tree]
      Output: The female tiger climbs the tree.
      Input: [wolf, male, run sit]
      Output: The male wolf runs before sitting down.
      Your input:[{species}], [{gender}], [{motion label}]"
  }
]
```

This process resulted in the generation of 234,447 textual descriptions. We then manually reviewed the data, removing duplicate and semantically incomplete sentences. Additionally, we refined the textual descriptions to better align with the actual motion sequences, yielding a final dataset of 185,435 descriptions.

C. Ablation Study of RVQ Parameters

We analyze the impact of the residual layer depth V and the codebook size $|C|$ on the experimental results. This includes evaluating both the *reconstruction* in the first stage and the complete *generation* pipeline. As shown in Table C1, increasing V improves the performance in the reconstruction stage. However, this comes at the cost of added complexity in the text-to-motion generation stage, which negatively affects the overall performance. Similarly, moderately increasing $|C|$ leads to better performance, but an excessively large $|C|$ results in performance degradation.

D. Case Study

To further evaluate the effectiveness of our model, we conducted case studies comparing AniMo with six baseline models: T2M [11], MDM [47], T2M-GPT [54], AttT2M

Table C1. Exploring the impact of RVQ parameter selection on the experimental results using the AniMo4D test set.

AniMo ($V, C $)	Reconstruction			Generation		
	FID↓	MPJPE↓	Top-1↑	FID↓	MM-Dist↓	Top-1↑
AniMo (2, 512)	0.097 \pm .000	0.330 \pm .002	0.726 \pm .002	0.063 \pm .001	1.196 \pm .002	0.708 \pm .002
AniMo (4, 512)	0.050 \pm .000	0.321 \pm .002	0.776 \pm .002	0.056 \pm .001	1.125 \pm .001	0.741 \pm .002
AniMo (6, 512)	0.035 \pm .000	0.206 \pm .002	0.798 \pm .002	0.029 \pm .000	1.063 \pm .002	0.774 \pm .002
AniMo (8, 512)	0.026 \pm .000	0.322 \pm .002	0.802 \pm .002	0.030 \pm .000	1.221 \pm .002	0.779 \pm .002
AniMo (10, 512)	0.032 \pm .000	0.237 \pm .002	0.804 \pm .002	0.032 \pm .000	1.067 \pm .001	0.782 \pm .002
AniMo (12, 512)	0.023 \pm .000	0.141 \pm .001	0.813 \pm .002	0.046 \pm .001	1.152 \pm .001	0.792 \pm .002
AniMo (6, 256)	0.037 \pm .000	0.228 \pm .002	0.786 \pm .002	0.037 \pm .000	1.083 \pm .001	0.766 \pm .002
AniMo (6, 512)	0.035 \pm .000	0.206 \pm .002	0.798 \pm .002	0.029 \pm .000	1.063 \pm .002	0.774 \pm .002
AniMo (6, 1024)	0.024 \pm .000	0.221 \pm .001	0.797 \pm .002	0.022 \pm .000	1.083 \pm .002	0.768 \pm .002
AniMo (6, 1536)	0.038 \pm .000	0.173 \pm .001	0.776 \pm .002	0.030 \pm .001	1.092 \pm .001	0.762 \pm .002
AniMo (6, 2048)	0.065 \pm .000	0.249 \pm .001	0.711 \pm .002	0.070 \pm .001	1.201 \pm .001	0.703 \pm .002

[61], MMM [34], and MoMask [12]. Three representative cases were designed for this comparison, as shown in the file `videos/Case_Study.mp4`.

One example, illustrated in Figure D2, evaluates the models’ ability to generate motions for the prompt: “The female cheetah fights, reacts swiftly to avoid danger, and then dies.” This scenario is challenging as it involves a sequence of three distinct motions. The results reveal notable differences among the models. Motions generated by MDM are largely unrelated to the given text description. T2M fails to capture the “die” motion entirely. Both T2M-GPT and AttT2M generate all three motions; however, their outputs exhibit inconsistent skeletal structures, with noticeable fluctuations in the animal’s bone lengths over time. MMM and MoMask perform better overall but are still affected by visible motion jitter. In contrast, AniMo produces the most stable and semantically accurate motions. Specifically, AniMo effectively produces rear-leg standing and front-leg swinging movements, which align closely with the actions described in the prompt, such as “fights” and “reacts swiftly to avoid danger.” These results underscore AniMo’s superior ability to generate complex, coherent, and high-quality motion sequences.

E. User Study

To further validate our conclusions, we conducted a user study. This study involved 23 volunteers and compared AniMo with several baseline methods, including T2M [11], MDM [47], T2M-GPT [54], AttT2M [61], MMM [34], and MoMask [12]. For each method, we generated 50 motions using the same text pool from the AniMo4D test set. Feedback was collected from three different users for each comparison to ensure diverse and unbiased evaluations. As shown in Figure E3, AniMo consistently achieved high rankings, demonstrating its ability to produce coher-

ent, accurate, and semantically aligned motion sequences that meet user expectations.

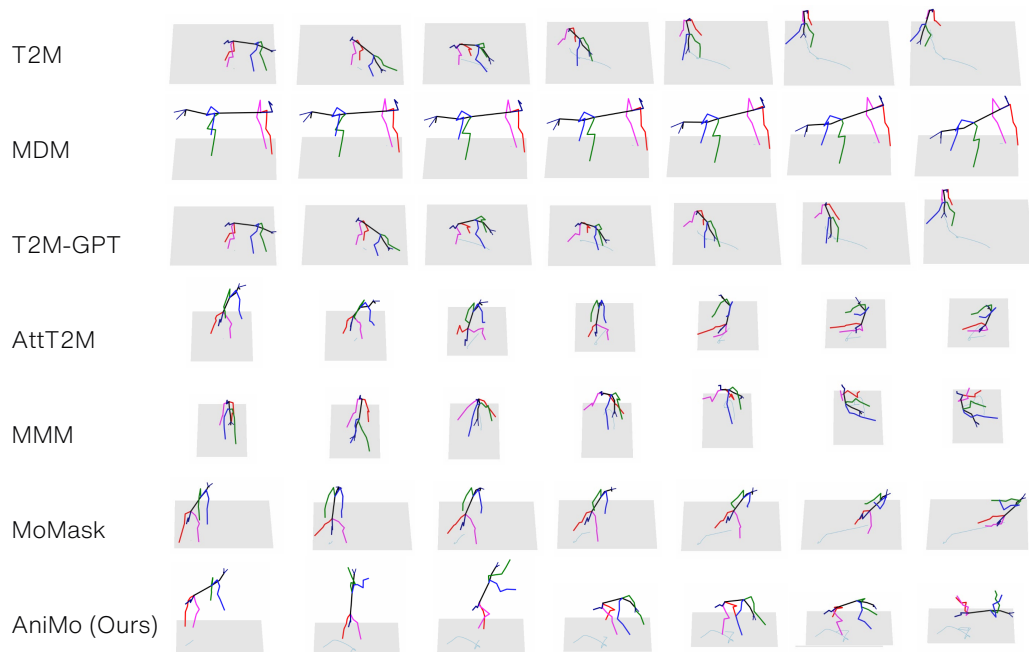


Figure D2. Motions generated by different models for the text description: “The female cheetah fights, reacts swiftly to avoid danger, and then dies.”

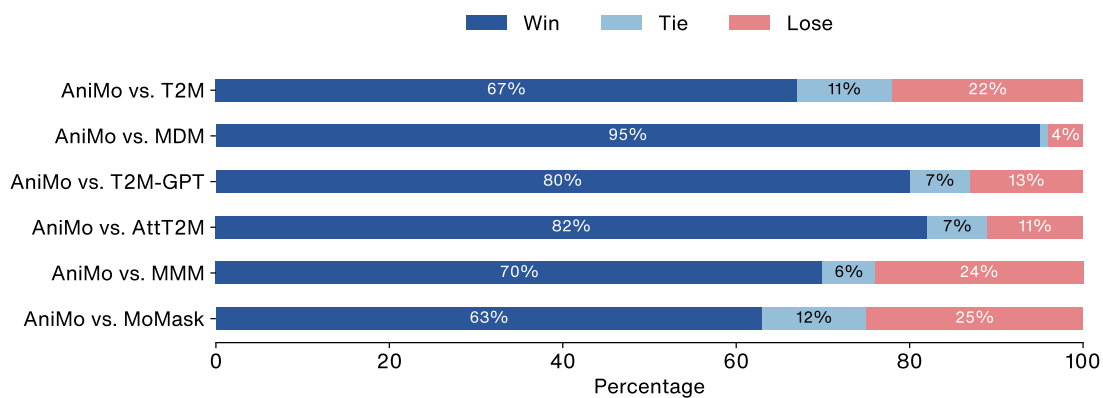


Figure E3. Human preference evaluation results comparing AniMo against baseline models using pairwise comparisons. Each bar represents the percentage of wins (blue), ties (light blue), and losses (pink) when comparing AniMo with other models. The results demonstrate AniMo’s superior performance in generating high-quality animal motions that align with human preferences.