AvatarGPT: All-in-One Framework for Motion Understanding, Planning, Generation and Beyond

— Supplementary Materials

Zixiang Zhou
zhouzixiang@xiaobing.ai

Yu Wan
wanyu@xiaobing.ai

Baoyuan Wang
wangbaoyuan@xiaobing.ai

Xiaobing.AI

A. Prompts for Instruction Tuning

We define seven major motion-related tasks, including 1) motion generation(MG), 2) motion understanding(MU), 3) motion-in-between(MiB), 4) task planning(CT2T), 5) task decomposition(T2S), 6) task summarization(S2T), and 7) scene estimation(T2C). We also define another four high-level tasks as auxiliary in training to assist in learning the comprehensive relationship between motion and language descriptions. These auxiliary tasks are: 1) steps planning conditioned on historical task(CT2S), 2) task planning conditioned on historical steps(CS2T), 3) steps planning conditioned on historical steps(CS2S), 4) scene estimation from steps(S2C). The notation of ‘C’, ‘S’, and ‘T’ are identical as previously stated.

The prompts employed for instruction tuning the shared LLM on these tasks are shown in Fig. 1.

B. Automatic Annotation Pipeline

The proposed annotation pipeline contains five modules covering video preprocessing and video describing at different levels of detail.

Video Processing The raw input videos could be up to hours long, which contain rich content. It is not practical to describe the content of the entire video at once. Therefore, we perform the video cropping to break down it into multiple short video segments. Experiments show that breaking down into 10-second chunks balances between efficiency and annotation richness.

Video Content Describing We use Ask-Anything[1] to describe the content of each video segment. It generates one long textual description with up to 500 words to depict the person’s activity, appearance, and environment shown in the video segment. Fig. 2 presents the prompts used for this task.

Motion Describing We use ChatGPT to extract key information from the video content description. What we are interested in are those describing human motion and activity. The experiment shows that hierarchical extraction obtains results that meet our requirements best. Specifically, we ask ChatGPT to extract information that describes how the person in the video executes the activity step-by-step. It produces several short descriptions at the finest text granularity(**step description**). Following this, we ask ChatGPT to summarize the contents of these step-by-step descriptions, the results depict the specific activities at medium-grained text granularity(**task description**). We also generate content description that depicts the scene information of the entire video at coarse-grained granularity(**scene description**). For this purpose, we feed the task descriptions to ChatGPT, we emphasize their order because temporal sequence matters.

We use the following prompts shown in Fig. 3 to extract step-, task- and scene-level descriptions.

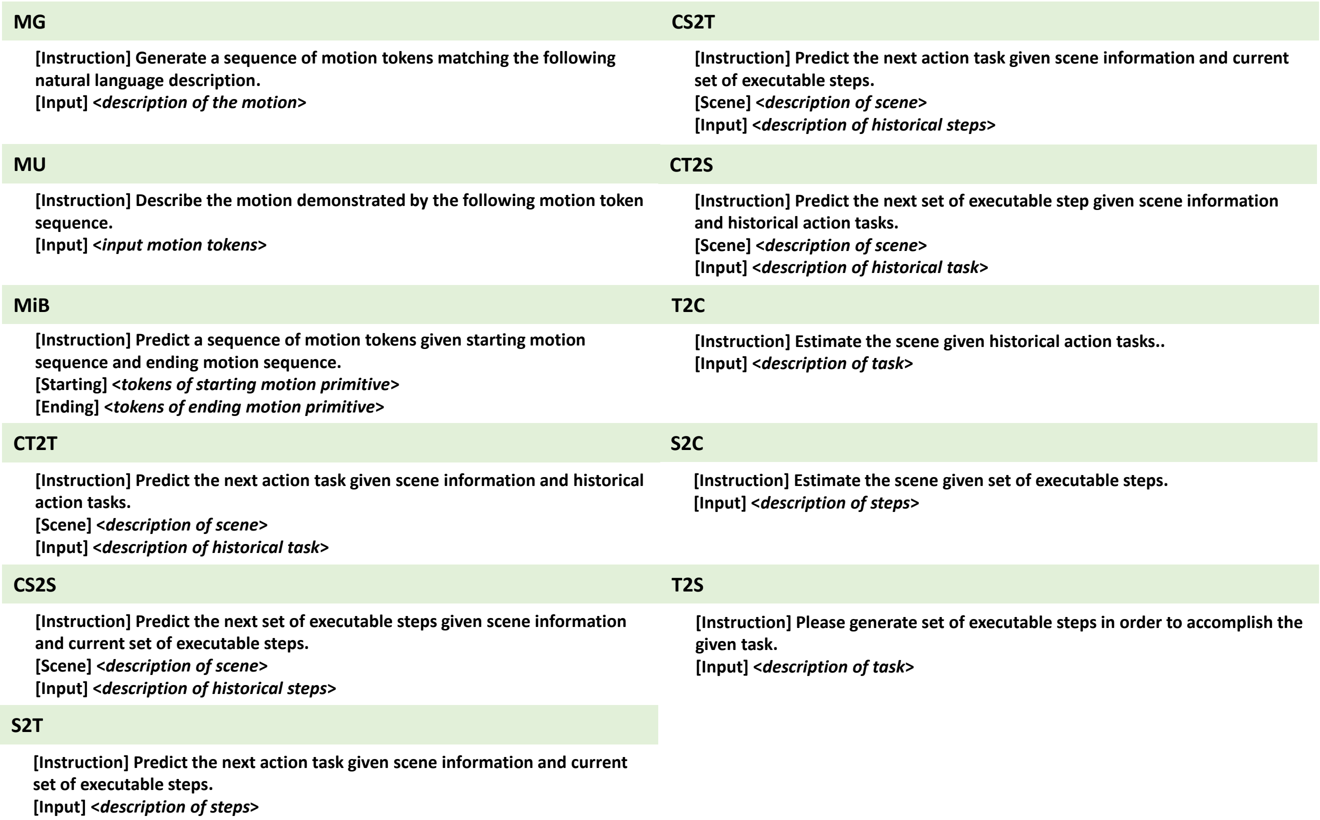
C. Metrics for Low-level Tasks

We describe the metrics for linguistic consistency evaluation here.

BertScore[6] uses a pretrained BERT model to evaluate the similarity between generated and reference text. It encodes the reference and target sentence to embeddings and estimates the cosine similarity between embeddings are the similarity score.

BLUE[3] measures the sub-sentence level(1-gram, 4-gram) correctness between candidate sentence and set of references. Although it was proposed for translation initially, it is widely adopted in assessing text generation.

ROUGE-L[2] assesses the similarity between two sentences according to the longest common sub-



MG

[Instruction] Generate a sequence of motion tokens matching the following natural language description.
[Input] <*description of the motion*>

MU

[Instruction] Describe the motion demonstrated by the following motion token sequence.
[Input] <*input motion tokens*>

MiB

[Instruction] Predict a sequence of motion tokens given starting motion sequence and ending motion sequence.
[Starting] <*tokens of starting motion primitive*>
[Ending] <*tokens of ending motion primitive*>

CT2T

[Instruction] Predict the next action task given scene information and historical action tasks.
[Scene] <*description of scene*>
[Input] <*description of historical task*>

CS2S

[Instruction] Predict the next set of executable steps given scene information and current set of executable steps.
[Scene] <*description of scene*>
[Input] <*description of historical steps*>

S2T

[Instruction] Predict the next action task given scene information and current set of executable steps.
[Input] <*description of steps*>

CS2T

[Instruction] Predict the next action task given scene information and current set of executable steps.
[Scene] <*description of scene*>
[Input] <*description of historical steps*>

CT2S

[Instruction] Predict the next set of executable step given scene information and historical action tasks.
[Scene] <*description of scene*>
[Input] <*description of historical task*>

T2C

[Instruction] Estimate the scene given historical action tasks..
[Input] <*description of task*>

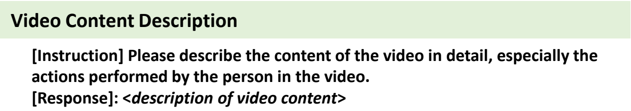
S2C

[Instruction] Estimate the scene given set of executable steps.
[Input] <*description of steps*>

T2S

[Instruction] Please generate set of executable steps in order to accomplish the given task.
[Input] <*description of task*>

Figure 1. **Prompts for Instruction Tuning.** We present the prompts we adopted for instruction tuning our model. $\langle text \rangle$ indicates the input text/motion tokens and output text/motion tokens.



Video Content Description

[Instruction] Please describe the content of the video in detail, especially the actions performed by the person in the video.
[Response]: <*description of video content*>

Figure 2. **Prompts for Video Describing.** We generate content descriptions using this prompt. $\langle text \rangle$ indicates the output description.

sequence(LCS).

CIDEr[5] measures the similarity based on the concept of consensus in terms of word, grammar, and text content. It first estimates the BLUE score between target and references and modifies the scores using inverse document frequency(IDF) weighting[4] to balance the weights between rare and common words. Finally, it averages the weighted scores as the final CIDEr score.

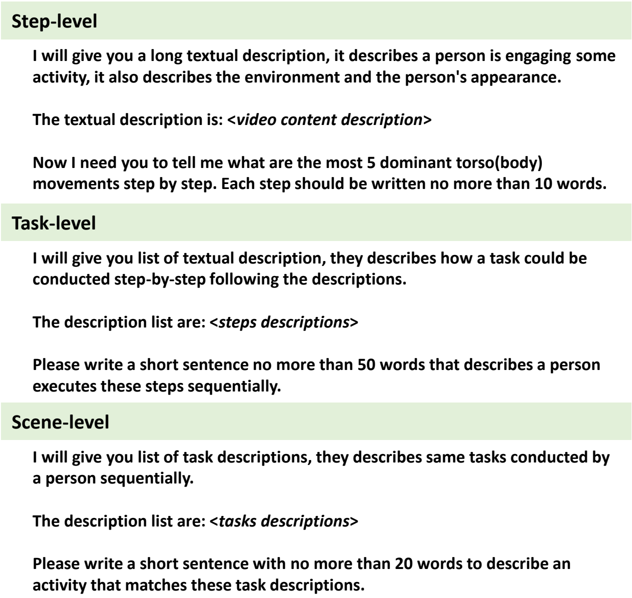
D. Metrics for High-level Tasks

The logical coherence between generated text and conditions is the primary concern. For instance, if the condition scene is *'indoor fitness exercise'* , and the historical activity is *'push-ups'*, one logically coherent activity could be *' jumping jacks'*, while *'organizing items on the ground'* is incoherent. We harness the capability of understanding and reasoning of ChatGPT to evaluate the consistency. Fig. 4 illustrates how ChatGPT is employed to assist the logical coherency evaluation on high-level tasks. Fig. 5 shows the prompts used for each evaluation task.

E. Full Pipeline Evaluation

The full pipeline of our method includes task planning, decomposition, motion generation, understanding, in-between, etc. We conduct quantitative evaluation and user study to assess the performance. The entire evaluation workflow is shown in Fig. 6. We define a forward path as scene → task → steps → motions, where task planning, decomposition, motion generation, and in-between are conducted sub-sequentially. The backward path executes the motion understanding and task summarization one after another, resulting in the following outputs: motion → steps → task.

We perform quantitative analysis and human evaluation to assess the linguistic consistency(Ling. Consis.) between planned and summarized tasks and decomposed and



Step-level

I will give you a long textual description, it describes a person is engaging some activity, it also describes the environment and the person's appearance.

The textual description is: <*video content description*>

Now I need you to tell me what are the most 5 dominant torso(body) movements step by step. Each step should be written no more than 10 words.

Task-level

I will give you list of textual description, they describes how a task could be conducted step-by-step following the descriptions.

The description list are: <*steps descriptions*>

Please write a short sentence no more than 50 words that describes a person executes these steps sequentially.

Scene-level

I will give you list of task descriptions, they describes same tasks conducted by a person sequentially.

The description list are: <*tasks descriptions*>

Please write a short sentence with no more than 20 words to describe an activity that matches these task descriptions.

Figure 3. **Prompts for Motion Describing.** We present the prompts for hierarchical annotation generation. $\langle text \rangle$ indicates the input/output texts.

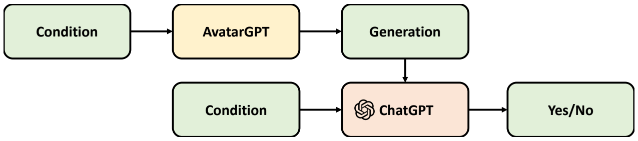


Figure 4. **Logical Coherency Evaluation.** We ask ChatGPT to determine whether the generation matches the condition in terms of logical coherency.

described step descriptions. Specifically, we estimate the BertScore, BLUE, ROUGE-L, and CIDEr scores as quantitative metrics, and we ask human participants to judge whether the target and reference sentences have similar semantic content. Score '1' is expected if semantic similarity exists, otherwise, '0'.

In addition, quantitative and human evaluations are also executed to measure the logical coherent score(LCS) in terms of task planning and decomposition in the forward path. For quantitative analysis, the same approach discussed in Appendix D is adopted. For human evaluation, users are expected to score '1' if they believe the generated texts are logically coherent to condition texts, otherwise, a score '0' is given.

We also conducted a user study to evaluate the consistency between generated long motion sequences and condition descriptions. Because the final motion is the combination of synthesized short motion segments corresponding to multiple condition descriptions, and the blended motions in-between, it is impractical to simply give a yes/no grade. We propose a rating scale of 1-5 for this task. Specifically, users are expected to score '5' if the generated motion presents all contents described in corresponding conditions. On the contrary, score '1' is expected if none of the descriptions are correctly presented in the generated motion. For those motions that present partial conditions, scores between '1' and '5' are expected according to the percentage of conditions synthesized.

F. Results of Automatic Video Annotation

We show one example of video annotation results in Fig. 7. Our annotation pipeline crops the input video into segments. We run multiple rounds on each segment to generate descriptions of diverse expressions. Specifically, we run six rounds per segment, resulting in 6 task descriptions and 30 step descriptions. We run twenty rounds on the entire video(task descriptions of all the video segments) to obtain multiple scene descriptions. For brevity, we show the results of two randomly selected segments, each with one task and five-step descriptions. We also show three scene descriptions.

G. More Results of Motion Generation, Understanding, and In-Between

We show more results of motion generation, motion understanding, and motion-in-between tasks in Fig. 8, 9 10. For motion generation and understanding tasks, we highlight the key words to emphasize the semantics. For motion-in-between, we use different colors for interpolated frames.

H. Motion Generation with Various Text Granularity

We show more results on generating motion sequences from descriptions of various granularity in Fig. 11, 12. The scene of Fig. 11 is some regular activities performed in an office space, and that of Fig. 12 is an indoor fitness workout. We show the motion synthesis conditioned on scene description(coarse-grained), task descriptions(medium-grained), and step descriptions(fine-grained), respectively. We observe high consistency between synthesized motion and conditions, regardless of their text granularities.

I. Motion Understanding of Various Levels of Detail

We show more results in describing motion sequences at various levels of detail in Fig. 13, 14. We first show the results of describing the activity displayed by long motion sequences(500+ frames) at medium-grained granularity. Then we present the results of describing the specific actions demonstrated in the short motion segments(≈200 frames) at fine-grained granularity. These results suggest that our method has a great ability to understand human movements at various levels of detail.

J. Limitation and Future Work

1) The training of high-level tasks and low-level tasks are not strictly joint. There is a domain gap between the datasets for low-level and high-level training. Building a larger dataset covering both low- and high-level is one of the top priorities of our future research. 2) Our method is only able to conduct torso movement-related tasks. We believe that facial expression-related tasks are of equal importance. For instance, generating head poses and facial expression sequences subject to the conversation is a good complementary to the torso and gesture movement synthesis. We argue these tasks have the potential to be integrated into one shared pipeline. We will investigate the unification of body, head movements, facial expression, and related high-level tasks in the future.

CT2T

Suppose it is under specific scenario:

[Scene]: <*condition scene description*>

There are two descriptions that describe a person is doing something activity.

[Task 1]: <*condition task description*>
[Task 2]: <*generated task description*>

Please tell me whether it is logically possible that this person can execute these two tasks sequentially under the specific scenario with word 'Yes' or 'No'.

CS2S

Suppose it is under specific scenario:

[Scene]: <*condition scene description*>

There are two descriptions that describe a person is doing something by executing the descriptions step by step.

[Steps 1]: <*condition steps description*>
[Steps 2]: <*generated steps description*>

Please tell me whether it is logically plausible that this person can execute these step descriptions sequentially under the provided scenario with word 'Yes' or 'No'.

CS2T

Suppose it is under specific scenario:

[Scene]: <*condition scene description*>

There are two descriptions, the first one are list of executable action steps, and the second one is a sentence that describes somebody is conducting some task.

[Step]: <*condition steps description*>
[Task]: <*generated task description*>

Please tell me whether it is logically plausible that this person can conduct this task after he(she) executes those action steps under the provided scenario with word 'Yes' or 'No'.

CT2S

Suppose it is under specific scenario:

[Scene]: <*condition scene description*>

There are two descriptions, the first one describes somebody is conducting some task, and the second one is a list of executable action steps.

[Task]: <*condition task description*>
[Steps]: <*generated steps description*>

Please tell me whether it is logically plausible that this person can execute these action steps sequentially after he(she) finished the first task under the provided scenario with word 'Yes' or 'No'.

T2C

I will give you a sentence that describes somebody is conducting some task.

[Task]: <*condition task description*>

I will also give you a sentence that describes a specific scene.

[Scene]: < *generated scene description*>

Please tell me whether this task is logically consistent with the given scenario with word 'Yes' or 'No'.

S2C

I will give you a list of action steps descriptions.

[Steps]: <*condition steps description*>

I will also give you a sentence that describes a specific scene.

[Scene]: < *generated scene description*>

Please tell me whether these action steps are logically consistent with the given scenario with word 'Yes' or 'No'.

T2S

I will give you a task description that describes somebody is conducting some task.

[Task]: <*condition task description*>

I will also give you a list of action step descriptions.

[Steps]: < *generated steps description*>

Please tell me whether a person can conduct this specific task by executing these action steps with word 'Yes' or 'No'.

S2T

I will give you a list of action step descriptions.

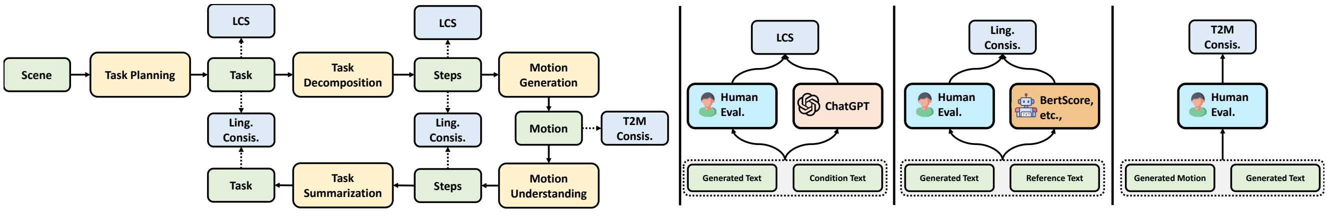
[Steps]: <*condition steps description*>

I will also give you a task description that describes somebody is conducting some task.

[Task]: < *generated task description*>

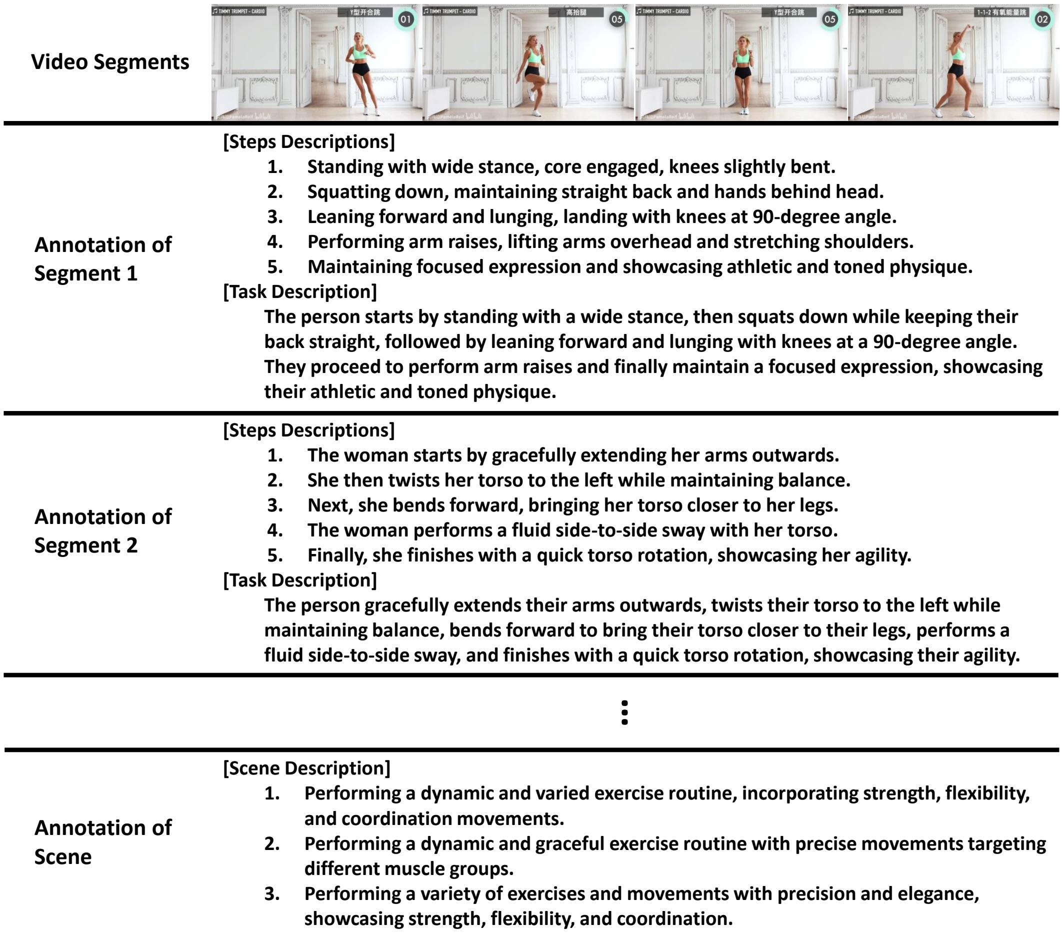
Please tell me whether a person can conduct this specific task by executing these action steps with word 'Yes' or 'No'.

Figure 5. **Prompts for High-level Tasks Evaluation.** We show prompts we employ to conduct automatic high-level task evaluation. $\langle text \rangle$ indicates the input/output texts.



(a) Full Evaluation Pipeline (b) LCS Evaluation (c) Ling. Consis. Evaluation (d) T2M Consis. Evaluation

Figure 6. **Method for Full Pipeline Evaluation.** We measure the performance of our method in the scope of the full pipeline based on the concept of cycle consistency. (a) Evaluation pipeline. (b) Evaluate the logical coherent score(LCS) using human evaluation and ChatGPT. (c) Evaluate the linguistic consistency(Ling. Consis.) using human evaluation and quantitative metrics(e.g. BertScore). (d) Evaluate the text-to-motion consistency(T2M Consis.) using human evaluation.



Video Segments

Annotation of Segment 1

[Steps Descriptions]

1. Standing with wide stance, core engaged, knees slightly bent.
2. Squatting down, maintaining straight back and hands behind head.
3. Leaning forward and lunging, landing with knees at 90-degree angle.
4. Performing arm raises, lifting arms overhead and stretching shoulders.
5. Maintaining focused expression and showcasing athletic and toned physique.

[Task Description]

The person starts by standing with a wide stance, then squats down while keeping their back straight, followed by leaning forward and lunging with knees at a 90-degree angle. They proceed to perform arm raises and finally maintain a focused expression, showcasing their athletic and toned physique.

Annotation of Segment 2

[Steps Descriptions]

1. The woman starts by gracefully extending her arms outwards.
2. She then twists her torso to the left while maintaining balance.
3. Next, she bends forward, bringing her torso closer to her legs.
4. The woman performs a fluid side-to-side sway with her torso.
5. Finally, she finishes with a quick torso rotation, showcasing her agility.

[Task Description]

The person gracefully extends their arms outwards, twists their torso to the left while maintaining balance, bends forward to bring their torso closer to their legs, performs a fluid side-to-side sway, and finishes with a quick torso rotation, showcasing their agility.

⋮

Annotation of Scene

[Scene Description]

1. Performing a dynamic and varied exercise routine, incorporating strength, flexibility, and coordination movements.
2. Performing a dynamic and graceful exercise routine with precise movements targeting different muscle groups.
3. Performing a variety of exercises and movements with precision and elegance, showcasing strength, flexibility, and coordination.

Figure 7. **Annotation Results.** We show one task description and five-step descriptions for each video segment. We show three scene descriptions of the entire video.

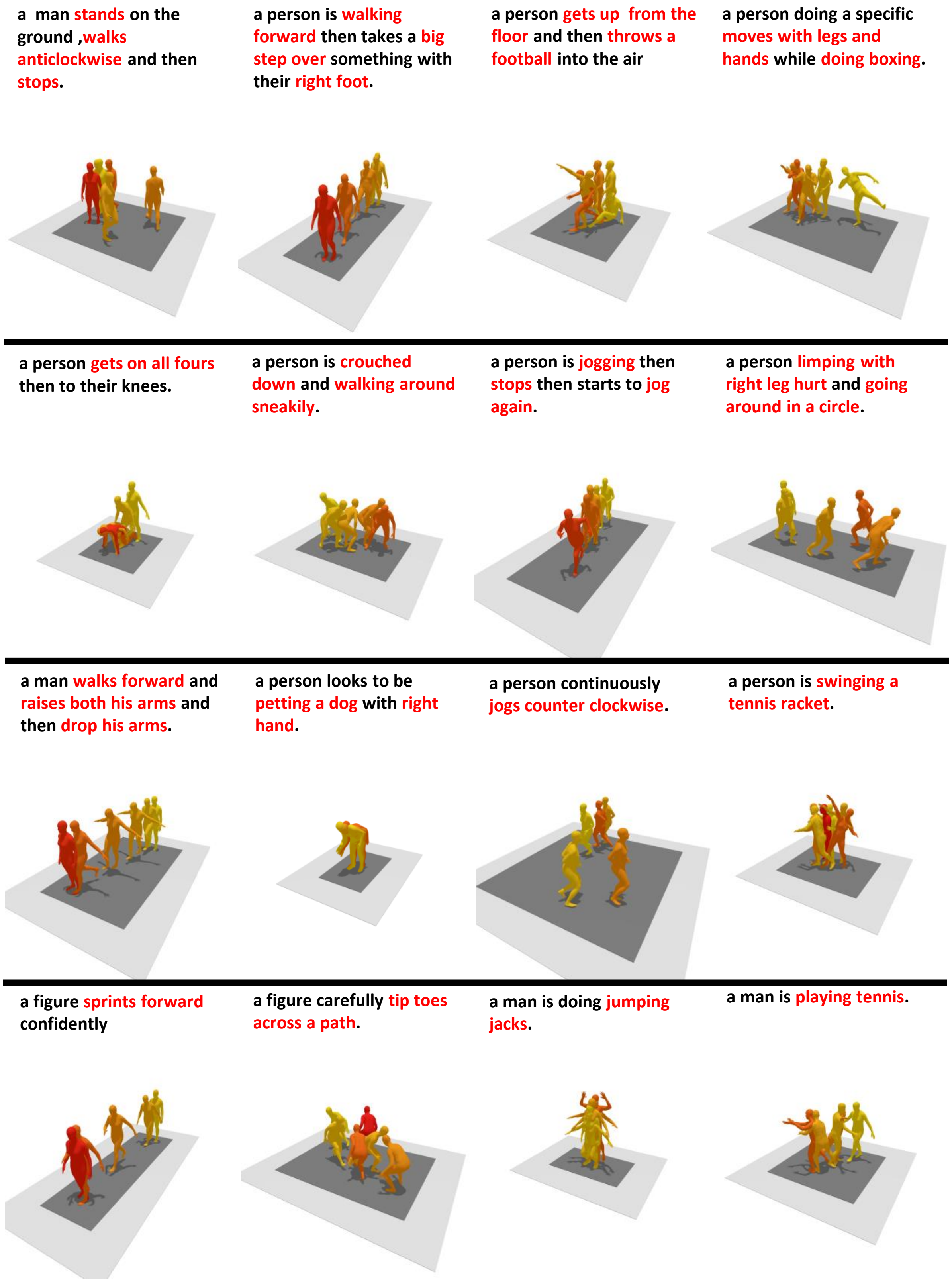
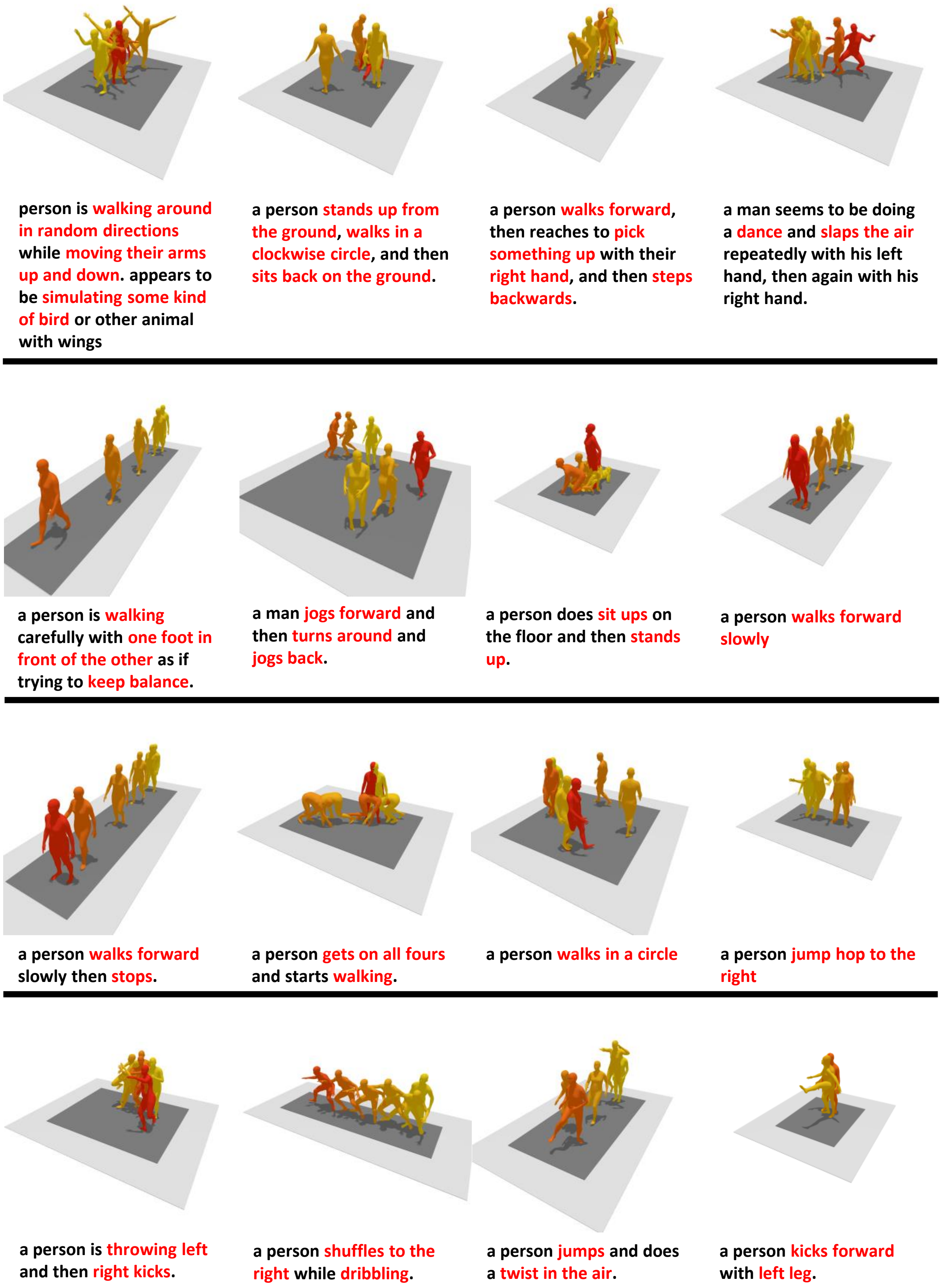

a man stands on the ground ,walks anticlockwise and then stops.
a person is walking forward then takes a big step over something with their right foot.
a person gets up from the floor and then throws a football into the air
a person doing a specific moves with legs and hands while doing boxing.
a person gets on all fours then to their knees.
a person is crouched down and walking around sneakily.
a person is jogging then stops then starts to jog again.
a person limping with right leg hurt and going around in a circle.
a man walks forward and raises both his arms and then drop his arms.
a person looks to be petting a dog with right hand.
a person continuously jogs counter clockwise.
a person is swinging a tennis racket.
a figure sprints forward confidently
a figure carefully tip toes across a path.
a man is doing jumping jacks.
a man is playing tennis.

Figure 8. **Results of Motion Generation.**



person is walking around in random directions while moving their arms up and down. appears to be simulating some kind of bird or other animal with wings

a person stands up from the ground, walks in a clockwise circle, and then sits back on the ground.

a person walks forward, then reaches to pick something up with their right hand, and then steps backwards.

a man seems to be doing a dance and slaps the air repeatedly with his left hand, then again with his right hand.

a person is walking carefully with one foot in front of the other as if trying to keep balance.

a man jogs forward and then turns around and jogs back.

a person does sit ups on the floor and then stands up.

a person walks forward slowly

a person walks forward slowly then stops.

a person gets on all fours and starts walking.

a person walks in a circle

a person jump hop to the right

a person is throwing left and then right kicks.

a person shuffles to the right while dribbling.

a person jumps and does a twist in the air.

a person kicks forward with left leg.

Figure 9. **Results of Motion Understanding.**

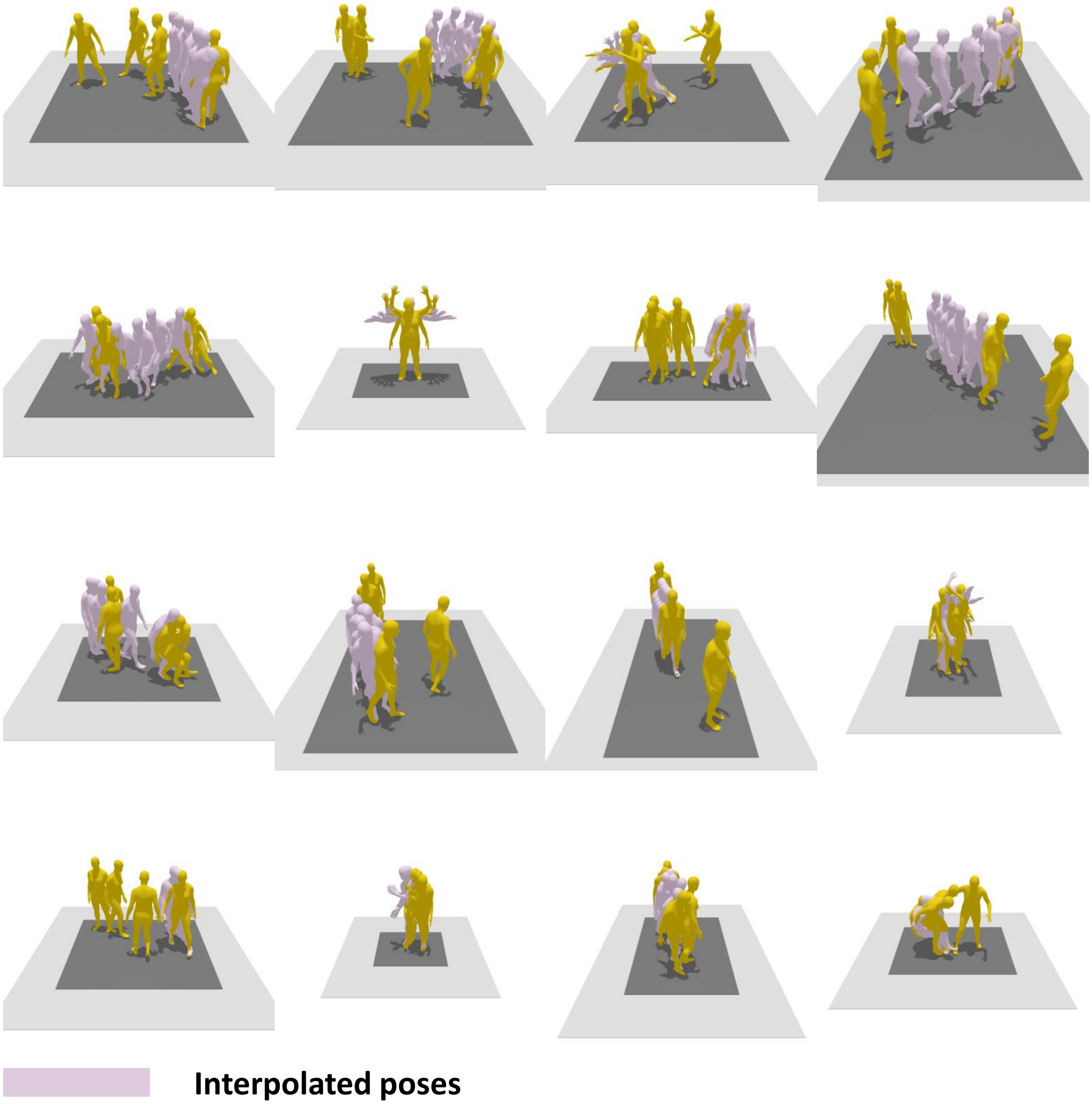
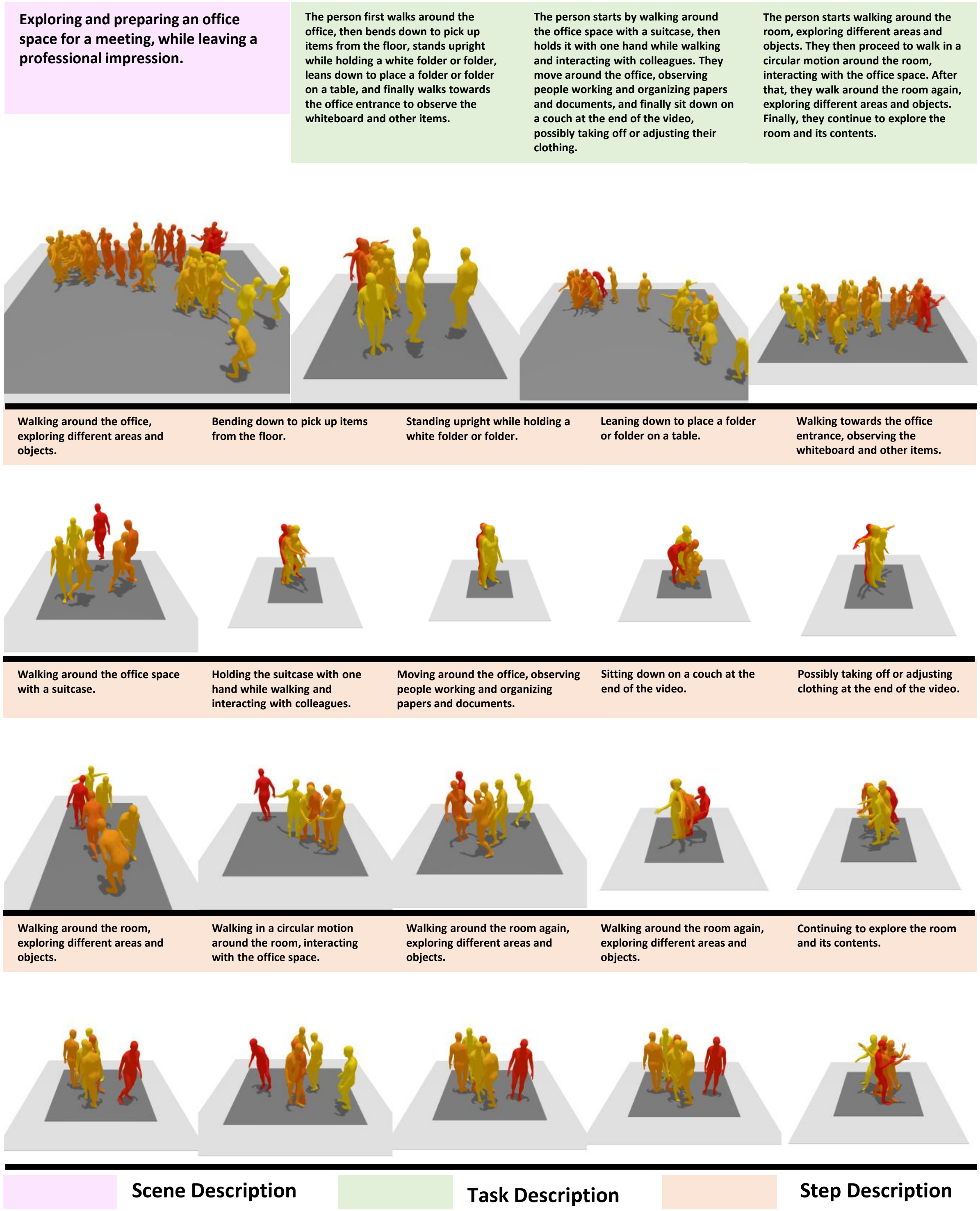


Figure 10. **Results of Motion-in-Between.**



Exploring and preparing an office space for a meeting, while leaving a professional impression.

The person first walks around the office, then bends down to pick up items from the floor, stands upright while holding a white folder or folder, leans down to place a folder or folder on a table, and finally walks towards the office entrance to observe the whiteboard and other items.

The person starts by walking around the office space with a suitcase, then holds it with one hand while walking and interacting with colleagues. They move around the office, observing people working and organizing papers and documents, and finally sit down on a couch at the end of the video, possibly taking off or adjusting their clothing.

The person starts walking around the room, exploring different areas and objects. They then proceed to walk in a circular motion around the room, interacting with the office space. After that, they walk around the room again, exploring different areas and objects. Finally, they continue to explore the room and its contents.

Walking around the office, exploring different areas and objects.

Bending down to pick up items from the floor.

Standing upright while holding a white folder or folder.

Leaning down to place a folder or folder on a table.

Walking towards the office entrance, observing the whiteboard and other items.

Walking around the office space with a suitcase.

Holding the suitcase with one hand while walking and interacting with colleagues.

Moving around the office, observing people working and organizing papers and documents.

Sitting down on a couch at the end of the video.

Possibly taking off or adjusting clothing at the end of the video.

Walking around the room, exploring different areas and objects.

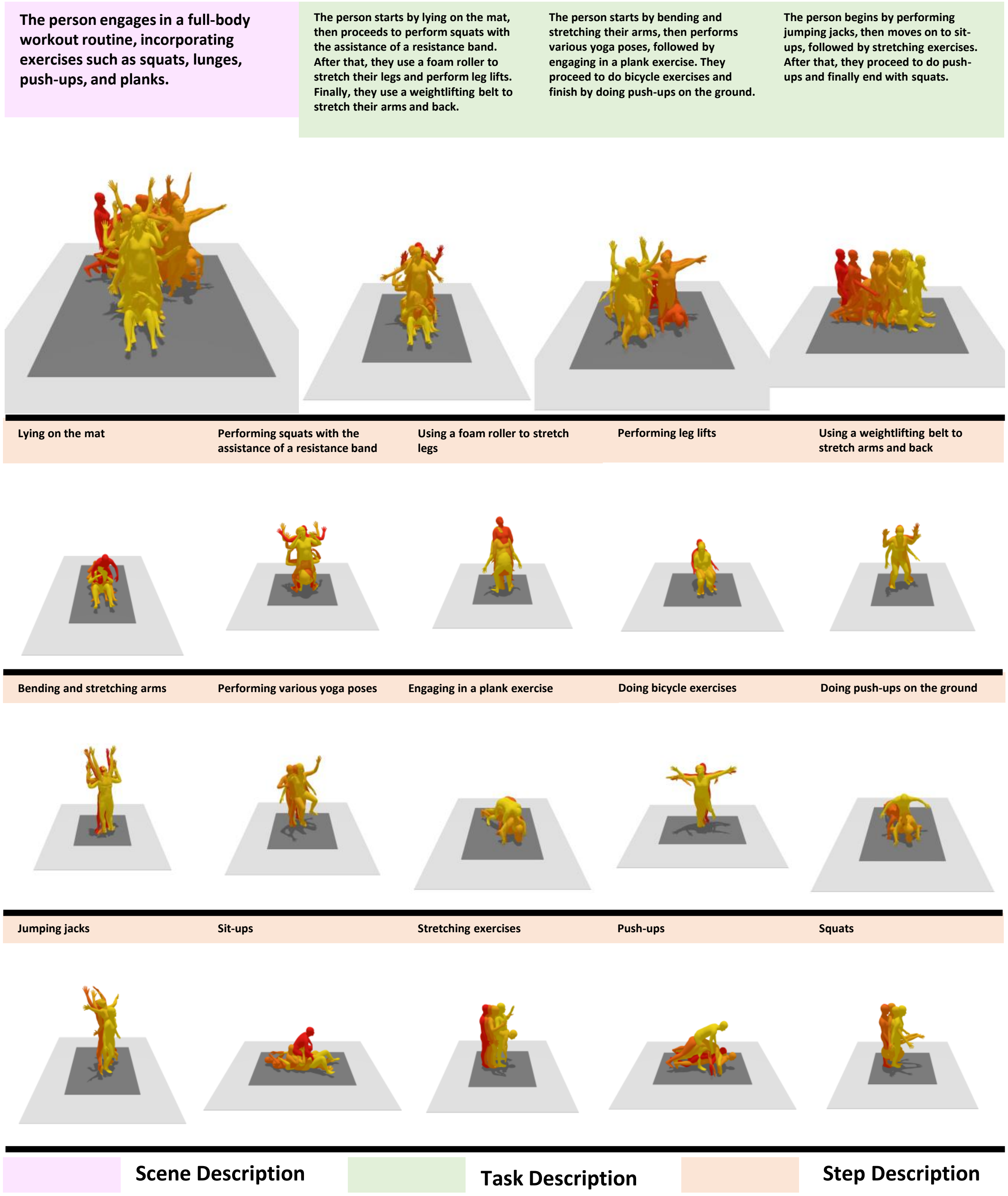
Walking in a circular motion around the room, interacting with the office space.

Walking around the room again, exploring different areas and objects.

Walking around the room again, exploring different areas and objects.

Continuing to explore the room and its contents.

Figure 11. **Motion Generation with Various Text Granularity.** Our method supports synthesizing human motion from various text granularity. Given a scene description(coarse-grained), our model generates a long motion sequence that matches the context. Given a task description(medium-grained), our method can synthesize a motion sequence that displays the corresponding activity. Given a step description(fine-grained), our method can generate a motion sequence that corresponds to the specific action.



The person engages in a full-body workout routine, incorporating exercises such as squats, lunges, push-ups, and planks.

The person starts by lying on the mat, then proceeds to perform squats with the assistance of a resistance band. After that, they use a foam roller to stretch their legs and perform leg lifts. Finally, they use a weightlifting belt to stretch their arms and back.

The person starts by bending and stretching their arms, then performs various yoga poses, followed by engaging in a plank exercise. They proceed to do bicycle exercises and finish by doing push-ups on the ground.

The person begins by performing jumping jacks, then moves on to sit-ups, followed by stretching exercises. After that, they proceed to do push-ups and finally end with squats.

Lying on the mat

Performing squats with the assistance of a resistance band

Using a foam roller to stretch legs

Performing leg lifts

Using a weightlifting belt to stretch arms and back

Bending and stretching arms

Performing various yoga poses

Engaging in a plank exercise

Doing bicycle exercises

Doing push-ups on the ground

Jumping jacks

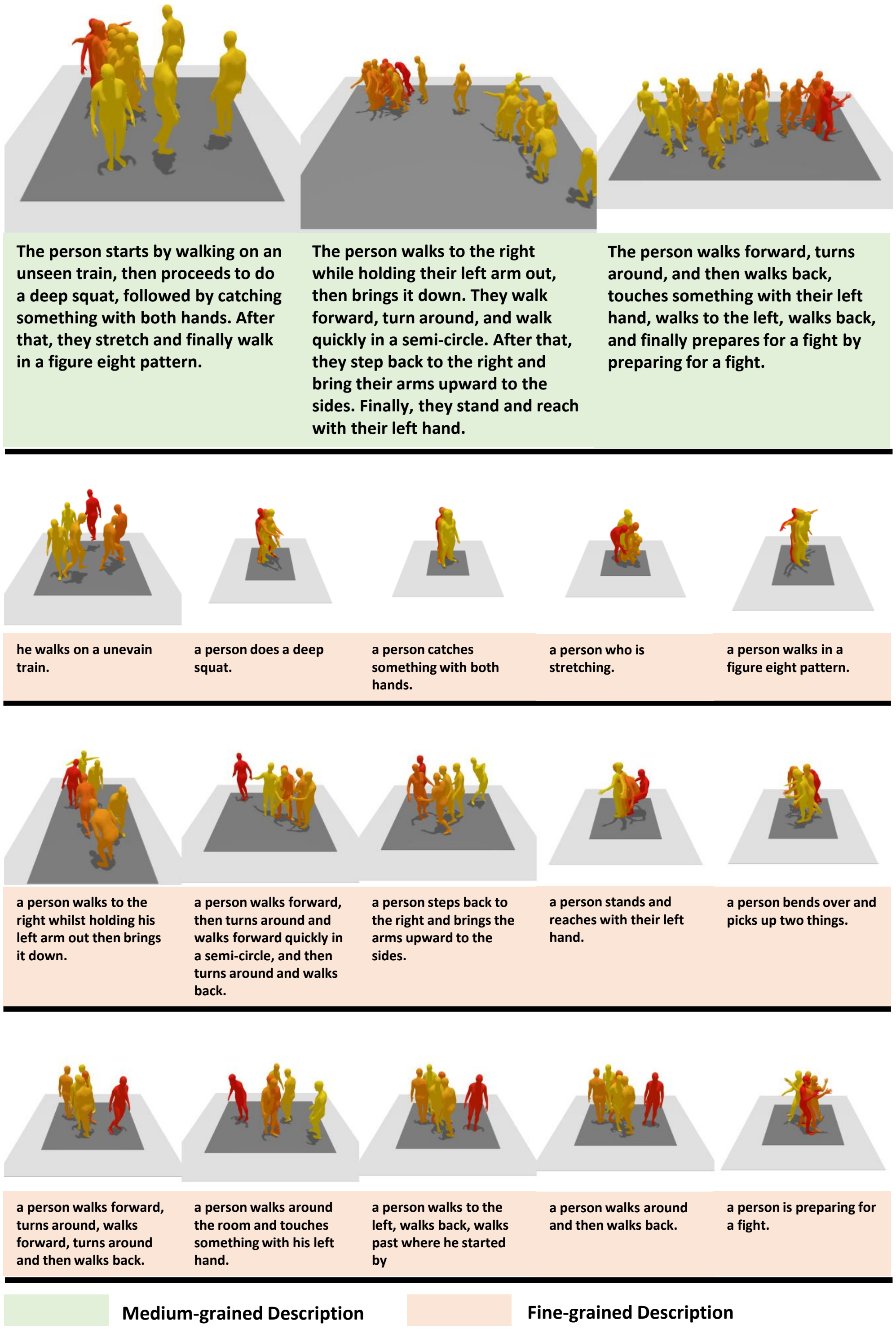
Sit-ups

Stretching exercises

Push-ups

Squats

Figure 12. **Motion Generation with Various Text Granularity.** Our method supports synthesizing human motion from various text granularity. Given a scene description(coarse-grained), our model generates a long motion sequence that matches the context. Given a task description(medium-grained), our method can synthesize a motion sequence that displays the corresponding activity. Given a step description(fine-grained), our method can generate a motion sequence that corresponds to the specific action.


The person starts by walking on an unseen train, then proceeds to do a deep squat, followed by catching something with both hands. After that, they stretch and finally walk in a figure eight pattern.

The person walks to the right while holding their left arm out, then brings it down. They walk forward, turn around, and walk quickly in a semi-circle. After that, they step back to the right and bring their arms upward to the sides. Finally, they stand and reach with their left hand.

The person walks forward, turns around, and then walks back, touches something with their left hand, walks to the left, walks back, and finally prepares for a fight by preparing for a fight.

he walks on a unevain train.

a person does a deep squat.

a person catches something with both hands.

a person who is stretching.

a person walks in a figure eight pattern.

a person walks to the right whilst holding his left arm out then brings it down.

a person walks forward, then turns around and walks forward quickly in a semi-circle, and then turns around and walks back.

a person steps back to the right and brings the arms upward to the sides.

a person stands and reaches with their left hand.

a person bends over and picks up two things.

a person walks forward, turns around, walks forward, turns around and then walks back.

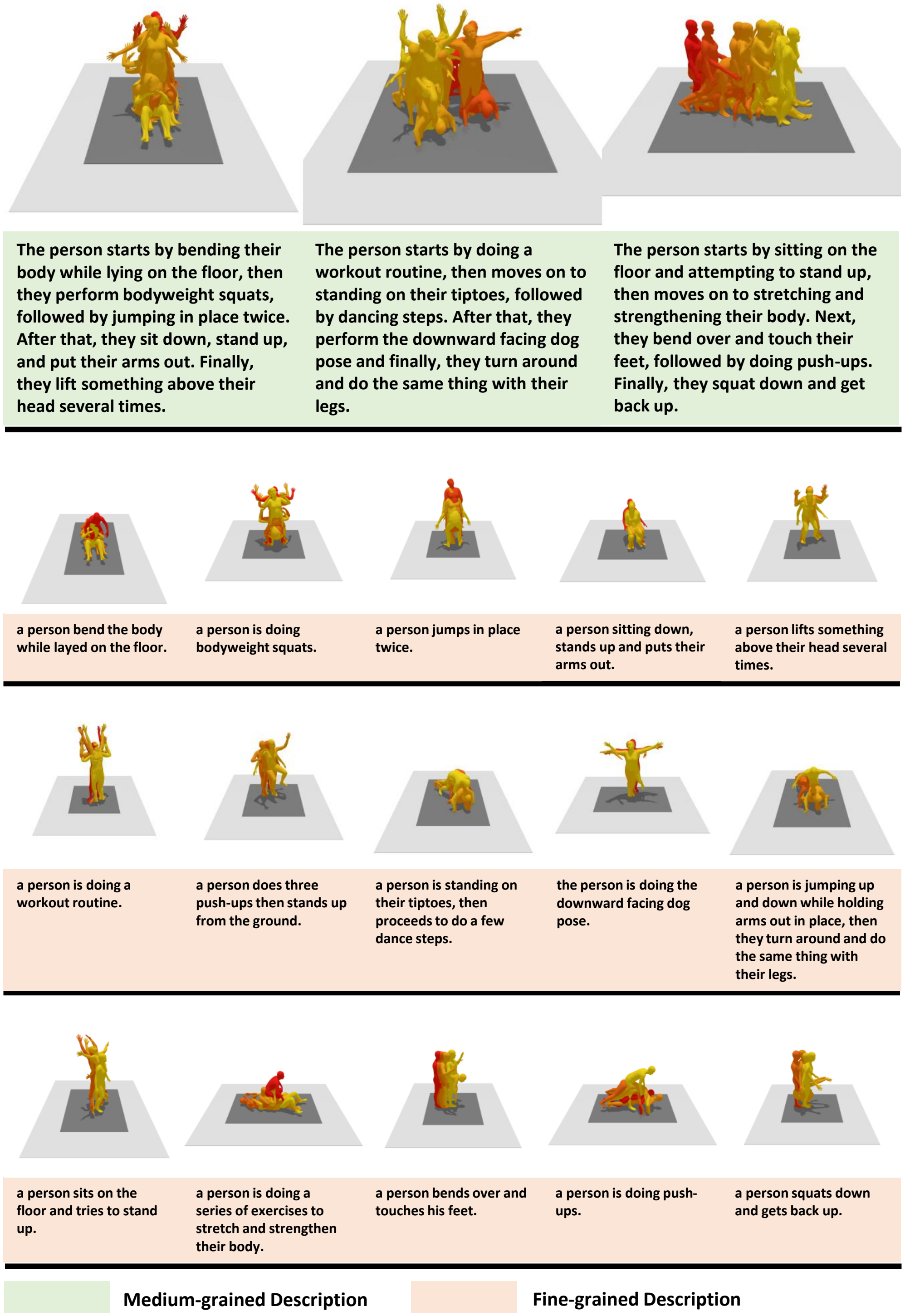
a person walks around the room and touches something with his left hand.

a person walks to the left, walks back, walks past where he started by

a person walks around and then walks back.

a person is preparing for a fight.

Figure 13. **Motion Understanding at Various Text Granularity.** Our method can describe motion at various levels of detail. Given long motions, our method can generate descriptions that depict the presented activity(medium-grained). Given a short motion sequence, our method can describe its specific action in a short sentence(fine-grained).



The person starts by bending their body while lying on the floor, then they perform bodyweight squats, followed by jumping in place twice. After that, they sit down, stand up, and put their arms out. Finally, they lift something above their head several times.

The person starts by doing a workout routine, then moves on to standing on their tiptoes, followed by dancing steps. After that, they perform the downward facing dog pose and finally, they turn around and do the same thing with their legs.

The person starts by sitting on the floor and attempting to stand up, then moves on to stretching and strengthening their body. Next, they bend over and touch their feet, followed by doing push-ups. Finally, they squat down and get back up.

a person bend the body while layed on the floor.

a person is doing bodyweight squats.

a person jumps in place twice.

a person sitting down, stands up and puts their arms out.

a person lifts something above their head several times.

a person is doing a workout routine.

a person does three push-ups then stands up from the ground.

a person is standing on their tiptoes, then proceeds to do a few dance steps.

the person is doing the downward facing dog pose.

a person is jumping up and down while holding arms out in place, then they turn around and do the same thing with their legs.

a person sits on the floor and tries to stand up.

a person is doing a series of exercises to stretch and strengthen their body.

a person bends over and touches his feet.

a person is doing push-ups.

a person squats down and gets back up.

🟩 Medium-grained Description 🟧 Fine-grained Description

Figure 14. **Motion Understanding at Various Text Granularity.** Our method can describe motion at various levels of detail. Given long motions, our method can generate descriptions that depict the presented activity(medium-grained). Given a short motion sequence, our method can describe its specific action in a short sentence(fine-grained).

References

[1] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1

[2] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 1

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 1

[4] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004. 2

[5] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 2

[6] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 1