ActAvatar: Temporally-Aware Precise Action Control for Talking Avatars

Supplementary Material

1. Introduction

This supplementary material provides detailed information to complement the main paper. We organize the content as follows:

- **Section 2** presents the comprehensive dataset construction pipeline.
- **Section 3** describes the Gemini-based evaluation framework design.
- **Section 4** discusses ethical considerations and limitations.

2. Dataset Construction Pipeline

2.1. Stage 1 Training Data

Data Sources and Statistics. For Stage 1 audio-driven lip synchronization training, we utilize 500K diverse talking head videos from two large-scale datasets: OpenHumanVid [1] and SpeakerVid [4].

Data Preprocessing Pipeline. Our preprocessing pipeline consists of several stages to ensure data quality and consistency:

1. **Video Segmentation:** Long videos are segmented into 5-second clips.
2. **Quality Filtering:** We apply multiple quality checks:
   - *Blur Detection:* Videos with HyperIQA [2] below 40 are discarded.
   - *Occlusion Detection:* Clips where facial landmarks are occluded for more than 30% of frames are removed.
   - *Face Detection:* Clips where face detection fails in more than 10% of frames are excluded.

After preprocessing, approximately 30% of the original videos are filtered out due to quality issues, resulting in the final 500K high-quality training samples.

2.2. Stage 2 Structured Annotation Dataset

2.2.1. DWPose-based Motion Selection

To construct a high-quality action control dataset, we need videos with significant body and hand movements. We employ DWPose [3] for motion magnitude calculation and filtering.

Motion Magnitude Calculation. For each video, we extract whole-body pose keypoints using DWPose, which provides 133 keypoints covering body (17), hands (21 per hand), and face (68). We focus on upper body and hand keypoints for motion calculation.

The motion magnitude for a video is computed as:

$$\text{Motion Magnitude} = \sum_{t=1}^{T-1} \sum_{j \in \mathcal{J}} \|p_j^{(t+1)} - p_j^{(t)}\|_2 \quad (1)$$

where T is the number of frames, $\mathcal{J}$ is the set of upper body and hand keypoints (excluding face keypoints to avoid bias from speaking-related movements), and $p_j^{(t)}$ denotes the 2D position of keypoint j at frame t.

To make motion magnitude comparable across videos of different lengths, we normalize by the number of frames:

$$\text{Normalized Motion} = \frac{1}{T-1} \sum_{t=1}^{T-1} \sum_{j \in \mathcal{J}} \|p_j^{(t+1)} - p_j^{(t)}\|_2 \quad (2)$$

Filtering Criteria. We apply the following criteria to select videos with significant actions:

- **Overall Motion Threshold:** Videos with normalized motion magnitude above the 60th percentile of the dataset distribution are retained.
- **Hand Motion Emphasis:** We separately compute hand motion magnitude. Videos where hand motion contributes less than 40% to the total motion are filtered out, as we prioritize expressive hand gestures.
- **Static Frame Filtering:** Videos containing more than 40% static frames (motion $<$1 pixel) are excluded.

2.2.2. MLLM-based Prompt Generation

After motion-based filtering, we use a Multimodal Large Language Model (MLLM) to generate structured prompts with temporal annotations.

MLLM Configuration. We employ Gemini 2.5 Pro, a state-of-the-art multimodal model capable of processing both video and audio inputs. The model takes two inputs for each 5-second clip:

1. **Video Content:** The complete 5-second video clip showing the talking avatar.
2. **Audio Content:** The corresponding audio waveform extracted from the video.

Prompt Engineering Template. We design a specialized system prompt to guide the MLLM in generating structured annotations for 5-second video clips. The prompt emphasizes temporal structure, motion amplitude specification, and audio-visual alignment.

System Prompt for 5-Second Clip Annotation:

```
You are a director analyzing a 5-second
```

```
talking avatar video clip.

You are given the complete 5-second video
(showing a frontal half-body view) and its
corresponding audio.

Write a structured prompt that describes the
video content to guide future generation of
similar clips.

Include a subtle dynamic background cue in
the Scene to avoid static backgrounds.

Use moderate-to-large, clearly visible motion
amplitude with clear direction and approximate
magnitudes aligned to the audio semantics.

Return ONLY the prompt text in this exact
format:

- Base
    Scene: <scene description + communicative
           intent + subtle dynamic background
           cue>
    Emotion: <short emotional state>
    Movement style: emotion: <emotional
                     quality>;
                    characteristics:
                    moderate-to-large amplitude,
                    <movement characteristics>;
                    energy: <energy level>;
                    social: <social context>.

- Phase-1 [0-2s]
    [SEG-A] <concise posture/gesture with
            clear direction + approximate
            magnitude, aligned with audio
            content in this segment>

- Phase-2 [2-4s]
    [SEG-B] <concise posture/gesture with
            clear direction + approximate
            magnitude, aligned with audio
            content in this segment>
```

Key Design Principles. The prompt template incorporates several important design principles:

1. **Hierarchical Structure:** The Base block captures time-invariant scene context, emotion, and movement style, while Phase blocks describe temporally-localized actions. This hierarchy enables the model to maintain global consistency while executing phase-specific actions.
2. **Explicit Temporal Anchors:** Each phase is annotated with precise time windows (e.g., [0-2s], [2-4s]) in a standardized format. This explicit temporal grounding is crucial for the Phase-Aware Cross-Attention mechanism to associate textual descriptions with specific temporal segments.
3. **Motion Amplitude Specification:** The prompt explicitly requests "moderate-to-large, clearly visible motion amplitude" to avoid the common issue of generated actions being too subtle or barely noticeable. This ensures that the annotated actions are sufficiently expressive for training.
4. **Directional and Magnitude Information:** Actions are described with clear direction (e.g., "outward", "downward", "to chest level") and approximate magnitudes (e.g., "to chest level", "45-degree angle"), providing concrete spatial guidance for future generation.
5. **Audio Semantic Alignment:** The prompt emphasizes aligning gestures with "audio content in this segment", encouraging the MLLM to consider the relationship between spoken content and physical actions within each temporal phase.
6. **Dynamic Background Cues:** Including subtle background dynamics (e.g., "gentle breeze moving curtains", "soft natural light shifting") prevents completely static backgrounds, which can appear unnatural in generated videos.
7. **Segment Markers:** The [SEG-A] and [SEG-B] markers serve as explicit temporal anchors that the model can use to associate actions with specific time windows during generation.
8. **Two-Phase Division:** We divide the 5-second clip into two phases ([0-2s], [2-4s]) to balance temporal granularity with annotation complexity. This division captures the natural rhythm of communicative gestures while maintaining manageable annotation overhead. The final second (4-5s) is typically used for transitioning or completing the gesture described in Phase-2.

Example Outputs. Here are representative examples of MLLM-generated prompts using our template:
Example 1 (Professional Presentation):

```
- Base
    Scene: A woman in a navy blazer standing
    in a modern conference room with soft
    natural light filtering through windows,
    speaking professionally about business
    strategy
    Emotion: confident, focused
    Movement style: emotion: professional and
    composed; characteristics: moderate-to-
    large amplitude, deliberate and controlled
    gestures; energy: steady and authoritative;
    social: engaging and persuasive.

- Phase-1 [0-2s]
    [SEG-A] Raises right hand from waist to
    chest level (approximately 40cm upward)
```

```
    with palm open facing forward, gesturing
    outward to emphasize opening statement

- Phase-2 [2-4s]
    [SEG-B] Lowers hand smoothly and extends
    index finger forward at shoulder height
    (approximately 30cm extension), pointing
    to emphasize key point while maintaining
    eye contact
```

Example 2 (Casual Conversation):

```
- Base
    Scene: A young man in casual gray sweater
    sitting in a home office with warm ambient
    lighting and subtle bookshelf background,
    speaking enthusiastically about a personal
    project
    Emotion: excited, friendly
    Movement style: emotion: enthusiastic and
    warm; characteristics: moderate-to-large
    amplitude, expressive and animated gestures;
    energy: high and dynamic; social: engaging
    and personable.

- Phase-1 [0-2s]
    [SEG-A] Waves right hand at head level
    (approximately 50cm lateral arc) in
    greeting motion with open palm, smiling
    broadly while leaning slightly forward

- Phase-2 [2-4s]
    [SEG-B] Brings both hands together at
    chest level (hands moving approximately
    40cm inward from sides), clasping fingers
    while nodding to emphasize agreement
```

Example 3 (Educational Explanation):

```
- Base
    Scene: A middle-aged instructor in smart
    casual attire standing before a whiteboard
    in a well-lit classroom with gentle shadows
    from overhead lighting, explaining a
    concept clearly
    Emotion: patient, instructive
    Movement style: emotion: calm and
    pedagogical; characteristics: moderate
    amplitude, measured and illustrative
    gestures; energy: moderate and consistent;
    social: instructive and supportive.

- Phase-1 [0-2s]
    [SEG-A] Extends left hand to the side at
    shoulder height (approximately 40cm
    lateral extension) with palm facing down,
    establishing first concept point

- Phase-2 [2-4s]
    [SEG-B] Shifts to right hand, raising it
    to head level (approximately 50cm upward
    from rest) with index finger extended,
    indicating second concept point and
    maintaining position through completion
```

3. Evaluation Framework Design

To provide comprehensive and reliable evaluation of action control quality, we develop an automated evaluation framework based on Gemini 2.5 Pro, a state-of-the-art multimodal large language model with strong video understanding capabilities. This section details the complete design of our evaluation system.

3.1. System Prompt Design

We design a comprehensive system prompt that guides Gemini to perform structured evaluation across multiple dimensions. The complete prompt is as follows:

```
You are an expert video quality assessor
specializing in talking avatar evaluation.

You will be provided with a video of a
talking avatar and a structured textual
prompt describing the expected actions.

Your task is to carefully analyze the video
and evaluate whether the avatar performs the
described actions accurately and at the
correct times.

## Input Format
You will receive:
1. A video of a talking avatar (approximately
   4-5 seconds, 125 frames at 25 FPS)
2. A structured prompt in the following
   format:
   - Base Block: Global scene description
     (identity, setting, emotion, style,
     constraints)
   - Phase Blocks: Temporal action
     descriptions with time ranges, e.g.,
     * Phase-1 [0-2s]: Description of action
       in first 2 seconds
     * Phase-2 [2-4s]: Description of action
       in next 2 seconds

## Evaluation Criteria

For each phase block, evaluate the following
aspects:

### 1. Action Occurrence (AO)
- Does the described action appear in the
  video during ANY time window?
- Return: 1 (action present) or 0 (action
  absent)
```

```
### 2. Action Accuracy (AA)
- How well does the executed action match
  the textual description?
- Consider: gesture type, hand shape, finger
  positions, movement direction
- Score: 0-10 (0=completely wrong, 10=perfect
  match)
- Deduct points for: incorrect gesture type,
  unclear hand shape, missing details

### 3. Temporal Correctness (TC)
- Does the action occur within the specified
  time window?
- If action occurs at correct time: 8-10
- If action occurs with slight delay/advance
  (±0.5s): 5-7
- If action occurs at wrong time (>1s offset):
  0-4
- Score: 0-10

### 4. Action Quality (AQ)
- How natural and smooth is the action
  execution?
- Consider: motion fluidity, speed
  appropriateness, naturalness
- Score: 0-10 (0=very unnatural/robotic,
  10=perfectly natural)
- Deduct points for: jerky motion, unnatural
  speed, abrupt transitions

### 5. Hand Clarity (HC)
- How clear and well-defined are the
  hands/fingers?
- Consider: finger separation, anatomical
  plausibility, visual sharpness
- Score: 0-10 (0=completely blurry/distorted,
  10=crystal clear)
- Deduct points for: blurry fingers, merged
  fingers, anatomically impossible poses,
  occlusion artifacts

## Special Considerations

- Temporal Windows: Pay close attention to
  when actions occur. An action that appears
  at the wrong time should receive low TC
  score even if the action itself is correct.

- Hand Gestures: Be strict about hand
  clarity. Common issues include:
  * Fingers blending together (reduce HC)
  * Extra/missing fingers (reduce both AA
    and HC)
  * Unnatural hand angles (reduce AQ and HC)

- Action Semantics: Different gestures have
  different requirements:
  * Counting gestures: Must show correct
    number of fingers clearly separated
  * Pointing: Index finger must be extended,
    others curled
  * Thumbs-up: Thumb must be clearly up,
    other fingers curled
  * Heart sign: Two hands forming heart
    shape with proper symmetry
  * Waving: Hand moving side-to-side with
    appropriate amplitude

- Base Block Compliance: Ensure the overall
  scene matches the base block description
  (identity, setting, emotion), but this
  does not affect the per-phase scores.

## Output Format

Return a JSON object with the following
structure:

{
  "overall_assessment": "Brief 1-2 sentence
    summary of video quality and action
    execution",
  "base_block_compliance": "Brief assessment
    of whether video matches base block
    description",
  "phases": [
    {
      "phase_id": 1,
      "time_range": "[0-2s]",
      "described_action": "Brief summary of
        what was described",
      "action_occurrence": 1,
      "action_accuracy": 8.5,
      "temporal_correctness": 9.0,
      "action_quality": 8.0,
      "hand_clarity": 9.5,
      "detailed_feedback": "Specific
        observations about this phase: what
        was done well, what could be improved,
        timing accuracy, hand quality details"
    },
    {
      "phase_id": 2,
      "time_range": "[2-4s]",
      "described_action": "Brief summary of
        what was described",
      "action_occurrence": 1,
      "action_accuracy": 7.5,
      "temporal_correctness": 8.5,
      "action_quality": 7.0,
      "hand_clarity": 8.0,
      "detailed_feedback": "Specific
        observations about this phase"
    }
  ],
```

```
  "temporal_analysis": "Analysis of temporal
    alignment across all phases, transitions
    between phases, and overall temporal
    coherence",
  "hand_quality_analysis": "Detailed analysis
    of hand/finger clarity throughout the
    video, noting any issues with blur,
    occlusion, or anatomical implausibility",
  "strengths": ["List of 2-3 main strengths"],
  "weaknesses": ["List of 2-3 main weaknesses
    or areas for improvement"]
}
```

3.2. Output Format and Parsing

The evaluation system returns structured JSON output that enables systematic analysis and aggregation of results. We design the output format to capture both quantitative scores and qualitative feedback.

JSON Structure. The output contains the following components:

1. **Overall Assessment:** A concise 1-2 sentence summary providing a holistic view of the video's quality and action execution fidelity.
2. **Base Block Compliance:** An assessment of how well the video matches the global scene description, including identity, setting, emotion, and style. This provides context but does not contribute to quantitative metrics.
3. **Phase-level Evaluations:** For each phase block, the system provides:
   - Phase identifier and time range
   - Summary of the described action
   - Five quantitative scores (AO, AA, TC, AQ, HC)
   - Detailed textual feedback explaining the scores
4. **Temporal Analysis:** A comprehensive analysis of temporal alignment across all phases, including transition quality between phases and overall temporal coherence.
5. **Hand Quality Analysis:** A detailed examination of hand and finger clarity throughout the video, noting specific issues with blur, occlusion, or anatomical implausibility.
6. **Strengths and Weaknesses:** Lists of 2-3 main strengths and areas for improvement, providing actionable insights.

Score Aggregation. For each video, we compute aggregate metrics by averaging phase-level scores:

$$\text{Hit@Segment (H@S)} = \frac{1}{K}\sum_{k=1}^{K} \text{AO}_k \tag{3}$$

$$\text{Action Accuracy (AA)} = \frac{1}{K}\sum_{k=1}^{K} \text{AA}_k \tag{4}$$

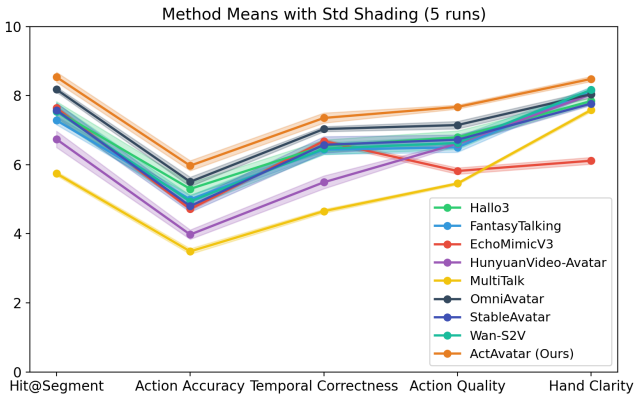

Figure 2. Comprehensive comparison across five Gemini-based action metrics on Action Bench. Shaded regions represent standard deviation across 5 runs. The small variance indicates high robustness of the Gemini-based evaluation framework. Hit@Segment is displayed as a percentage.

$$\text{Temporal Correctness (TC)} = \frac{1}{K}\sum_{k=1}^{K} \text{TC}_k \tag{5}$$

$$\text{Action Quality (AQ)} = \frac{1}{K}\sum_{k=1}^{K} \text{AQ}_k \tag{6}$$

$$\text{Hand Clarity (HC)} = \frac{1}{K}\sum_{k=1}^{K} \text{HC}_k \tag{7}$$

where K is the number of phases in the video. Hit@Segment represents the percentage of phases where the described action actually occurs (AO=1).

3.3. Reliability Validation

We assess the consistency of Gemini's evaluations by running the same prompt and video through the model 5 times independently. For each metric, we compute the mean and standard deviation across the 5 runs. Figure 2 presents the consistency analysis results on Action Bench. The standard deviations are relatively small across all metrics, indicating high consistency.

4. Ethical Considerations and Limitations

4.1. Ethical Considerations

The ability to generate realistic talking avatars presents significant risks including identity impersonation, misinformation propagation, and privacy violations through unauthorized use of likenesses. These concerns are inherent to any realistic video generation technology and require careful consideration by the research community and potential deployers. We acknowledge that comprehensive safeguards such as watermarking systems, detection tools, and usage

authentication mechanisms are necessary but remain areas for future development. The research community should prioritize developing technical solutions including robust watermarking that survives common video manipulations, reliable detection methods to identify synthetic content, and authentication systems to verify ownership of reference images. From a policy perspective, clear guidelines prohibiting malicious uses and legal frameworks establishing accountability are essential.

ActAvatar is intended for beneficial applications including virtual assistants, education, accessibility services, and research. However, the technology could be misused for impersonation and fraud, misinformation creation, privacy violations, harmful content generation, and unauthorized commercial exploitation. We strongly advocate that any deployment should clearly disclose the synthetic nature of content, obtain explicit consent for using individuals' likenesses, and implement monitoring to prevent abuse. Ultimately, responsible use depends on deployers adopting ethical practices and policymakers establishing appropriate regulatory frameworks.

4.2. Limitations

ActAvatar faces several technical limitations that constrain its applicability. Generation quality challenges include blurred high-frequency details such as finger joints and skin texture, particularly during rapid movements. Hair rendering tends to produce smoothed appearances rather than individual strands, and motion blur can exceed natural levels during fast actions. Lighting and shadow handling struggles with complex illumination conditions, often producing softened or inconsistent shadows.

The system performs optimally with single-person frontal or semi-frontal views showing the upper body in controlled indoor settings with moderate-amplitude standard communicative gestures. It is not well-suited for full-body content including walking or dancing, extreme camera angles, complex environments with dynamic backgrounds, or specialized actions requiring fine motor skills.

References

[1] Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7752–7762, 2025. 1

[2] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3667–3676, 2020. 1

[3] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 1

[4] Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. *arXiv preprint arXiv:2507.09862*, 2025. 1