

000 MICROVERSE: A PRELIMINARY EXPLORATION TO- 001 002 WARD A MICRO-WORLD SIMULATION 003 004

005 **Anonymous authors**

006 Paper under double-blind review

007 008 ABSTRACT 009

010
011 Recent advances in video generation have opened new avenues for macroscopic
012 simulation of complex dynamic systems, but their application to microscopic phe-
013 nomena remains largely unexplored. Microscale simulation holds great promise
014 for biomedical applications such as drug discovery, organ-on-chip systems, and
015 disease mechanism studies, while also showing potential in education and inter-
016 active visualization. In this work, we introduce **MicroWorldBench**, a multi-level
017 rubric-based benchmark for microscale simulation tasks. MicroWorldBench en-
018 ables systematic, rubric-based evaluation through 459 unique expert-annotated
019 criteria spanning multiple microscale simulation task (e.g., organ-level processes,
020 cellular dynamics, and subcellular molecular interactions) and evaluation dimen-
021 sions (e.g., scientific fidelity, visual quality, instruction following). MicroWorld-
022 Bench reveals that current SOTA video generation models fail in microscale simu-
023 lation, showing violations of physical laws, temporal inconsistency, and misalign-
024 ment with expert criteria. To address these limitations, we construct **MicroSim-10K**,
025 a high-quality, expert-verified simulation dataset. Leveraging this dataset,
026 we train **MicroVerse**, a video generation model tailored for microscale simula-
027 tion. MicroVerse can accurately reproduce complex microscale **mechanism**. Our
028 work first introduce the concept of **Micro-World Simulation** and present a **proof**
029 of **concept**, paving the way for applications in biology, education, and scientific
030 visualization. Our work demonstrates the potential of educational microscale sim-
031 ulations of biological mechanisms.

032 1 INTRODUCTION

033
034 World models LeCun (2022); Bruce et al. (2024); Lu et al. (2024) have been extensively studied
035 for their ability to simulate environments and agent interactions. They offer a unified computational
036 framework for perceiving surroundings, controlling actions, and predicting outcomes, thereby re-
037 ducing reliance on real-world trials. This not only robotics engines Luo & Du (2024); Lu et al.
038 (2024) engines and reinforcement learning planners Hafner et al. (2020); Agarwal et al. (2025), but
039 also enhances decision-making, supports safe exploration, and enables scalable learning.

040 Recently, video generative models have demonstrated strong potential to acquire commonsense
041 knowledge directly from raw video data, ranging from physical laws in the real world to embodi-
042 ed behavioral patterns Brooks et al. (2024), laying the foundation for their use as real-world sim-
043 ulators. For example, prior work Luo & Du (2024) employs video-guided goal-conditioned explo-
044 ration, grounding large-scale video generation model priors into continuous action spaces through
045 self-supervision, enabling robots to master complex manipulation skills without explicit actions or
046 rewards; and other works Lu et al. (2024) leverage video generation models for embodied decision-
047 making, allowing agents to imaginatively explore their environment with high generative quality
048 and consistent exploration.

049 Despite tremendous progress in video generation for natural scenes and human-centered do-
050 mains OpenAI (2024); Google DeepMind (2025); Kong et al. (2024); Wan et al. (2025); Yang et al.
051 (2024), research efforts have remained predominantly focused on the macroscopic scale. This suc-
052 ceess has not translated effectively to the microscopic scale, where current state-of-the-art models
053 fail to produce physically plausible or biologically meaningful dynamics, as shown in Figure 1.
Microscopic simulation, which tracks the interactions of atoms, molecules, and cells to uncover

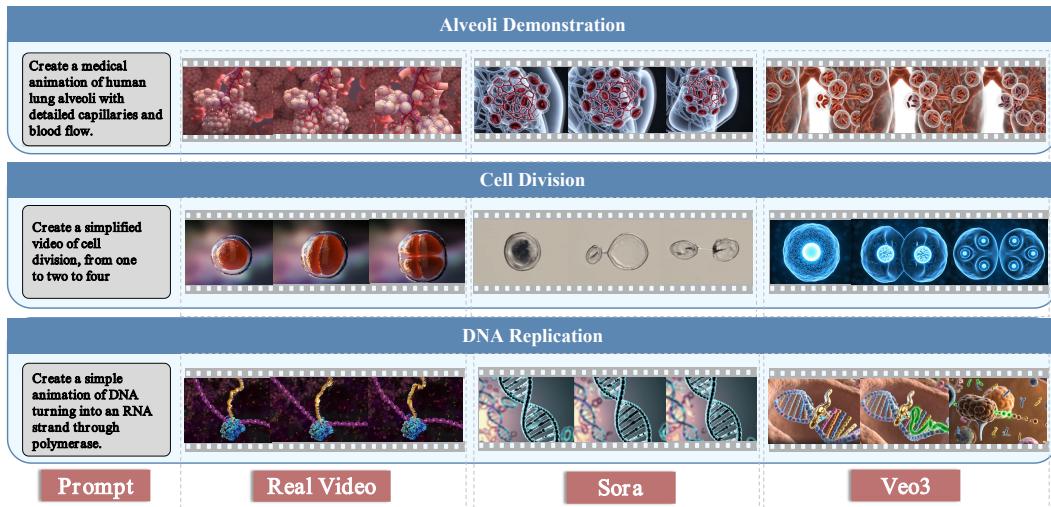


Figure 1: Failure cases of Sora and Veo3 on [Microscale](#) Simulation. Although Sora and Veo3 generate results that appear visually correct, their violations of physical laws are particularly evident.

underlying mechanisms, is crucial for applications in materials science, biomedical research Dario et al. (2000), education Romme (2002), and interactive visualization White (1992). The failure of existing models, primarily due to a lack of incorporated biomedical knowledge, highlights a critical gap despite the strong potential of microscale simulation for generating clinically realistic dynamics in fields like drug discovery and disease modeling. [To address this, we aim to explore the potential of educational microscale simulations of biological mechanisms.](#)

In this work, we introduce **MicroWorldBench**, a multi-level rubric-based benchmark for microscale simulation tasks comprising 459 real-world tasks that span organ-level, cellular, and subcellular processes. These tasks were jointly selected from a large candidate pool by LLMs and domain experts for their diversity and relevance, with each task paired with self-contained, objective evaluation criteria specifying the essentials for valid simulation. Our extensive experiments across a broad spectrum of video generation models reveal that while most maintain superficial visual coherence and adhere to prompts, they perform poorly in microscale settings, consistently failing to generate biologically plausible dynamics. These failures indicate that current models, trained predominantly on human-scale videos, lack grounding in microphysical principles and knowledge.

To mitigate the gap, we introduce **MicroVerse**, a video generation model tailored for microscale simulation. MicroVerse is built on Wan2.1 Wan et al. (2025) model and trained with **MicroSim-10K**, [the first microscale dataset containing 9,601 expert-verified scenarios](#). Unlike human-scale datasets, MicroSim-10K emphasizes physical plausibility and biological fidelity across diverse microscale mechanisms. On MicroWorldBench, MicroVerse surpasses original model by more than +2.7 in scientific fidelity, highlighting the importance of domain-specific data.

Our contributions are summarized as follows: (i) We introduce the concept of **Micro-World Simulation** and present a **proof of concept**, which includes a clear objective, a dedicated benchmark, a training dataset, and a tailored model. (ii) We propose MicroWorldBench, the first rubric-based benchmark specifically designed for evaluating microscale simulation in video generation; (iii) [we construct MicroSim-10K, a large-scale, expert-verified dataset of microscale simulation videos](#); (iv) We introduce MicroVerse, a fine-tuned video generation model built upon MicroSim-10K, achieving competitive performance on MicroWorldBench by reducing violations of scientific constraints and improving temporal and spatial consistency.

2 MICROWORLDBENCH: A RUBRIC-BASED BENCHMARK FOR MICROSCALE SIMULATION

Generic Evaluation Fails to Capture Microscale Simulation Dynamics Existing evaluation methods for video models often rely on generic scoring rules or high-level principles Huang et al. (2024);

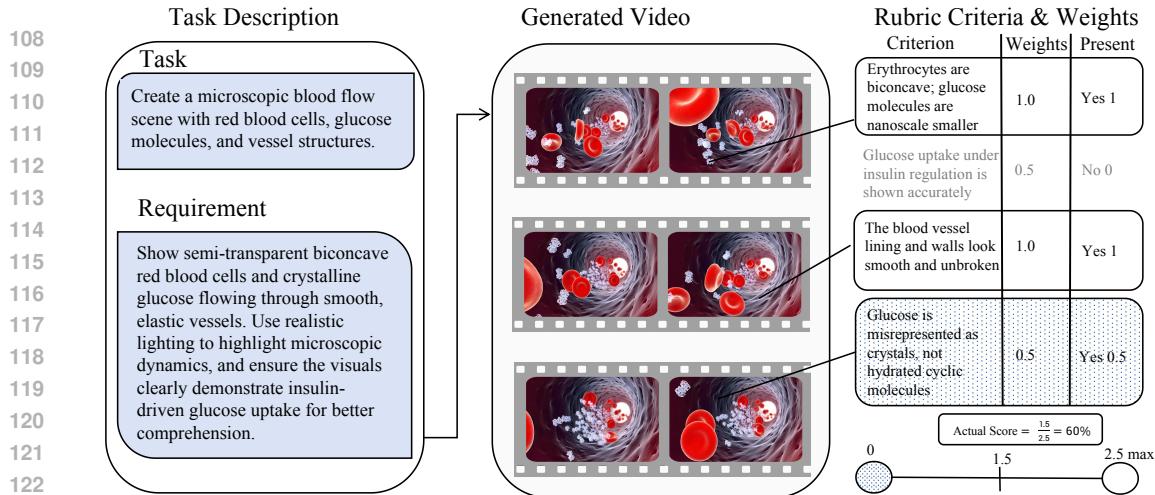


Figure 2: Illustration of MicroWorldBench Evaluation Process.

Zheng et al. (2025); He et al. (2024), which are insufficient for microscale simulation. Such methods overlook the need for fine-grained microscopic simulations, resulting in misaligned outcomes and failing to capture deficiencies in physical plausibility and biological fidelity. In this work, the proposed Rubric evaluation addresses this gap by introducing task-specific criteria with differentiated weights. Rubrics highlight the most critical dimensions identified by experts and ensure that evaluations emphasize substantive shortcomings rather than being diluted by aggregate scoring.

In this section, we introduce the core structure of the rubric-based benchmark, covering task selection (Sec. 2.1), prompt design (Sec. 2.2), and rubric construction (Sec. 2.3), and describe the methodology for model evaluation (Sec. 2.4).

2.1 TASK CHOICE

Biological systems are inherently hierarchical, encompassing levels from society, body, organ, and tissue to cell, organelle, protein, and gene Qu et al. (2011). Given constraints of practicality impact and data availability, in this work we focus on three representative levels as a principled sampling of this hierarchy. Importantly, this choice does not discard existing scientific frameworks, but rather reflects a consensus-based selection of the most representative and tractable scales.

1. **Organ-level simulations** are essential because they connect microscale behaviors with macroscopic physiological functions. Dynamic processes such as cardiac contraction or vascular deformation are directly related to medical diagnosis, surgical planning, and education. A benchmark that evaluates these dynamics provides a direct path toward clinically relevant applications.
2. **Cellular-level simulations** are central to biology and medicine, as cell migration, proliferation, and interaction underpin processes such as tissue growth, wound healing, and immune response. Accurate modeling at this level enables researchers and students to visualize and understand the driving forces of health and disease, creating opportunities for both discovery and pedagogy.
3. **Subcellular-level simulations** present the most fine-grained view, capturing biochemical and biophysical mechanisms that govern life at its foundation—fusion, apoptosis, signaling cascades. Evaluating generative models at this level is particularly important, as these processes are both visually subtle and mechanistically complex, requiring high fidelity and physical plausibility.

2.2 PROMPT SUITE

Both the sampling process of diffusion-based video generation models and the development of expert-driven evaluation rubrics are computationally expensive. To ensure efficiency, we control

162 the number of tasks while maintaining diversity and coverage. The construction follows a two-stage
 163 pipeline: (1) *Collecting tasks related to microscale simulation from YouTube*; and (2) *Expert filtering*
 164 to retain only scientifically meaningful tasks. The final suite contains 459 tasks: 238 at the organ
 165 level, 189 at the cellular level, and 32 at the subcellular level. The proportion of tasks is consistent
 166 with the distribution of levels in the collected videos.

167 **Collecting and Generating Prompts** We retrieved over 8,000 YouTube videos using topic-specific
 168 queries related to organ-level, cellular-level, and subcellular-level simulations. For each video, we
 169 collected metadata including titles and descriptions. This information was then provided to GPT-4o,
 170 which generated tasks describing the microscale **mechanism**. Finally, we generated 8,162 tasks. The
 171 prompts used to instruct GPT-4o refer to Appendix A.

172 **Expert Filtering** We filtered the generated tasks based on two criteria: (1) the diversity of the tasks,
 173 and (2) the practical relevance of the tasks. For diversity, we asked GPT-4o to classify each task into
 174 one of the following categories: Organ-level simulations, Cellular-level simulations, or Subcellular-
 175 level simulations. For practical relevance, we invited three biology experts, and each task had to
 176 receive agreement from at least two of the three experts. A task was retained in MicroWorldBench
 177 only if it satisfied both criteria. Classification prompts are in Appendix B.

179 2.3 RUBRIC CRITERIA

181 As shown in Figure 2, each MicroWorldBench example includes a task instruction and rubric cri-
 182 teria, drafted by LLMs and refined by experts. These criteria evaluate scientific fidelity, visual
 183 quality, and instruction following. **Scientific fidelity emphasizes mechanistic accuracy rather than**
 184 **visual realism**. An LLM-based grader then scores the output, providing a standardized, interpretable
 185 assessment.

186 Due to limited expert availability and efficiency concerns, we adopt a collaborative approach where
 187 LLMs generate initial rubric drafts and experts perform revision and validation. This method not
 188 only improves the efficiency of rubric construction but also ensures broader coverage and more
 189 comprehensive consideration despite the small number of experts.

191 **Stage 1: Rubric Drafts Generation** For each task, GPT-5 generates a set of fine-grained criteria:
 192 $P = (a_i, d_i, s_i, w_i)_{i=1}^N$, where a_i denotes the evaluation dimension, d_i is the description of the i -th
 193 criterion, $s_i \in +1, -1$ is the polarity indicating whether the point contributes (+1) or deducts (-1),
 194 and $w_i \in (0, 1]$ is the weight reflecting its importance (e.g., $w_i = 1.0$ for core scientific require-
 195 ments, $w_i = 0.5$ for key but secondary requirements, and $w_i = 0.2$ for auxiliary or presentational).

196 The score for each task is defined as: $S = \sum_{i=1}^N s_i \cdot w_i$. To ensure comparability across tasks, we
 197 normalize it: $S_{\text{norm}} = \frac{S}{\sum_{i=1}^N w_i^+} \times 100$ where $\sum w_i^+$ is the maximum score from positive criteria,
 198 ensuring a maximum of 100 and preventing minor positives from offsetting severe scientific errors.”

200 **Stage 2: Expert Revision and Validation** Domain experts refine the LLM-generated rubric through
 201 the following actions:

- 203 • Deleting or filtering criteria: Experts refine the criteria by modifying or removing d_i that
 204 are redundant, irrelevant, or scientifically trivial.
- 205 • Adjusting weights: When the weight of certain criteria does not align with the scientific
 206 validity of the task, experts modify the corresponding weight w_i .
- 208 • Supplementing criteria: If the automatically generated criteria fail to cover essential sci-
 209 entific dimensions, experts can introduce new tuples (a_j, d_j, s_j, w_j) .

211 We invited three experts to participate in the revision and validation process. Each expert first in-
 212 dependently reviewed and modified the evaluation criteria, including adjusting weights, removing
 213 redundant items, and supplementing any missing dimensions. All modifications were documented
 214 with clear rationale to ensure transparency. The proposed changes from all experts were then aggre-
 215 gated, and conflicts were resolved through discussion, majority voting. For more analysis on expert
 revision and validation, refer to the Appendix D.

216 Table 1: Performance comparison of different video generation models on MicroWorldBench. Bold
 217 indicates the best performance.
 218

Model	Average	Organ-level	Cellular-level	Subcellular-level
Open-Source Video Generation Models				
HunyuanVideo	23.2	23.1	23.8	19.4
CogVideoX-5B	43.5	39.9	47.0	38.6
Wan2.1-T2V-1.3B	49.4	45.9	51.7	52.4
Wan2.2-TI2V-5B	51.6	46.6	53.9	49.5
Wan2.1-T2V-14B	54.8	55.7	54.4	52.8
Wan2.2-T2V-A14B	53.8	56.3	52.0	53.3
MicroVerse-1.3B (Ours)	50.2	47.6	51.7	53.3
Commercial Video Generation Models				
Sora	50.7	55.9	46.1	55.0
Veo3	77.2	77.5	76.9	78.2

231 Table 2: Performance comparison of different video generation models on MicroWorldBench
 232 (dimension-wise scores). Bold indicates the best performance.
 233

Model	Average	Scientific Fidelity	Visual Quality	Instruction Following
Open-Source Video Generation Models				
HunyuanVideo	23.2	15.6	48.2	23.4
CogVideoX-5B	43.5	37.4	64.1	38.6
Wan2.1-T2V-1.3B	49.4	40.3	71.8	50.1
Wan2.2-TI2V-5B	51.6	40.7	82.7	47.0
Wan2.1-T2V-14B	54.8	42.7	86.0	53.8
Wan2.2-T2V-A14B	53.8	37.8	92.8	55.4
MicroVerse-1.3B (Ours)	50.2	43.0	68.5	49.3
Commercial Video Generation Models				
Sora	50.7	35.3	96.4	37.9
Veo3	77.2	65.7	97.0	77.0

248 249 250 2.4 EVALUATION RESULTS AND ANALYSIS

251 **Settings** We evaluated video generation models on microscopic simulation tasks using MicroWorld-
 252 Bench, including open-source models (e.g., Wan2.1 Wan et al. (2025), HunyuanVideo Kong et al.
 253 (2024)) and commercial models (e.g., Sora OpenAI (2024), Veo3 Google DeepMind (2025)). Infer-
 254 ence was conducted once per model under default settings to ensure fairness and consistent resolu-
 255 tion. Rubric evaluation employed LLM-as-a-Judge Zheng et al. (2023), with GPT-5 serving as the
 256 Judge. The configurations and sampling details in the Appendix F.

257 **Overall Results** As shown in Table 1, the performance of different models varies significantly across
 258 organ-level, cellular-level, and subcellular-level tasks. Although commercial closed-source models,
 259 such as Veo3, substantially outperform open-source models in overall scores, their advantage is
 260 mainly confined to the visual quality dimension rather than scientific fidelity.

261 **Visual Quality vs. Scientific Fidelity** Table 2 shows that nearly all models achieve high scores in
 262 visual quality (80–97), yet their scientific fidelity lags far behind (most open-source models score
 263 only 15–43). This result demonstrates that current models often generate videos that “look right”
 264 but fail to strictly adhere to physical and biological laws.

265 **Performance Differences Across Hierarchical Tasks** Both advanced open-source models (e.g.,
 266 Wan2.2-T2V-A14B) and top commercial models (Sora, Veo3) exhibit lower performance on cellu-
 267 lar and subcellular tasks compared to organ-level simulations. This may be attributed to the higher
 268 requirements for physical and biological consistency in these tasks, as well as the scarcity of mi-
 269 croscale training data that can capture complex dynamics.

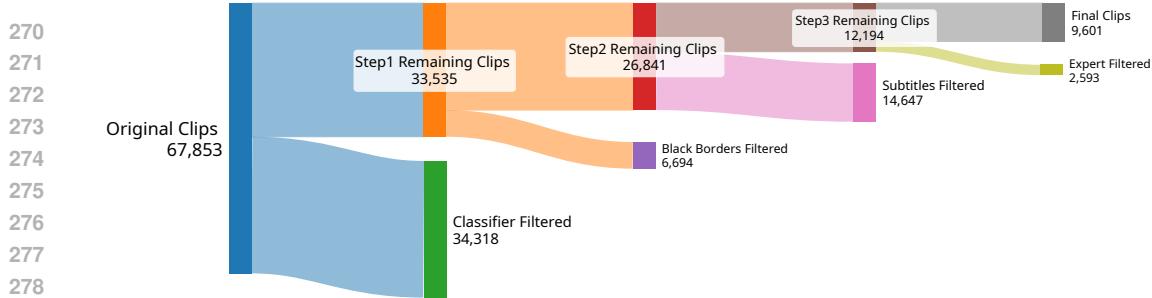


Figure 3: Overview of our data filtering pipeline.

Scale Effects in Open-Source Models Within the Wan series, increasing model size from 1.3B to 14B mainly improves visual quality, while scientific fidelity shows little significant growth. This suggests that expanding model parameters alone is not sufficient to solve the core scientific fidelity challenges in microscale simulation.

3 MICROVERSE: TOWARD MICROSCALE SIMULATION VIA A EXPERT-VERIFIED DATASET

The results of MicroWorldBench indicate that current models remain limited in their ability to model microscale **mechanism** governed by physical and biological principles. Most large-scale video datasets—such as InternVid Wang et al. (2023b), UCF101 Soomro et al. (2012), and OpenVid-1M Nan et al. (2024)—primarily consist of natural scenes or human activities, offering little relevance to microscopic processes. To address this challenge, we propose a new microscale simulation models, termed *MicroVerse*, which explicitly incorporate physical grounding and fine-grained biological dynamics. A key prerequisite for developing such models is the availability of domain-specific data that accurately capture microscopic processes with physical fidelity.

3.1 DATA CONSTRUCTION: MICROSIM-10K

Collecting videos from YouTube We used the official YouTube API to search for videos related to microsimulation and filtered them based on the following criteria: (1) resolution of at least 720p; and (2) licensed under Creative Commons. These requirements ensure that the collected videos are suitable and freely available for training. In total, we obtained 12,848 relevant videos.

Splitting videos After obtaining the videos, we segmented them into multiple semantically consistent and short clips. We used OpenCLIP Ilharco et al. (2021) for video segmentation: whenever the similarity between adjacent frames fell below 0.85, a split was made. In total, 67,853 clips were generated. Since not all clips were related to microsimulation, we trained a classifier based on VideoMAE Tong et al. (2022) to filter them. The model achieved an accuracy of over 92%, significantly improving the quality of the dataset. With the help of the classifier, 34,318 clips were filtered out. For details of the clip classification model related to microsimulation, refer to the Appendix C.

Automatic and expert filtering To improve the quality and physical consistency of the clips, we first applied OpenCV ¹ to detect black borders and used EasyOCR ² to detect subtitles in order to filter out those affecting semantic representation, retaining 12,194 clips. Experts then reviewed the data, removing meaningless or physically inconsistent clips, resulting in 9,601 clips.

Generating captions We leverage a multimodal LLM (GPT-4o) to generate detailed captions. Due to context limits, we uniformly sampled 8 frames per clip as visual input. To minimize hallucinations, we supply the video title and description.

¹<https://github.com/opencv/opencv-python>

²<https://github.com/JaideedAI/EasyOCR>

324 **Prompt MLLM to Generate Video Caption**

325

326 The provided images are sampled from a video clip (8 evenly spaced frames). This clip is taken from
327 a video with the following metadata:

328 Video Title: [{Video Title}](#); Video Description: [{Video Description}](#)

329 Using the visual content of the clip, together with the title and description, please generate a clear,
330 detailed, and accurate description of what is shown. Focus on the subject, explains the scene and
331 actions, and emphasizes visible details, textures, and fine structures.

332 **3.2 DATA STATISTICS**

333

334 **3.2.1 FUNDAMENTAL ATTRIBUTES**

335

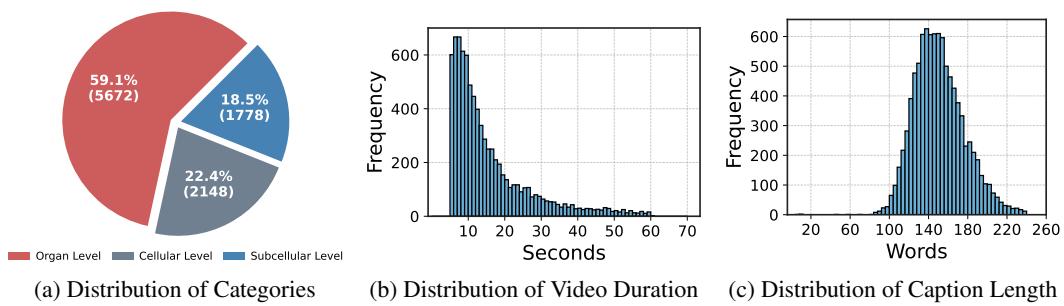


Figure 4: Distributions of fundamental video attributes in the MicroSim-10K.

348 MicroSim-10K is the first large-scale dataset dedicated to microscale simulation, comprising 9,601
349 high-quality video clips. As shown in Figure 4, all clips have a resolution of at least 720p and a
350 duration of 5–60 seconds, ensuring that each captures a complete and coherent microscopic process.
351 The dataset spans diverse biological [mechanisms](#) across organ, cellular, and subcellular levels, of-
352 fering broad coverage of key scenarios. Each clip is paired with a detailed caption generated by a
353 multimodal LLM and validated by experts, with an average length of around 150 words, providing
354 precise semantic alignment for model training.

355 **3.2.2 POPULARITY AND RELEVANCE**

356

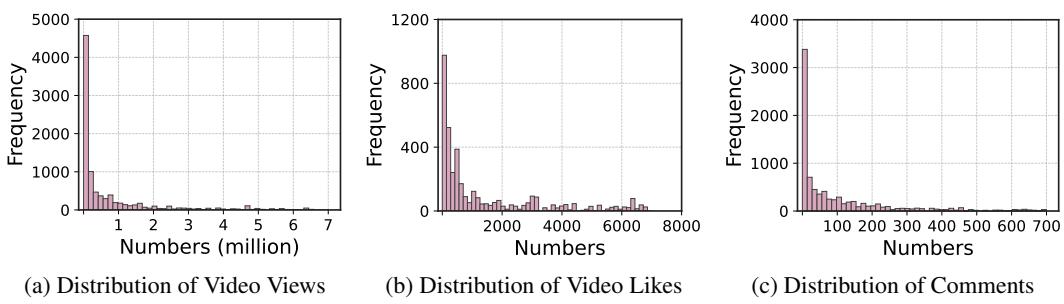


Figure 5: Distributions of video popularity indicators in the MicroSim-10K.

367 To capture the educational and communicative value of microscale simulations, MicroSim-10K re-
368 tains metadata such as views, likes, and comments. As shown in Figure 5, the videos in MicroSim-
369 10K have been widely viewed, with many reaching hundreds of thousands of views, and they have
370 received substantial likes and comments, reflecting strong popularity and broad accessibility across
371 both scientific and public communities.

372 **3.3 TRAINING MICROVERSE**

373

374 For training, we fine-tune the Wan2.1 model. A text prompt P is encoded as a sequence: $P =$
375 (p_0, p_1, \dots, p_m) , while the target video V is decomposed into T frames. Each frame is mapped

378 into the latent space via a VAE Kingma & Welling (2013) encoder, yielding the sequence: $L =$
 379 (l_0, l_1, \dots, l_T) . The text input P is transformed into embeddings E using CLIP text encoder, and
 380 the latent sequence L is processed by a Diffusion Transformer (DiT) Peebles & Xie (2023).
 381

382 The training objective is to predict the latent representation of the video through a denoising diffu-
 383 sion process. At timestep t , the loss function is defined as:

$$\mathcal{L} = \mathbb{E} \left[\|\varepsilon - \varepsilon_\theta(L_t, t, E)\|^2 \right], \quad (1)$$

386 where L_t is the noisy latent representation at timestep t , ε denotes the injected noise, ε_θ is the
 387 model’s noise prediction, t is the current diffusion timestep, and E is the text embedding.
 388

389 During fine-tuning, with probability defined by the 10%, the text conditioning is entirely masked,
 390 enabling Classifier-Free Guidance (CFG) Ho & Salimans (2022) training. This mixture of uncondi-
 391 tional and conditional training improves the generation quality of the model during inference.
 392

393 4 EXPERIMENTS

395 **Experiment Settings** We train MicroVerse using 8 NVIDIA H200 GPUs, fully fine-tuning all pa-
 396 rameters of Wan2.1-T2V-1.3B Wan et al. (2025) with a learning rate of 1e-5 and a batch size of
 397 8. The training process is designed to improve the model’s capability to generate microscopic sim-
 398 ulation videos conditioned on text prompts. We conducted a comparative with other models on
 399 MicroWorldBench. Additional training details are provided in the Appendix E.

400 **Human Evaluation** To evaluate alignment with human preferences, we conducted a human study
 401 comparing MicroVerse with Sora and Veo3. The evaluation included 60 samples across three levels
 402 of microsimulation (20 samples per level), all sourced from the 20 most popular microsimulation
 403 videos on YouTube. Model outputs were randomly shuffled, and three evaluators independently
 404 selected the preferred result based on instruction fidelity and visual clarity, or marked a tie. The final
 405 results were reported as preference ratios.
 406

407 4.1 RESULTS OF OUR MICROVERSE

409 **Improvement in Scientific Fidelity** Table 2 shows that MicroVerse achieves a significant improve-
 410 ment in Scientific Fidelity, reaching a score of 43.0 and outperforming all open-source models. This
 411 enhancement is attributed to the training on the physics-grounded MicroSim-10K dataset, which en-
 412 ables the model to better adhere to biological and physical laws. Although there is a slight decrease
 413 in Visual Quality (68.5) and Instruction Following (49.3), this does not affect our core objective:
 414 advancing scientific fidelity.

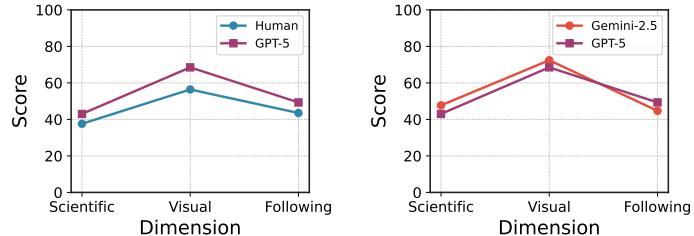
415 **Breakthrough in Subcellular-Level Tasks** According to Table 1, on the highly challenging
 416 subcellular-level tasks, MicroVerse achieves a score of 53.3, surpassing all open-source models.
 417 This demonstrates that our dataset enables MicroVerse to make notable progress on microscale sim-
 418 ulation tasks where existing models typically struggle.
 419

420 4.2 ANALYSIS

422 **Human Evaluation Results** Figure 6 shows the results of human evaluation. Compared with
 423 Wan2.1-1.3B models, MicroVerse performs excellently in the dimension of Scientific Fidelity. Its
 424 outstanding performance in Scientific Fidelity further validates the effectiveness of MicroSim-10K.
 425 In addition, the Cohen’s Kappa coefficient among the three independent experts was above 0.80,
 426 indicating strong interrater agreement and confirming the reliability of the scoring process. More
 427 details on Cohen’s Kappa coefficient can be found in Appendix H.

428 **Consistency among Judgers in MicroWorldBench** To ensure that MicroWorldBench’s evaluation
 429 aligns closely with human judgment across all dimensions, we conducted human preference labeling
 430 on a large set of generated videos. Specifically, we computed the consistency of evaluation tasks
 431 across different models as well as between the models and humans. Figure 6 shows the consistency
 relationships among different models and between the models and humans.

	Wins	Tie	Loses
Scientific Fidelity	47	5	8
Visual Quality	27	11	22
Instruction Following	29	7	24



(a) Human Evaluation Result of MicroVerse and Wan2.1-1.3B (b) Consistency between LLMs and Humans (c) Consistency across different LLMs

Figure 6: Human Evaluation and Consistency Results.

5 RELATED WORK

World Model World models LeCun (2022); Bruce et al. (2024); Lu et al. (2024) have garnered significant attention. They simulate dynamic environments by predicting future states and estimating rewards based on current observations and actions. Their ability to model state transitions has been extended to real-world scenarios through joint learning of policies and world models, improving sample efficiency in simulated robotics Seo et al. (2023), real-world robots Wu et al. (2022), clinical decision Yang et al. (2025), and autonomous driving Wang et al. (2023a). For example, some work Du et al. (2023) explores long-horizon video planning by combining vision–language and text-to-video models. Others Luo & Du (2024) focus on linking video models to continuous actions through goal-conditioned exploration. Recent works Lu et al. (2024) also use video generative models to let agents explore environments more effectively. MeWM Yang et al. (2025) applies world modeling to medical image analysis and clinical decision-making.

Video Generation Video generation has seen rapid progress in the past two years. The release of Sora OpenAI (2024) has ignited strong research interest in text-to-video generation, leading to breakthroughs in quality, coherence, and controllability Blattmann et al. (2023). Other commercial systems such as Veo3, Kling, HunyuanVideo Kong et al. (2024), and Hailuo HailuoAI (2024) have achieved impressive performance and are widely applied in video production, advertising, and education. With the technology maturing, domain-specific models are emerging to address specialized needs. For instance, MedGen Wang et al. (2025) generates accurate, high-quality medical videos for health education, while AniSora Jiang et al. (2025) focuses on producing detailed and stylistically rich animated content. Despite these advances, the use of video generation for microscale simulation remains largely unexplored.

Rubric Evaluation Rubric-based evaluation has become a standard approach for assessing LLMs on open-ended tasks, offering task-specific and interpretable criteria that improve grading consistency. HealthBench Arora et al. (2025) scales this paradigm to 5,000 multi-turn conversations with 48k clinician-authored rubrics covering accuracy, safety, and communication. Building on this, Baichuan-M2 Team (2025) dynamically generates case-specific rubrics as verifiable reward signals for reinforcement learning, enabling adaptive and context-aware supervision. Rubrics as Rewards (RaR) Gunjal et al. (2025) further formalizes rubric-based RL and shows significant gains over Likert-style scoring. These efforts highlight rubric-guided evaluation and training as a promising methodology for developing reliable, aligned, and LLMs.

6 CONCLUSIONS

Video generation excel at natural and human-centered macroscopic scenes but fail to capture faithful microscale dynamics. This work introduces MicroWorldBench, the first rubric-based benchmark for microscale video generation with 459 expert-curated tasks and well-defined rubric criteria. In addition, we build MicroSim-10K and develop MicroVerse which demonstrate remarkable performance on microscale simulation tasks. By integrating physical constraints and expert supervision, MicroVerse not only improves visual fidelity but also advances toward biologically meaningful dynamics, enabling applications in biomedical research, education, and interactive scientific visualization.

486
487

LIMITATION

488
489
490
491
492
493
494

Our work aims to explore the potential of educational microscale simulations of biological mechanisms, rather than the reproduction of results observed in wet lab experiments. However, our current approach does not explicitly incorporate the underlying physical laws that govern biomedical microscale dynamics, such as fluid mechanics in blood flow, diffusion–reaction equations in molecular transport, or biomechanical constraints in cellular processes. Consequently, this limitation restricts the applicability of the model in scenarios that require high-precision scientific simulation and prediction.

495

496
497

ETHICS STATEMENT

498
499
500
501
502

All data are publicly available, compliant with YouTube’s terms, and we exclude personal/sensitive content. Captions were auto-generated (MLLMs) and manually verified to remove inappropriate/identifiable material. The dataset is intended solely and strictly for research purposes and should not be used for non-research settings. We do not own the copyright of these data and will only publicly release the URLs linked to the data instead of the raw data.

503

504
505

REPRODUCIBILITY STATEMENT

506
507
508
509
510
511
512
513

We have made every effort to ensure the reproducibility of our results. The proposed benchmark, along with the training and evaluation code, has been made publicly available in an anonymous repository to facilitate verification and replication. The paper provides a detailed description of the experimental setup, including evaluation procedures, training steps, and hardware information, as well as a complete specification of the benchmark to help researchers accurately understand and reproduce our experiments. We believe these measures will further advance research in this field. Please visit our anonymous Github: <https://anonymous.4open.science/r/rsrsyzyz/>

514

515
516

LARGE LANGUAGE MODELS USAGE STATEMENT

517
518
519
520
521

In this work, large language models (LLMs) served as a writing aid to polish the manuscript’s language. We utilized ChatGPT specifically to refine sentence structure and improve grammatical correctness. Importantly, the LLM played no role in generating the research itself; the intellectual contributions, experimental design, data interpretation, and findings are entirely the work of the authors.

522

523
524

REFERENCES

525
526
527

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chat-topadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

528
529
530

Rahul K. Arora, Jason Wei, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

531
532

Andreas Blattmann, Sergey Frolov, Robin Rombach, and Patrick Esser. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

533
534
535
536
537

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. <https://openai.com/research/video-generation-models-as-world-simulators>, 2024. OpenAI Research.

538
539

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

540 Paolo Dario, Maria Chiara Carrozza, Antonella Benvenuto, and Arianna Menciassi. Micro-systems
 541 in biomedical applications. *Journal of Micromechanics and Microengineering*, 10(2):235, 2000.
 542

543 Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet,
 544 Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Kaelbling, Andy Zeng, and Jonathan
 545 Tompson. Video language planning, 2023. URL <https://arxiv.org/abs/2310.10625>.

546 Google DeepMind. Veo 3, 2025. URL <https://deepmind.google/models/veo/>.

547

548 Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Bing Liu, and Sean Hendryx. Rubrics as
 549 rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*,
 550 2025.

551 Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning
 552 behaviors by latent imagination, 2020. URL <https://arxiv.org/abs/1912.01603>.

553 HailuoAI. Hailuoai, 2024. URL <https://hailuoai.video/>.

554

555 Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil
 556 Chandra, Ziyang Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate
 557 fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024.

558 Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint*
 559 *arXiv:2207.12598*, 2022.

560

561 Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianx-
 562 ing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for
 563 video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*
 564 *Pattern Recognition*, pp. 21807–21818, 2024.

565 Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori,
 566 Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali
 567 Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL <https://doi.org/10.5281/zenodo.5143773>. If you use this software, please cite it as below.

568

569 Yudong Jiang, Baohan Xu, Siqian Yang, Mingyu Yin, Jing Liu, Chao Xu, Siqi Wang, Yidi Wu,
 570 Bingwen Zhu, Xinwen Zhang, Xingyu Zheng, Jixuan Xu, Yue Zhang, Jinlong Hou, and Huyang
 571 Sun. Anisora: Exploring the frontiers of animation video generation in the sora era, 2025. URL
 572 <https://arxiv.org/abs/2412.10255>.

573

574 Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint*
 575 *arXiv:1312.6114*, 2013.

576

577 Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li,
 578 Bo Wu, Jianwei Zhang, et al. Hunyuanyvideo: A systematic framework for large video generative
 579 models. *arXiv preprint arXiv:2412.03603*, 2024.

580

581 Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open*
 582 *Review*, 62(1):1–62, 2022.

583

584 Taiming Lu, Tianmin Shu, Alan Yuille, Daniel Khashabi, and Jieneng Chen. Generative world
 585 explorer. *arXiv preprint arXiv:2411.11844*, 2024.

586

587 Yunhao Luo and Yilun Du. Grounding video models to actions through goal conditioned exploration.
 588 *arXiv preprint arXiv:2411.07223*, 2024.

589

590 Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang,
 591 and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv*
 592 *preprint arXiv:2407.02371*, 2024.

593

594 OpenAI. Video generation models as world simulators, February 2024. URL <https://openai.com/index/video-generation-models-as-world-simulators/>.

595

596 William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of*
 597 *the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

594 Zhilin Qu, Alan Garfinkel, James N Weiss, and Melissa Nivala. Multi-scale modeling in biology:
 595 how to bridge the gaps between scales? *Progress in biophysics and molecular biology*, 107(1):
 596 21–31, 2011.

597 Georges Romme. The educational value of microworld simulation. *Tilburg University, Netherlands*,
 598 2002.

600 Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter
 601 Abbeel. Masked world models for visual control, 2023. URL <https://arxiv.org/abs/2206.14244>.

603 Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions
 604 classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

606 Baichuan-M2 Team. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv
 607 preprint arXiv:2509.02208*, 2025.

609 Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-
 610 efficient learners for self-supervised video pre-training. *Advances in neural information process-
 611 ing systems*, 35:10078–10093, 2022.

612 Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu,
 613 Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative
 614 models. *arXiv preprint arXiv:2503.20314*, 2025.

616 Rongsheng Wang, Junying Chen, Ke Ji, Zhenyang Cai, Shunian Chen, Yunjin Yang, and Benyou
 617 Wang. Medgen: Unlocking medical video generation by scaling granularly-annotated medical
 618 videos, 2025. URL <https://arxiv.org/abs/2507.05675>.

619 Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drive-
 620 dreamer: Towards real-world-driven world models for autonomous driving, 2023a. URL <https://arxiv.org/abs/2309.09777>.

622 Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan
 623 Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal under-
 624 standing and generation. *arXiv preprint arXiv:2307.06942*, 2023b.

626 Barbara Y White. A microworld-based approach to science education. In *New directions in educa-
 627 tional technology*, pp. 227–242. Springer, 1992.

628 Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and Pieter Abbeel. Daydreamer:
 629 World models for physical robot learning, 2022. URL <https://arxiv.org/abs/2206.14176>.

631 Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei
 632 Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, et al. Medical world model: Generative simulation
 633 of tumor evolution for treatment planning. *arXiv preprint arXiv:2506.02327*, 2025.

635 Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang,
 636 Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models
 637 with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

638 Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen
 639 He, Wei-Shi Zheng, Yu Qiao, et al. Vbench-2.0: Advancing video generation benchmark suite
 640 for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

642 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
 643 Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica.
 644 Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL <https://arxiv.org/abs/2306.05685>.

648 **A GENERATE PROMPT FROM THE VIDEO TITLE AND DESCRIPTION**
649650 **Generate Prompt from the Video Title and Description.**

651 Your task: Based on the “video title” and “video description” I provide, craft an extremely detailed,
652 professional, and keyword-rich English prompt specifically for generating breathtaking microscopic-
653 world videos. Do not give any explanation, output directly. Don’t use bullet points. Write it as a single,
654 complete paragraph.

655 Generate a complete, ready-to-use video-generation prompt. This prompt must include all of the
656 following sections:

657

- 658 1. Main Subject: Clearly describe the central object within the microscopic scene.
- 659 2. Scene & Action: Describe what is happening.
- 660 3. Details & Textures: Emphasize the details that should be visible.

661 The video title is:

662 {video_title}

663 The video description is:

664 {video_description}

665 **B PROMPT LLM CLASSIFIES BASED ON TASK DESCRIPTIONS**
666667 **Prompt LLM Classifies Based on Task Descriptions.**

668 Your task: You are an expert in scientific video classification. Given a task description, classify it into
669 one of the following categories for diversity:

670

- 671 1. Organ-level simulations – tasks focusing on the behavior, dynamics, or interactions at the
672 scale of whole organs or organ systems.
- 673 2. Cellular-level simulations – tasks focusing on the behaviors and interactions of single cells
674 or collections of cells, such as cell division, cell fusion, cell migration, or cell signaling.
- 675 3. Subcellular-level simulations – tasks focusing on molecular, genetic, or biochemical pro-
676 cesses within cells, such as protein folding, gene regulation, or intracellular signaling.

677 The task description is:

678 {Task_Description}

679 Please output only the most appropriate category label based on the task description provided.

680 **C VIDEOMAE-BASED MICROSIMULATION CLASSIFIER**
681

682

683 To filter out video clips related to microsimulation, we trained a classifier using 2,580 manually
684 annotated samples based on the VideoMAE model. The training was implemented within the Trans-
685 formers³, with a learning rate of 5e-5 and a total of 10 epochs, enabling the model to effectively
686 capture video features and achieve accurate classification. Finally, our classifier achieved an accu-
687 racy of 92% on the test set.

688

689

690

691

692

693

694

695

696 Table 3: Dataset statistics for microsimulation classification.

697

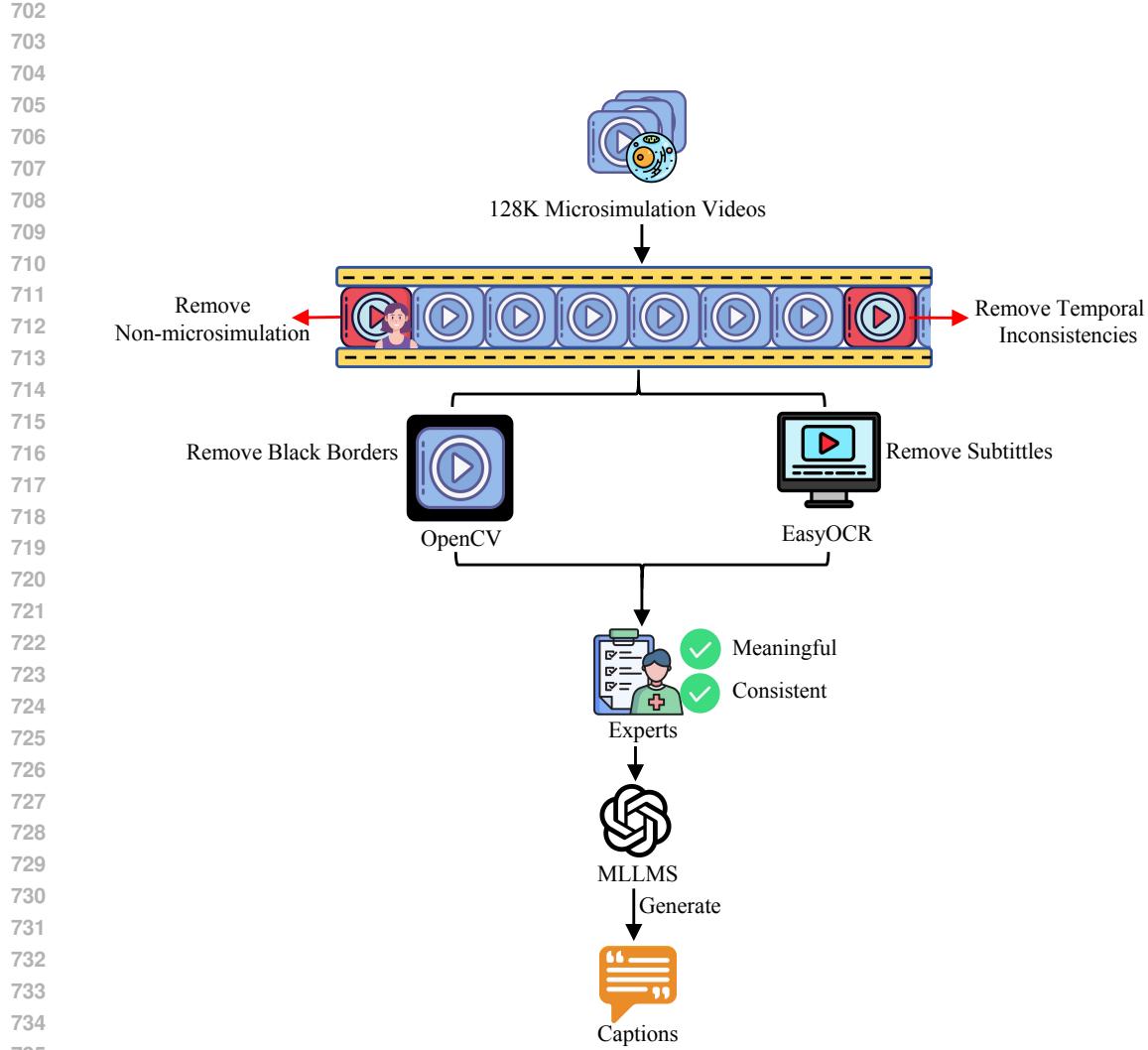
698

699

700

701

Category	Total	Train (80%)	Test (20%)
Microsimulation-related	1107	885	222
Non-microsimulation	1473	1178	295
Total	2580	2063	517



736 Figure 7: Overview of our data filtering pipeline. Each stage applies specific filters and shows the
737 volume of data removed and retained.

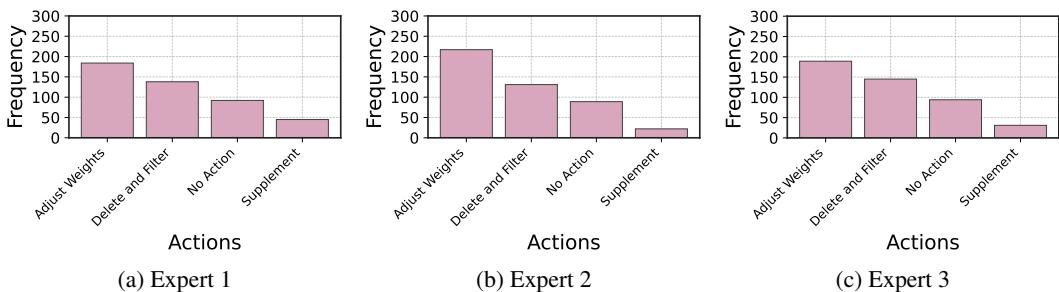


Figure 8: Actions Log of Three Independent Experts.

756 D ANALYSIS OF THE PROCESS OF EXPERT REVISION AND VALIDATION
757
758

759 We analyzed the frequency with which three experts employed the four types of rubric operations
760 when handling different tasks. As shown in Figure 8, all experts tended to favor Adjust Weights,
761 while Supplement was used relatively infrequently. Follow-up interviews with the three experts
762 revealed that the Supplement operation is more cumbersome, as it requires identifying additional
763 evaluation criteria beyond those automatically generated by the LLM, which can introduce extra
764 burden.

765
766 E TRAINING SETTINGS ON MICROVERSE
767
768

770 We train MicroVerse using 8 NVIDIA H200 GPUs, fully fine-tuning all parameters of Wan2.1-T2V-
771 1.3B. Table 4 shows the detailed training parameter settings used to train MicroVerse.
772
773

774 Table 4: Training parameter settings.
775

776 Parameter	777	778 Value
778 --train_batch_size	779	780 8
779 --max_train_steps	780	781 5000
780 --learning_rate	781	782 1e-5
781 --mixed_precision	782	783 bf16
782 --training_cfg_rate	783	784 0.1
783 --num_height	784	785 480
784 --num_width	785	786 832
785 --num_frames	786	787 81
786 --weight_decay	787	788 0.01
787 --dit_precision		789 fp32
		790 full

791 F INFERENCE SETTINGS ON MICROVERSE
792
793794 Table 5: Inference parameter settings.
795
796

797 Parameter	798	799 Value
800 --height	801	802 480
801 --width	802	803 832
802 --num_frames	803	804 81
803 --guidance_scale	804	805 5.0
804 --num_inference_steps		806 50

807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2379
2380
2381
2382
2383
2384
2385
2386
238

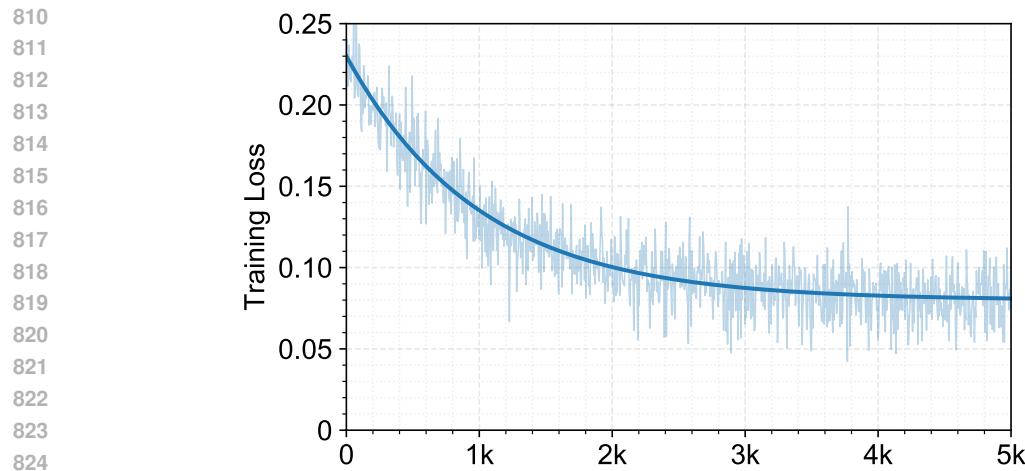


Figure 9: Training loss decreases steadily over 5k iterations.

864 **G PROMPT GPT-5 TO GENERATE RUBRIC CRITERIA**
865866 **Prompt GPT-5 to Generate Rubric Criteria**
867868 **Task:**869 You are a biology expert. Your task is to **design a set of rubrics** to evaluate the completion of a given
870 task based on the provided **Prompt**.871 The rubric should consist of multiple **triplets** in the form:

872
$$\{a_i, d_i, s_i, w_i\}$$

- 873 • a_i : the evaluation aspect, restricted to one of the following three categories:
 - 874 – Scientific Fidelity: Accurate representation of organs, cells, and subcellular structures
875 in scale, morphology, and spatial relationships, with dynamic processes consistent with
876 biological and physical laws.
 - 877 – Visual Quality: Emphasis on clarity, detail, and aesthetics, including model precision,
878 rendering, lighting, and color balance.
 - 879 – Instruction Following: Generated videos strictly follow the prompt description.
- 880 • d_i : description of the i -th evaluation criterion.
- 881 • s_i : polarity of the criterion, either +1 (contributes positively) or -1 (deducts points); leave
882 this field empty.
- 883 • w_i : weight of importance in the range (0, 1]:
 - 884 – 1.0 → core scientific requirements
 - 885 – 0.5 → important but secondary requirements
 - 886 – 0.2 → auxiliary or presentational requirements

888 **Output example:**

```

889 {
890   "a1": "Scientific Fidelity",
891   "d1": "Key cell structures are clearly defined and proportionally
892   accurate",
893   "s1": "+1",
894   "w1": "1.0"
895 }
```

896 **Constraint:**

- 897 1. Do not give any explanation, output directly.
- 898 2. Please describe the evaluation criterion (d_i) in as much detail as possible.
- 899 3. Directly describe the key rubrics in the evaluation criterion, and do not use words such as
'whether'.
- 900 4. Only English output is allowed.

901 **Given prompt:**

```
{prompt}
```

904 **H INTER-RATER RELIABILITY AMONG HUMAN EVALUATORS**
905906 We used Cohen's Kappa coefficient to measure agreement among the three experts. Table 6 indicate
907 strong agreement among the experts, confirming the reliability of the scoring process.911 Table 6: Comparison of Cohen's Kappa values between experts.
912

913 Comparison Pair	914	915 Cohen's Kappa
915 Expert1 vs. Expert2		0.87
916 Expert1 vs. Expert3		0.83
917 Expert2 vs. Expert3		0.81

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971

I A RUBRIC EXAMPLE FROM MICROWORLDBENCH

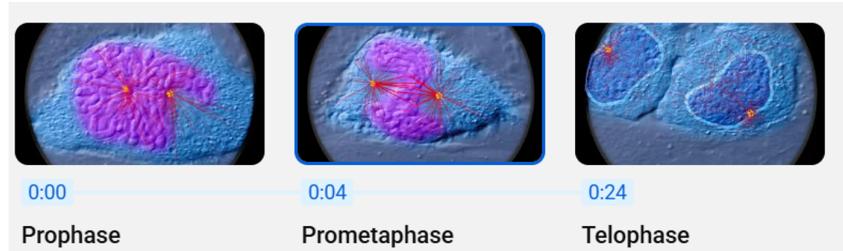
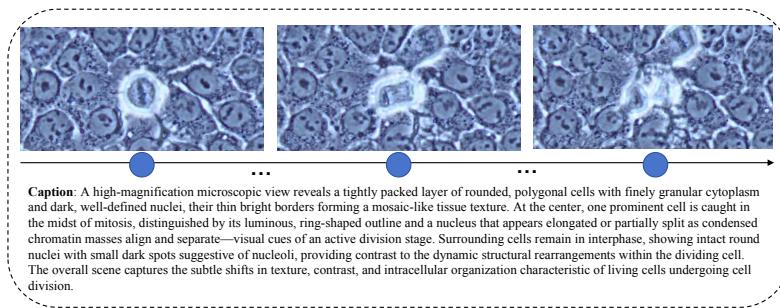


Figure 10: Example of a cell mitosis simulation video frame, taken from an excellent example video on YouTube.

Table 7: Rubric Example for Cell Mitosis Evaluation

Dimension	Criteria
972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025	<p>Scientific Fidelity</p> <p>Weight = +1.0 The sequence of stages, segregation patterns, and ploidy changes in mitosis and meiosis are accurately represented; mitosis produces two genetically identical diploid daughter cells, whereas meiosis involves two successive divisions resulting in four genetically diverse haploid gametes.</p> <p>Weight = +1.0 The alignment of chromosomes at the metaphase plate and their segregation from metaphase to anaphase occur correctly; sister chromatids are distinctly differentiated from homologous chromosomes; the attachment of kinetochores to spindle microtubules and the direction of their tension conform to biological principles.</p> <p>Weight = +1.0 Structural details and dynamic coordination between the spindle apparatus and the centrosome (centriole) are accurate; spindle pole positioning, microtubule polarity, and force distribution are appropriate; the relationship between the microtubule-organizing center and cell polarity is correctly established.</p> <p>Weight = +1.0 The timing and mechanisms of DNA replication and genetic recombination are correctly presented; DNA replication occurs during the pre-mitotic S phase, homologous pairing and crossing over take place in prophase I of meiosis, and no DNA replication occurs between meiosis I and II.</p> <p>Visual Quality</p> <p>Weight = +0.5 The image demonstrates high clarity and fine presentation of microstructural details, with sharp edges of subcellular structures, well-defined layer separation, and absence of wax artifact noise.</p> <p>Weight = +0.5 The animation exhibits coherent motion with stable temporal rhythm, smooth phase transitions, and natural movement trajectories, without any stuttering or tearing.</p> <p>Weight = +0.2 The 3D modeling and material texture are credible, with consistent form proportions and scale hierarchy; the textures are detailed, and surface microstructures are discernible.</p> <p>Weight = +0.2 Coordination of lighting, shadows, and depth of field; controlled volumetric scattering and highlights without excess; clear subject contours with well-defined micro-scale detailing.</p> <p>Instruction Following</p> <p>Weight = +0.5 Accurately present the key stages of mitosis in a single somatic cell, ensuring a clear transition and coherent progression between meiotic divisions I and II in gonadal germ cells.</p> <p>Weight = +0.5 Accurate reproduction of subcellular elements and dynamics: chromosome separation following metaphase plate alignment, coordinated movement of spindle fibers and centrioles, and continuous changes and details of the cell membrane/cytokinesis.</p> <p>Weight = +0.2 Accurately describe genetic outcomes and differences: mitosis produces two genetically identical diploid daughter cells, while meiosis results in four genetically diverse haploid gametes, highlighting the mechanistic distinctions.</p> <p>Weight = +0.2 Compliance with technical specifications and viewing angle requirements: within an approximate total duration of 5 seconds, information is organized clearly; microscopic close-up focuses on the single-cell subject; the camera remains stable, transitions are clear, and the subject is unobstructed.</p> <p>Weight = -0.5 Presence of fundamental conceptual and procedural errors: confusion between mitosis and meiosis, incorrect sequencing of stages, inaccurate ploidy descriptions, omission of the two meiotic divisions, or failure to represent the single-cell focus.</p>

1026 **J EXAMPLE OF REAL BIOLOGICAL VIDEO CLIPS**
1027
1028

1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079

Caption: A high-magnification microscopic view reveals a tightly packed layer of rounded, polygonal cells with finely granular cytoplasm and dark, well-defined nuclei, their thin bright borders forming a mosaic-like tissue texture. At the center, one prominent cell is caught in the midst of mitosis, distinguished by its luminous, ring-shaped outline and a nucleus that appears elongated or partially split as condensed chromatin masses align and separate—visual cues of an active division stage. Surrounding cells remain in interphase, showing intact round nuclei with small dark spots suggestive of nucleoli, providing contrast to the dynamic structural rearrangements within the dividing cell. The overall scene captures the subtle shifts in texture, contrast, and intracellular organization characteristic of living cells undergoing cell division.

Figure 11: Example of real biological video clips.