# Appendix for DynaCartoon: 3D Dynamics-Guided Cartoon Video Generation from a Single Illustration and Forces

Anonymous Author(s)
Submission Id: 693

## Overview

The appendix provides detailed implementations, additional analyses, and supplementary experiments. Specifically, we include the following sections:

- Appendix A: More Baselines.
- Appendix B: Additional Results.
- Appendix C: Simulation Parameter Sets.
- Appendix D: Details and Validation of VLMs.

## A More Baselines

Ideally, our method should compare with PhysAnimator [10], which is based on 2D mesh simulation for cartoon video generation. However, because this method is closed-source, we limited our discussion to a theoretical comparison in the related work section.

Furthermore, since our method is a trajectory-driven video generation framework, the comparative analysis should include this type of model. We select the advanced methods MOFA-Video [7] and DragAnything [9] for comparison. Due to space constraints, the results were not presented in the main paper. As illustrated in Figure 1, these baselines struggle to handle complex 3D motions such as barrel rolls due to the lack of three-dimensional physical modeling, resulting in severe physical artifacts. Table 1 demonstrates the superiority of our method.




**Figure 1: Additional Qualitative Results. We compared DynaCartoon (Ours) against several trajectory-driven baselines, including MOFA-Video (MOFA), DragAnything (DA), and ForcePrompting (FP).**

**Table 1: Quantitative Comparison With Trajectory-driven Baselines. Left: VBench Metrics [2] including Subject Consistency (SC), Motion Smoothness (MS), Aesthetic Quality (AQ), and Imaging Quality (IQ). Right: VLM Evaluation (GLM-4.5V) [? ] covering Physical Realism (PR) and Visual Quality (VQ). Bold indicates best, underline indicates second best.**

| Model | VBench Metrics | | | | VLM | |
|---|---|---|---|---|---|---|
| | SC↑ | MS↑ | AQ↑ | IQ↑ | PR↑ | VQ↑ |
| MOFA | 0.161 | 0.985 | 0.585 | 0.671 | 0.493 | 0.612 |
| DA | 0.172 | 0.988 | 0.579 | 0.653 | 0.459 | 0.635 |
| FP | 0.181 | 0.994 | 0.566 | 0.692 | 0.598 | 0.667 |
| **Ours** | **0.215** | **0.996** | **0.614** | **0.719** | **0.704** | **0.742** |

**Table 2: Hyperparameter study. A study on the sketch extraction hyper-parameters $\tau_{bg}$ and $\tau_{dyn}$.**

| Thresholds | Imaging↑ | Aesthetic↑ | Motion↑ | Consist↑ | PhysReal↑ |
|---|---|---|---|---|---|
| $\tau_{bg} = 0.05, \tau_{dyn} = 0.05$ | 0.672 | 0.577 | 0.993 | 0.209 | 0.674 |
| $\tau_{bg} = 0.1, \tau_{dyn} = 0.05$ | 0.678 | 0.585 | 0.994 | 0.212 | 0.685 |
| $\tau_{bg} = 0.2, \tau_{dyn} = 0.2$ | 0.624 | 0.576 | 0.991 | 0.205 | 0.654 |
| $\tau_{bg} = 0.1, \tau_{dyn} = 0.2$ | 0.691 | 0.587 | 0.993 | 0.212 | 0.693 |
| Ours ($\tau_{bg} = 0.1, \tau_{dyn} = 0.1$) | **0.699** | **0.614** | **0.995** | **0.215** | **0.704** |

## B Additional Results

### B.1 More Qualitative Visualizations

To further demonstrate the robustness of our method, we present a broader range of generated results in Figure 2. These examples encompass various physical phenomena and material types, including cloth, rigid bodies, and fluids.

### B.2 Hyperparameter Study

We provide additional ablation studies on the sketch extraction hyper-parameters, specifically the background filtering threshold $\tau_{bg}$ and the dynamic motion threshold $\tau_{dyn}$, as presented in Table 2. Note that prior to processing, we normalize the sketch images to a range of $[0, 1]$. The performance of different configurations is consistently inferior to that of our optimal setting ($\tau_{bg} = 0.1$, $\tau_{dyn} = 0.1$). Specifically, lower threshold values (e.g., 0.05) introduce excessive noise and capture minor irrelevant variations into the structural conditioning, thereby degrading the aesthetic and imaging quality. Conversely, larger thresholds (e.g., 0.2) lead to the omission of crucial geometric details and motion boundaries, which diminishes both motion smoothness and physical realism.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



**Figure 2: Additional qualitative results.**

## C Simulation Parameter Sets

Our composite simulator is implemented based on the Genesis Framework [1], primarily utilizing its Rigid Body Solver, MPM Solver [3], and PBD Solver [6]. Each solver requires a unique set of physical parameters. We referenced and optimized the initial parameter sets of WonderPlay [4]. Table 3 presents the physical parameters and their corresponding default values for these solvers. These default values are utilized in the simulation process when physical parameters cannot be estimated using a VLM, and are subsequently adjusted manually.

To provide a deeper understanding of the simulation configuration, we briefly describe the impact of key parameters listed in Table 3:

- Sub-steps number: Sets the quantity of internal calculation loops per frame to enhance stability and prevent simulation errors such as tunneling.
- Sampled particle size: Determines the physical radius of the discrete particles used to discretize the mesh volume for simulation.
- Grid density: Specifies the resolution of the background computation grid in MPM solvers, directly impacting the fineness of the simulation details.
- Poisson's ratio: Describes the material deformation response by measuring the ratio of transverse contraction to longitudinal extension.
- Young's modulus: Defines the stiffness for objects, determining how much they deform under a specific load.
- Particle material Friction angle: Critical for granular media like sand, determining the internal friction that allows piles to hold their shape.
- Cloth material stretch compliance: Represents the inverse stiffness for stretching, defining how easily the fabric elongates when pulled.
- Cloth material bending compliance: Represents the inverse stiffness for bending, distinguishing between stiff materials like leather and soft drapes like silk.
- Smoke material viscosity coefficient: Quantifies the internal friction within the smoke fluid, affecting how smoothly the smoke diffuses and flows.

## D Details and Validation of VLMs

### D.1 Foreground Object Segmentation

Precisely identifying and segmenting foreground objects from a static image is the initial and critical step for 3D world reconstruction. We utilize GPT-5.2 combined with the image to perform high-level semantic reasoning, recognizing the object categories relevant to the target motion, and then input these to the Grounded-SAM [8] model to achieve open-set instance detection and segmentation. Figure 3 shows the detailed prompt used for object detection.

**Table 3: Simulation Parameters, Default Values, and Units.**

| Parameter | Default Value | Unit |
|---|---|---|
| **General Simulation** | | |
| Step time | $1 \times 10^{-2}$ | s |
| Sub-steps number | 10 | – |
| Sampled particle size | $1 \times 10^{-2}$ | m |
| Gravity | $(0, 0, -9.8)$ | $m/s^2$ |
| **Rigid Body Solver** | | |
| Mass | 1 | kg |
| Friction coefficient | 0.5 | – |
| **MPM Solver** | | |
| Grid density | 128 | – |
| Poisson's ratio | 0.1 | – |
| Elastic material Young's modulus | $1 \times 10^5$ | Pa |
| Liquid material Young's modulus | $1 \times 10^7$ | Pa |
| Particle material Young's modulus | $1 \times 10^6$ | Pa |
| Particle material Friction angle | 45 | ° |
| **PBD Solver** | | |
| Cloth material stretch compliance | $1 \times 10^{-6}$ | $Pa^{-1}$ |
| Cloth material bending compliance | $1 \times 10^{-4}$ | $Pa^{-1}$ |
| Smoke material viscosity coefficient | 0.3 | $Pa \cdot s$ |

We validated the reliability of our segmentation pipeline on a balanced subset of 60 images, 10 per material category. To conduct a comprehensive evaluation, we manually annotated both the target object categories to verify recognition accuracy and the ground truth masks to assess segmentation precision. As shown in Table 4, our method demonstrates robust semantic understanding with an overall Accuracy of 0.98. Furthermore, it achieves an average mIoU of 0.89, ensuring precise geometry for subsequent 3D mesh reconstruction.

**Table 4: Quantitative validation of foreground object recognition and segmentation across different material types.**

| | Rigid | Elastic | Cloth | Liquid | Smoke | Particle | Overall |
|---|---|---|---|---|---|---|---|
| **Accuracy** | 1.00 | 1.00 | 1.00 | 0.90 | 1.00 | 1.00 | 0.98 |
| **mIoU** | 0.92 | 0.94 | 0.92 | 0.86 | 0.84 | 0.85 | 0.89 |

## D.2 Physics Parameter Estimation

Perceiving physical parameters is a crucial step in simulations. Our method leverages the multimodal capabilities of GPT-5.2 to infer potential physical properties of objects directly from input images. A specific prompt is designed to achieve this goal, as shown in Figure 4. We adopt the parameter estimation strategy from PhysGen [5]. Given its proven reliability in LLM-based physical reasoning, we implement this approach without further validation.

## D.3 Physical Plausibility Evaluation Metric

To introduce a different quality evaluation method, we additionally employ GLM-4.5V[**?** ] to score the physical realism of generated animations. The specific prompt is shown in Figure 5.

> Analyze the animated image to determine the main dynamic action or interaction implied by the image and relevant context. Identify all foreground objects that are either involved in the action or essential to the scene structure for physics simulation. Classify each identified object into its semantic category (e.g., "Green bucket", "Basketball"). Output format requirement: Present the object categories as a clear numbered list.
>
> Example output format: {Green bucket, Basketball}

**Figure 3: Prompt for GPT-5.2 foreground object segmentation**

> Answer the questions for each object in the illustration. Provide answers in the following format for each object: 'Object number, name, material type (rigid, elastic, liquid, cloth, smoke, particle), mass, density, Young's modulus, Poisson's ratio, requires internal filling (yes/no), friction coefficient, friction angle, stretch compliance, bending compliance, viscosity coefficient. For parameters that are non-applicable to an object's material type, use 'N/A'.
>
> If there are multiple objects in the picture, respond for each object on a new line in the specified format.
>
> What is each object's name ?
> What is its material type ? What is its mass in kg ?
> What is its density in kg/m³ ?
> What is its Young's modulus in Pa ?
> What is its Poisson's ratio ?
> Does the object require internal filling for simulations ?
> What is its friction coefficient ? (rigid/elastic)
> What is its friction angle ? (particle)
> What is its stretch compliance ? (cloth )
> What is its bending compliance ? (cloth )
> What is its viscosity coefficient ? (smoke)

**Figure 4: Prompt for GPT-5.2 physics reasoning**

To validate the reliability of our automated PhysReal metric derived from GLM-4.5V, we compared it against human Motion Plausibility preferences. We observed a clear agreement where higher PhysReal scores consistently aligned with the majority of human votes across all comparisons. This ranking consistency confirms that PhysReal serves as a reliable proxy for evaluating human-perceived physical realism in video generation.

Please evaluate the quality of the following twelve generated animations with a focus on physical realism. This criterion specifically measures how accurately the animations adhere to the laws of physics and whether they realistically reflect physical properties such as elasticity, friction, gravity, momentum, and collision dynamics.

For each animation, your evaluation will be based on 10 evenly sampled frames provided. These frames are representative of the video's content and should be used to assess the consistency of physical behavior across the video's duration.

Please assign a score to each video on a scale of 0 to 1, where 1 indicates perfect physical realism and 0 indicates a complete lack of adherence to physical laws.

Strictly follow this format for your output, with each line ending in a semicolon (;). Do not include any additional explanations, comments, or formatting:

Video 1: Physical Realism Score: [score];
Video 2: Physical Realism Score: [score];
.....
Video 12: Physical Realism Score: [score];

Note: The first illustration provided is the input image, which establishes the initial state of the scene. The 10 evenly spaced frames that follow correspond to generated video number {idx + 1}, and should be analyzed in sequence to evaluate the physical consistency of the video's progression.

**Figure 5: Prompt for GLM-4.5V video evaluation**

## References


[1] Genesis Authors. 2024. Genesis: A universal and generative physics engine for robotics and beyond. (2024). https://genesis-embodied-ai.github.io/

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

[3] Chenfanfu Jiang, Craig Schroeder, Joseph Teran, Alexey Stomakhin, and Andrew Selle. 2016. The material point method for simulating continuum materials. In *Acm siggraph 2016 courses*. 1–52.

[4] Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. 2025. WonderPlay: Dynamic 3D Scene Generation from a Single Image and Actions. *arXiv preprint arXiv:2505.18151* (2025).

[5] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. 2024. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*. Springer, 360–378.

[6] Matthias Müller, Bruno Heidelberger, Marcus Hennix, and John Ratcliff. 2007. Position based dynamics. *Journal of Visual Communication and Image Representation* 18, 2 (2007), 109–118.

[7] Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. 2024. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European Conference on Computer Vision*. Springer, 111–128.

[8] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024).

[9] Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. 2024. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*. Springer, 331–348.

[10] Tianyi Xie, Yiwei Zhao, Ying Jiang, and Chenfanfu Jiang. 2025. Physanimator: Physics-guided generative cartoon animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 10793–10804.