Animating the Past: Reconstruct Trilobite via Video Generation

Authors: XiaoranWu, ZienHuang, ChonghanYu

Date: 2024-11-12T00:00:00+00:00

Abstract

Paleontology, the study of past life, fundamentally relies on fossils to reconstruct ancient ecosystems and understand evolutionary dynamics. Trilobites, as an important group of extinct marine arthropods, offer valuable insights into Paleozoic environments through their well-preserved fossil records. Reconstructing trilobite behaviour from static fossils will set new standards for dynamic reconstructions in scientific research and education. Despite the potential, current computational methods for this purpose like text-to-video (T2V) face significant challenges, such as maintaining visual realism and consistency, which hinder their application in science contexts. To overcome these obstacles, we introduce an automatic T2V prompt learning method. Within this framework, prompts for a fine-tuned video generation model are generated by a large language model, which is trained using rewards that quantify the visual realism and smoothness of the generated video. The fine-tuning of the video generation model, along with the reward calculations make use of a collected dataset of 9,088 Eoredlichia intermedia fossil images, which provides a common representative of visual details of all class of trilobites. Qualitative and quantitative experiments show that our method can generate trilobite videos with significantly higher visual realism compared to powerful baselines, promising to boost both scientific understanding and public engagement.

Full Text

Preamble

Animating the Past: Reconstructing Trilobites via Video Generation

Xiaoran Wu*
AI Lab, Yishi Inc., Hangzhou, China
wuxr18@tsinghua.org.cn

Zien Huang*
The International Department, Experimental High School Attached to Beijing Normal University
keane.huangzien@gmail.com

Chonghan Yu
School of Ocean Sciences, China University of Geosciences, Beijing, China
yuchonghan@foxmail.com

*These authors contributed equally to this work.

Fig. 1 [Figure 1: see original paper]: We design and train the first text-to-video framework that automatically learns to refine prompts to generate visually realistic trilobites adhering closely to pronounced and authentic trilobite characteristics in more fluid and lifelike videos. The first prompting image is courtesy of [1].

Abstract

Paleontology, the study of past life, fundamentally relies on fossils to reconstruct ancient ecosystems and understand evolutionary dynamics. Trilobites, as an important group of extinct marine arthropods, offer valuable insights into Paleozoic environments through their well-preserved fossil records. Reconstructing trilobite behavior from static fossils will set new standards for dynamic reconstructions in scientific research and education. Despite this potential, current computational methods such as text-to-video (T2V) face significant challenges, including maintaining visual realism and consistency, which hinder their application in scientific contexts.

To overcome these obstacles, we introduce an automatic T2V prompt learning method. Within this framework, prompts for a fine-tuned video generation model are generated by a large language model, which is trained using rewards that quantify the visual realism and smoothness of the generated video. The fine-tuning of the video generation model, along with the reward calculations, makes use of a collected dataset of 9,088 *Eoredlichia intermedia* fossil images, which provides a common representative of visual details across all classes of trilobites. Qualitative and quantitative experiments demonstrate that our method can generate trilobite videos with significantly higher visual realism compared to powerful baselines, promising to enhance both scientific understanding and public engagement.

Index Terms—Trilobite, *Eoredlichia intermedia*, Text-to-Video, Multimodal Large Language Model, Learning from Human Feedback

---

I. Introduction

Paleontology, the study of prehistoric life, relies heavily on the fossil record to reconstruct past ecosystems, understand evolutionary processes, and decipher the biology of extinct organisms [2], [3]. As an extinct group of marine arthropods, trilobites are among the most iconic and well-studied fossils [2], [4], [5], providing critical insights into Paleozoic ecosystems. Reconstructing the behavior and locomotion of trilobites is of great research and educational interest [1], as such dynamic reconstructions help formulate hypotheses about trilobites' living environments and the functional morphology and ecological roles of these ancient creatures [4]–[7]. Furthermore, from an educational perspective, reconstruction provides tangible visualization of trilobite appearance and behavior, thus bridging the gap between abstract scientific knowledge and public understanding [8].

Despite the abundance of trilobite fossils, reconstructing their behavior and movement remains challenging, primarily due to the static nature of fossil remains. Fortunately, recent advancements in generative artificial intelligence (AI) and computational techniques provide new opportunities to address these challenges [9]–[13]. Integrating AI into paleontological research not only showcases the potential of extending machine learning into a natural research field that AI has not studied extensively before [14] but also can hopefully enhance our understanding of trilobite ethology and shed new light on its study.

Among generative AI techniques, video generation [1], [15], [16] is particularly suitable for simulating trilobite movement in a dynamic, visually engaging manner. However, current video generation methods encounter several challenges that hinder their application to paleontological reconstructions. Primarily, as demonstrated in our qualitative studies, existing methods struggle with maintaining the realism of depicted trilobites, with creatures appearing unrealistic or oddly shaped [3]. This lack of realism significantly detracts from viewer engagement and reduces the educational and research value of the visualizations. Moreover, the consistency of generated videos often falls short, with noticeable discrepancies between consecutive frames [17], [18]. Such inconsistencies are particularly problematic in longer sequences, leading to choppy transitions that disrupt the fluid simulation of trilobite movement.

To tackle these issues, we propose a novel approach that embeds the evaluation of trilobite realism and video smoothness directly into the video generation workflow. Our solution leverages diffusion models [19]–[22], which have demonstrated impressive capabilities in producing realistic images and videos from textual descriptions. We employ these models to create animated segments that capture various aspects of trilobite movement, guided by descriptive prompts generated by a large language model (LLM) [23]–[25]. The cornerstone of our method involves assessing the smoothness of transitions and the accuracy of trilobite appearance in these animations, compared against a curated collection of trilobite fossil images. This assessment acts as a feedback mechanism to fine-tune the LLM that generates prompts for the text-to-animation model [26]–[28],

enhancing animation fidelity. The objective is twofold: to produce animations that accurately depict trilobite appearance and movement while ensuring seamless transitions, adding complexity to the model's training but proving crucial for high-quality video output.

Our methodology encompasses several stages: initially, we generate basic animated segments from LLM-generated prompts. These segments are then pieced together, and the composite video is evaluated for transition quality and content realism. The evaluation results serve as reward signals to update the LLM with preference optimization [29], [30] to refine the animations. This cycle of generation, evaluation, and enhancement repeats until the video meets our criteria for smoothness and realism.

We comprehensively evaluate our method both qualitatively and quantitatively against state-of-the-art text-to-animate and text-to-video academic research [27] and commercial tools [31], [32]. The results show clear advances in paleontological visualization in terms of content realism and video continuity. Furthermore, we provide ablation studies to demonstrate the contribution of each component in our learning framework. We hope that this pioneering integration of technology and paleontology makes significant contributions to the field of synthetic media generation and opens new pathways for visualizing and understanding prehistoric entities and exploring ancient life.

II. Related Work

Our method of training the Large Language Model (LLM) that generates prompts relates to Reinforcement Learning from Human Feedback (RLHF), an important technique for ensuring LLM outputs align with human preferences [33]–[35]. In our work, the counterpart of human preference is defined by metrics regarding content realism and video continuity. Typically, RLHF initially learns a reward model (RM) [36] from human preferences and then optimizes the supervised fine-tuned LLM model with reinforcement learning algorithms (e.g., PPO [37]) to maximize cumulative rewards from the RM. However, training the reward model is time-consuming and computation-intensive [36].

Direct Preference Optimization (DPO) [29] avoids training the reward model by directly aligning LLMs to best satisfy human preferences using a simple classification objective. The recently proposed KTO [30] extends DPO by maximizing utility functions derived from prospect theory [38] for accurate human utility modeling. To achieve better stability and robustness, we utilize calculated realism and continuity rewards to order different LLM outputs (prompts to the animation generation model) and use KTO for preference optimization.

Video generation methods like Tune-a-Video [13] extend text-to-image (T2I) models to generate multiple images simultaneously by incorporating a tailored spatio-temporal attention mechanism and an efficient one-shot tuning strategy

to learn continuous motion among generated images. Text2Video-Zero [12] proposes a cost-effective approach requiring no training or optimization by leveraging existing T2I synthesis methods adapted for video generation. CogVideo [16] proposes a multi-frame-rate hierarchical training strategy to better align text and video clips on large-scale text-video datasets. Commercial video generation tools are also setting significant benchmarks; we empirically compare our method against Pika [31] and Gen3 [32] for evaluation.

Text-to-Animation (T2A) is another video generation approach that extends pretrained T2I models by incorporating temporal structures [26], [27]. AnimateDiff [27] introduces a plug-and-play motion module enabling T2A model training without model-specific tuning. In this paper, we employ the T2A method to generate trilobite animations from user prompts, focusing on enhancing temporal coherence and content realism.

We now introduce the preliminaries of RLHF and T2A techniques upon which we develop our method.

III. Preliminaries

RLHF. The training of modern Large Language Models (LLMs) involves three phases as outlined in [23], [33], [39], [40]. (1) **Pretraining:** This phase involves training an initial model π_0 on a large text corpus to optimize prediction of the next token based on preceding text [41]–[43]. (2) **Supervised Fine-tuning (SFT):** The model is further trained on task-specific data that generally includes targeted instructions and expected responses to refine its utility for practical applications [44]. This fine-tuned model is denoted as π_{ref}. (3) **RLHF:** This step uses a preference dataset $\mathcal{D}$ containing tuples (x, y_w, y_l) where x is the input and y_w, y_l are the preferred and less preferred outputs, respectively [29], [34]. A Bradley-Terry model [45] calculates preferences: $p^*(y_w \succ y_l|x) = \sigma(r^*(x, y_w) - r^*(x, y_l))$, where σ is the logistic function. A reward model r_ϕ is trained by minimizing the negative log-likelihood of preference data in set $\mathcal{D}$ [33]:

$$\mathcal{L}_R(r_\phi) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}[-\log\sigma(r_\phi(x, y_w) - r_\phi(x, y_l))].$$

To balance reward maximization with linguistic correctness, a KL divergence penalty prevents the model from deviating excessively from the reference model π_{ref}:

$$\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}[r_\phi(x, y)] - \beta D_{KL}[\pi_\theta(\cdot|x)||\pi_{ref}(\cdot|x)].$$

This non-differentiable objective requires an RL approach like PPO [37] for optimization. The computational demands and instability of training the reward

model r_ϕ led to Direct Preference Optimization (DPO) [29], which provides a stable alternative that trains directly on preference pairs with similar optimal policy convergence performance:

$$\mathcal{L}_{DPO}(\pi_\theta, \pi_{ref}) = \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right].$$

T2A. One approach for short video generation animates a text-to-image (T2I) [46] model by integrating temporal dynamics. Following related literature [16], [27], a batch of video data is represented as 5-dimensional tensors $x \in \mathbb{R}^{b\times c\times f\times h\times w}$, where b denotes the batch axis, f represents the frame-time axis, and c, h, w are the channels, height, and width of each video frame, respectively. The text-to-animation process begins by encoding each frame of a video data batch $x_{1:f} \in \mathbb{R}^{b\times c\times f\times h\times w}$ into latent representations $z_{1:f,0}$ using a pre-trained auto-encoder. These representations are subsequently perturbed by noise according to the forward diffusion schedule [20], [47]:

$$z_{1:f,t} = \sqrt{\bar{\alpha}_t}z_{1:f,0} + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

In this inflated model, the noisy latent representations along with corresponding text prompts serve as inputs for predicting the noise added during the diffusion process. The training objective for T2As can be formulated as:

$$\mathcal{L} = \mathbb{E}_{\mathbb{E}(x_{1:f}),y,\epsilon\sim\mathcal{N}(0,I),t}[\|\epsilon - \epsilon_\theta(z_{1:f,t}, t, \tau_\theta(y))\|^2].$$

By inflating the model with the additional temporal axis [48], this formulation emphasizes the critical role of temporal coherence and dynamic content adaptation in generating animations from textual descriptions. However, this approach may yield animations that are not only brief in duration but also suffer from less smooth transitions between frames and inconsistencies in object appearances across different frames. These issues primarily arise from the model's limitations in maintaining consistent motion patterns and visual quality throughout the sequence [17]. This challenge is particularly pronounced when the model attempts to interpolate complex dynamics, a task that demands high fidelity in temporal and spatial representations. The difficulty lies in the model's capacity to accurately generate and link successive frames where each must evolve naturally from its predecessor while adhering to the dynamics specified by the textual description.

IV. Method

This section describes our method for addressing the challenge of maintaining motion smoothness and visual realism throughout video sequences. The proposed framework synergizes the power of a large language model (SCRIPT WRITER) that generates prompts and a fine-tuned text-to-animation model (VIDEO GENERATOR). Our main technical novelty lies in the design of the optimization algorithm for SCRIPT WRITER. In effect, we design a contextual bandit learning task for SCRIPT WRITER, a concept popular in cutting-edge LLM research such as direct preference optimization (DPO [29]).

For fine-tuning the VIDEO GENERATOR and training the SCRIPT WRITER, we collect a set $\mathcal{R}$ of 9,088 *Eoredlichia intermedia* fossil images, which include numerous specimens covering different stages of individual development. These images provide a common representative of visual traits across all classes of trilobites to enhance the visual details of generated content. While these real trilobite fossil images do not mean that the videos produced in this study can fully reproduce real trilobite structure, using these real fossil details greatly supplements the scarcity and errors of trilobite images in web footage, enhancing trilobite structure and details in the videos.

A. Prompt and Video Initialization

As the first step, SCRIPT WRITER $\pi(\theta_0)$, where θ_0 represents the initial parameters, generates an initial prompt y_0 for the text-to-animation model with the format $y_0 = (t_1 : y_0^1; t_2 : y_0^2; \cdots, t_N : y_0^N)$, where $y_0^n, n \in [N]$ is a textual description of the appearance and expected movement of a trilobite in animation clip n, and $t_n, n \in [N]$ is the frame index where animation clip n will start in the final video.

This initial prompt y_0 directs the text-to-animation diffusion model VIDEO GENERATOR to generate N initial animation clips $(c_0^1(y_0), \cdots, c_0^N(y_0))$ that are concatenated sequentially to produce an initial video $z_0^{1:f}(y_0)$. Before generating this initial video, the VIDEO GENERATOR has been fine-tuned on the collected fossil image dataset $\mathcal{R}$ to enhance the model's ability to generate detailed textures and structures observed in fossil images.

B. Reward Design for SCRIPT WRITER

The second step involves designing reward signals to train SCRIPT WRITER, with the goal of refining the initial prompt y_0 so that the resulting video achieves better quality in terms of transition smoothness and visual realism. Specifically, the reward for prompt y_0 is designed as a summation of two components: $r(y_0) = r_s(y_0) + r_a(y_0)$, where r_s measures frame transition smoothness and r_a measures visual realism of trilobites in the generated video.

Smoothness of Frame Transition. To assess video smoothness, we compute the Fréchet Inception Distance (FID) [50] between adjacent frames. Consider

two frames $x_t \in \mathbb{R}^{c\times h\times w}$ and $x_{t+1} \in \mathbb{R}^{c\times h\times w}$, where c, h, w represent channel, height, and width, respectively. We first use a pre-trained InceptionV3 network [51] to extract image features (pool3 layer) z_t and z_{t+1} from frames x_t and x_{t+1}, then compute the FID score for consecutive frames:

$$\mathrm{FID}_t = \|z_t - z_{t+1}\|^2.$$

After obtaining FID scores for all consecutive frames, we derive the transition smoothness reward $r_s(y_0) = -\sum_{t=1}^{f} \mathrm{FID}_t$. For fine-grained control, the reward can be calculated per clip: $r_s(y_0^n) = -\sum \mathrm{FID}_t$.

Visual Realism. To ensure scientific rigor, we compare visual details of generated content against real trilobite fossil samples from $\mathcal{R}$. For a video consisting of multiple frames, we expect no frame to contain trilobites with morphological details deviating significantly from realistic data. We therefore design a max-min objective:

$$r_a(y_0^n) = -\max_{x\in[f]} D(x, r).$$

Here, D is a distance function measuring morphological similarity between a generated trilobite and a reference image. The reference image set contains trilobite fossils from different growth stages, various sizes, different preservation states, and different geological periods. Therefore, we have clear evidence that a generated trilobite is visually realistic if it is morphologically similar to at least one reference image, captured by the minimum operation in Eq. 7. We then find the frame most different from the reference set; minimizing this guarantees no frame deviates too far from the reference set.

In practice, we use the ORB (Oriented FAST and Rotated BRIEF) detector [52] as the distance function D. ORB is a fast, efficient feature detection algorithm combining the FAST keypoint detector and BRIEF descriptor, providing robust performance suitable for extracting morphological details. We then use the BF (Brute-Force) method [53], [54] for matching features, which computes distances between every descriptor pair, typically employing Hamming distance for binary descriptors as utilized in our case.

C. Training SCRIPT WRITER

Having defined the reward signals, we now introduce the third step: training SCRIPT WRITER. We note that the rewards defined above are all negative, can be large in magnitude, and are prone to noise, indicating these rewards may be ineffective for training the SCRIPT WRITER LLM with algorithms like PPO [37], which are sensitive to specific reward values. To address this, we propose ordering prompts based on rewards then applying preference optimization, which has proven effective in RLHF literature [33] and is more robust when reward values are noisy.

Specifically, we collect a training dataset $\mathcal{D}$ where each sample contains a query x, a desirable generation y_d, and an undesirable generation y_u, with $r(y_d) > r(y_u)$. We use Kahneman-Tversky Optimization (KTO) [30] to train SCRIPT WRITER. Letting λ_y denote λ_D (λ_U) when y is desirable (undesirable), where λ_D and λ_U are constants, the KTO loss is:

$$\mathcal{L}_{KTO}(\theta_0) = \mathbb{E}_{x,y\sim\mathcal{D}}[\lambda_y - v(x,y)],$$

where

$$r_{\theta_0}(x,y) = \log \frac{\pi_{\theta_0}(y|x)}{\pi_{ref}(y|x)},$$

$$z_0 = \mathbb{E}_{x'\sim\mathcal{D}}[KL(\pi_{\theta_0}(y'|x') \parallel \pi_{ref}(y'|x'))],$$

and

$$v(x,y) = \begin{cases} \lambda_D \sigma(\beta(r_{\theta_0}(x,y) - z_0)) & \text{if } y \sim y_d|x; \\ \lambda_U \sigma(\beta(z_0 - r_{\theta_0}(x,y))) & \text{if } y \sim y_u|x. \end{cases}$$

The following section presents examples showing how KTO training improves prompts. After KTO training, SCRIPT WRITER parameters update to θ_1, leading to updated prompts y_1. θ_1 can be further improved by running KTO given the new video generated by y_1, repeating this process until the new video is satisfactory.

V. Experiments

This section presents experiments testing our method's effectiveness, providing both qualitative and quantitative evaluation of generated videos. We compare against strong baselines including previous text-to-animation work (AnimateDiff [27]) and powerful commercial text-to-video tools (Pika Labs [31] and Gen-3 [32]). We also conduct ablation studies to show the separate contributions of SCRIPT WRITER training and VIDEO GENERATOR fine-tuning.

A. Qualitative Results I: Visual Realism

Comparison with baselines. Fig. 2 [Figure 2: see original paper] showcases qualitative comparisons of trilobite renderings produced by different models, evaluating each model's ability to generate realistic trilobites in various dynamic backgrounds.

The first row shows trilobites on the ocean floor with a volcanic eruption background. The Pika model generates a trilobite with unrealistic segmentation. The Runway model shows more realistic structure but fails to capture authentic trilobite exoskeleton texture. The AnimateDiff model produces an oversimplified trilobite, with the volcano dominating the image. In contrast, our model generates trilobites displaying intricate segmentation, realistic texturing, and coloration that blends well with the naturalistic ocean floor setting, making them the most lifelike.

The second row depicts three trilobites among aquatic plants on the seabed. Pika's trilobites resemble no known types. Runway's versions show better background integration but remain somewhat artificial. AnimateDiff's trilobites lack depth and textural detail. Our model, however, shows trilobites with precise, well-defined segmentation and natural colors that harmonize with the underwater environment, enhancing scene realism.

The third row captures a single trilobite moving along the ocean floor, focusing on environmental interaction such as sediment displacement. Pika's rendition again lacks realistic appearance. AnimateDiff's trilobite appears round. Meanwhile, our model produces a realistic trilobite interacting with its surroundings, showing sediment displacement that suggests natural weight and presence in the water.

In summary, our model outperforms others in creating trilobites with realistic anatomical features, textural fidelity, and appropriate environmental interactions. This qualitative analysis underscores our method's ability to generate video content that closely mirrors true trilobite appearance.

Comparison with ablations. Two components contribute to visual realism: fine-tuning the T2A model and performing preference optimization for SCRIPT WRITER. Fig. 3 [Figure 3: see original paper] shows the influence of these components, presenting two examples demonstrating results before and after preference optimization, focusing on visual quality of trilobite renderings. Each example shows a video frame, the most closely matching dataset image, and corresponding prompts.

In the first example, the pre-optimization frame shows a trilobite on the ocean floor with a volcanic eruption background. The trilobite appears somewhat blended into the background, lacking distinct features and resulting in a low match score of 0.07. The prompt focuses on general trilobite presence amid a dynamic background. After optimization, the frame exhibits a trilobite with more emphasized and defined hard shells, enhancing visibility and structural integrity against the complex background. This improvement stems from the updated prompt, which now specifically highlights the trilobite's hard shell. The match score significantly improves to 0.35, indicating closer resemblance to the most similar reference image showing clearer, more detailed trilobite features.

In the second example, the pre-optimization frame captures a trilobite moving across the ocean floor with anomalocaris in the background. Initially, the trilo-

bite lacks prominent distinguishing features, yielding a match score of 0.158. After preference optimization, the frame shows the trilobite with enhanced distinguishing features such as longitudinal lobes and textural details, making it more realistic and akin to the reference image. Again, the updated prompt drives these changes by specifically pointing out these features, contributing to a raised match score of 0.33.

In both cases, preference optimization adjusts model rendering to enhance specific trilobite features contributing to greater visual realism. The targeted adjustments in post-optimization prompts are pivotal in directing the model to produce outputs that adhere more closely to reference images while showcasing more pronounced and authentic trilobite characteristics. This approach demonstrates the model's capability to adapt and refine its output by learning from preferences, ultimately yielding higher match scores and visually richer renderings.

B. Qualitative Results II: Smoothness

Fig. 4 [Figure 4: see original paper] displays frame sequences before and after preference optimization. In initial pre-optimization frames, the trilobite's movement appears somewhat jerky, particularly in its antennae. The corresponding prompt focuses on the trilobite gliding through the landscape. After preference optimization, frames show noticeable improvement in movement fluidity, with the trilobite seamlessly integrating into surrounding plant motion, creating a more naturalistic and visually appealing scene. This change occurs because the optimized prompt adds frames and emphasizes the smooth, effortless glide of the trilobite and its streamlined body, highlighting how these characteristics should be reflected in animation. This directive likely influenced the rendering process to focus on creating smoother, more coherent movement patterns.

Fig. 5 [Figure 5: see original paper] provides another example where SCRIPT WRITER learns to add words enhancing video smoothness. The comparison clearly demonstrates that post-optimization prompt changes lead to significant improvements in video smoothness.

C. Quantitative Results

We conduct quantitative comparisons to further evaluate our method.

Smoothness after KTO prompt training. Fig. 6 [Figure 6: see original paper] illustrates Fréchet Inception Distance (FID) scores between adjacent frames in a generated video sequence, comparing results before and after preference optimization. Before optimization, the blue line shows several peaks, particularly around frames 15 and 60-80, suggesting less smooth transitions with more noticeable visual discrepancies. After optimization, the dark line generally maintains lower FID scores throughout the sequence with fewer and lower peaks, indicating greater visual consistency and smoother transitions between frames. The overall trend demonstrates that preference optimization effectively reduces FID

scores across most of the video sequence, signifying improved smoothness and more visually coherent frames.

User study. We generate videos using four methods and conduct a user study evaluating performance on three criteria: smoothness, visual realism, and consistency with the prompt. Participants rate videos on a scale from 1 to 4, where 4 indicates the highest score. This scoring system is equivalent to Average User Ranking (AUR), with higher scores indicating superior performance across evaluated metrics.

TABLE I shows quantitative comparison results. Our method outperforms the other three methods in all evaluation criteria, indicating significant improvement in video generation quality. This is evident from higher scores across all three categories, confirming our approach's effectiveness in producing smooth, visually realistic videos consistent with given prompts. Particularly, our method's much higher scores regarding prompt consistency highlight the effectiveness of our prompt learning method.

VI. Conclusion

Our study leverages advanced generative AI techniques to address challenges in reconstructing trilobite behavior from fossil records. By integrating computational methods with paleontological research, we demonstrate potential for enhancing understanding of these ancient creatures. Our proposed video generation framework, which incorporates realism and smoothness assessments into the workflow, produces more accurate and dynamic visualizations of trilobite movements. These enhanced animations improve scientific insights while making the prehistoric world more accessible to the public. This interdisciplinary approach marks an advancement in both paleontology and (multi-modal) artificial intelligence, opening new avenues for future research and educational opportunities.

Acknowledgment

We deeply appreciate Qiang Ou (China University of Geosciences, Beijing), Degan Shu (Northwest University, Xi'an), Jian Han (Northwest University, Xi'an), and Meirong Cheng (Northwest University, Xi'an) for generously providing trilobite images that supported this study.

References

[1] A. El Albani, A. Mazurier, G. D. Edgecombe, A. Azizi, A. El Bakhouch, H. O. Berks, E. H. Bouougri, I. Chraiki, P. C. Donoghue, C. Fontaine et al.,

"Rapid volcanic ash entombment reveals the 3d anatomy of cambrian trilobites," *Science*, vol. 384, no. 6703, pp. 1429–1435, 2024.

[2] R. Fortey, "The palaeoecology of trilobites," *Journal of zoology*, vol. 292, no. 4, pp. 250–259, 2014.

[3] C. trilobite fossil from Utah, "Trilobites and end of cambrian explosion," *PNAS*, vol. 116, no. 10, pp. 3935–3937, 2019.

[4] J. Bergström, "Organization, life, and systematics of trilobites," in *Organization, life, and systematics of trilobites*, 1973, pp. 1–69.

[5] N. C. Hughes, "The evolution of trilobite body patterning," *Annu. Rev. Earth Planet. Sci.*, vol. 35, pp. 401–434, 2007.

[6] R. Levi-Setti, *Trilobites*. University of Chicago Press, 1995.

[7] Q. Ou, D. Shu, J. Han, X. Zhang, Z. Zhang, and J. Liu, "A juvenile redlichiid trilobite caught on the move: Evidence from the cambrian (series 2) chengjiang lagerstatte, southwestern china," *Palaios*, vol. 24, no. 7, pp. 473–477, 2009.

[8] M. J. Hopkins, "Development, trait evolution, and the evolution of development in trilobites," *Integrative and Comparative Biology*, vol. 57, no. 3, pp. 488–498, 2017.

[9] E. B. Hunt, *Artificial intelligence*. Academic Press, 2014.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni et al., "Make-a-video: Text-to-video generation without text-video data," in *The Eleventh International Conference on Learning Representations*, 2022.

[12] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15954–15964.

[13] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. [page numbers].

[14] A. Sohrabi, A. Kadkhodaie, and R. Kadkhodaie-Ilkhchi, "Artificial intelligence approach to palaeogeography and evolutionary trend analysis of laurentian brachiopod fauna in the rhynchotrema-hiscobeccus lineage," *Palaeogeography, Palaeoclimatology, Palaeoecology*, vol. 562, p. 110114, 2021.

[15] Y. Li, M. Min, D. Shen, D. Carlson, and L. Carin, "Video generation from text," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[16] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," arXiv preprint arXiv:2205.15868, 2022.

[17] W. Ren, H. Yang, G. Zhang, C. Wei, X. Du, S. Huang, and W. Chen, "Consisti2v: Enhancing visual consistency for image-to-video generation," arXiv preprint arXiv:2402.04324, 2024.

[18] R. Henschel, L. Khachatryan, D. Hayrapetyan, H. Poghosyan, V. Tadevosyan, Z. Wang, S. Navasardyan, and H. Shi, "Streamingt2v: Consistent, dynamic, and extendable long video generation from text," arXiv preprint arXiv:2403.14773, 2024.

[19] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 21696–21707, 2021.

[20] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[21] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.

[22] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850–10869, 2023.

[23] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[24] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand et al., "Mixtral of experts," arXiv preprint arXiv:2401.04088, 2024.

[25] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288, 2023.

[26] N. Bouali and V. Cavalli-Sforza, "A review of text-to-animation systems," *IEEE Access*, 2023.

[27] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," arXiv preprint arXiv:2307.04725, 2023.

[28] L. Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. [page numbers].

[29] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[30] K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela, "KTO: Model alignment as prospect theoretic optimization," 2024. [Online]. Available: https://arxiv.org/abs/2402.01306

[31] P. Lab, "Pika lab," https://pika.art/, 2024.

[32] Gen-2, "Gen-2: The forward generative model," https://research.runwayml.com/gen2, 2024.

[33] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray et al., "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.

[34] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan et al., "Training a helpful and harmless assistant with reinforcement learning from human feedback," arXiv preprint arXiv:2204.05862, 2022.

[35] Y. Wang, Q. Liu, and C. Jin, "Is RLHF more difficult than standard RL? A theoretical perspective," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[36] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," in *International Conference on Machine Learning*. PMLR, 2023, pp. 10835–10866.

[37] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[38] A. Tversky and D. Kahneman, "Advances in prospect theory: Cumulative representation of uncertainty," *Journal of Risk and uncertainty*, vol. 5, pp. 297–323, 1992.

[39] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier et al., "Mistral 7b," arXiv preprint arXiv:2310.06825, 2023.

[40] C. Wang, Y. Deng, Z. Lv, S. Yan, and A. Bo, "Q*: Improving multi-step reasoning for LLMs with deliberative planning," arXiv preprint arXiv:2406.14283, 2024.

[41] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu et al., "Deepseek llm: Scaling open-source language models with longtermism," arXiv preprint arXiv:2401.02954, 2024.

[42] T. Wei, L. Zhao, L. Zhang, B. Zhu, L. Wang, H. Yang, B. Li, C. Cheng, W. Lü, R. Hu et al., "Skywork: A more open bilingual foundation model," arXiv

preprint arXiv:2310.19341, 2023.

[43] T. Wei, B. Zhu, L. Zhao, C. Cheng, B. Li, W. Lü, P. Cheng, J. Zhang, X. Zhang, L. Zeng et al., "Skywork-moe: A deep dive into training techniques for mixture-of-experts language models," arXiv preprint arXiv:2406.06563, 2024.

[44] K. Lu, H. Yuan, Z. Yuan, R. Lin, J. Lin, C. Tan, C. Zhou, and J. Zhou, "# instag: Instruction tagging for analyzing supervised fine-tuning of large language models," in *The Twelfth International Conference on Learning Representations*, 2023.

[45] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[46] B. Li, X. Qi, T. Lukasiewicz, and P. Torr, "Controllable text-to-image generation," *Advances in neural information processing systems*, vol. 32, [page numbers].

[47] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 1415–1428, 2021.

[48] X. Shen, X. Li, and M. Elhoseiny, "Mostgan-v: Video generation with temporal motion styles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5652–5661.

[49] W. Myers, "Phacops trilobites print," https://fineartamerica.com/featured/phacops-trilobites-walter-myers.html?product=art-print, phacops Trilobites Art Print.

[50] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen, "The role of imagenet classes in fréchet inception distance," arXiv preprint arXiv:2203.06026, 2022.

[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. [page numbers].

[52] P. Aglave and V. S. Kolkure, "Implementation of high performance feature extraction method using oriented fast and rotated brief algorithm," *Int. J. Res. Eng. Technol*, vol. 4, pp. 394–397, 2015.

[53] N. Antony and B. R. Devassy, "Implementation of image/video copy-move forgery detection using brute-force matching," in *2018 2nd International conference on trends in electronics and informatics (ICOEI)*. IEEE, 2018, pp. 1085–1090.

[54] F. K. Noble, "Comparison of opencv's feature detectors and feature matchers," in *2016 23rd International Conference on Mechatronics and Machine Vision in Practice (M2VIP)*. IEEE, 2016, pp. 1–6.

[55] A. Photo, "2hg309h," https://www.alamy.com/an-illustration-of-a-trilobite-moving-about-on-a-cambrian-period-400-million-years-ago-sea-bottom-trilobites-are-a-well-known-fossil-group-image457370189.html, 2016, an illustration of a Trilobite moving about on a Cambrian Period (400 million years ago) sea bottom.

Note: Figure translations are in progress. See original paper for figures.