NeurIPS Paper Appendix for *ChronoMagic-Bench: A Benchmark for Metamorphic Evaluation of Text-to-Time-lapse Video Generation*

A Related Works: Text-to-Video Generation Models

The emergence of large-scale text-to-image models [92, 60, 59, 58, 42, 5, 94, 14, 54, 40] has significantly advanced the field of Text-to-Video (T2V) generation [66, 6, 7, 21, 73, 90]. Existing T2V architectures can be categorized into two types: U-Net-based and DiT-based. The former typically builds on Stable Diffusion [62], extending the 2D U-Net to a 3D U-Net by adding temporal layers, thereby achieving high-quality video generation [74, 15, 23, 4, 11, 41]. The latter focuses on recreating open-source structures similar to Sora [9], using the DiT (Diffusion-Transformer) [57] framework for T2V generation [43, 95, 93, 20]. However, the generation quality of DiT-based architectures still lags behind that of U-Net-based architectures. MagicTime [88] notes that although these models have achieved basic video generation, the videos are typically limited to simple actions and scenes, resulting in the production of general videos rather than those enriched with physical priors like metamorphic/time-lapse videos. For a more intuitive representation, we have detailed a comparison of the metamorphic video generation capabilities of different algorithms.

B More details about Automatic Metrics

B.1 Construction of retrieval sentences for Metamorphic Score

To obtain an effective Metamorphic Score (MTScore), we meticulously designed ten distinct retrieval texts to differentiate between time-lapse and normal videos. Although, in theory, only two retrieval sentences are needed to distinguish between general and time-lapse videos, multiple texts were used to enhance the model's robustness and accuracy. This approach also provides diverse linguistic representations for each video category, ensuring comprehensive coverage and minimizing bias. As shown in Table 5, the first five sentences (Index 0-4) describe general videos, capturing standard, unaltered video content in unique phrasings. The last five sentences (Index 5-9) describe time-lapse videos, characterized by accelerated playback or condensed time sequences, also phrased in various ways to capture different nuances. When calculating the MTScore, the video retrieval model uses these texts to evaluate each frame of the video, assigning probabilities based on the matches. The final result is obtained by summing the general probability and the metamorphic probability. For GPT4o-MTScore, we used a five-point rating scale and provided detailed scoring guidelines in the prompt, as shown in Table 6.

Table 5: **Retrieval sentences for coarse-grained score (MTScore)**

Index	Sentence
1	A conventional video, not a time-condensed video.
2	A usual video, not an accelerated video sequence.
3	A normal video, not a time-lapse video.
4	A standard video, not a time-lapse.
5	An ordinary video, different from a fast-motion video.
6	A time-lapse video, distinct from a regular recording.
7	A time-lapse footage, not your typical video.
8	A fast-motion video, unlike a standard video.
9	A time-condensed video, not a conventional video.
10	An accelerated video sequence, not a usual video.

Table 6: **Scoring Criteria for GPT4o-MTScore.** We set guidelines for each score to ensure that GPT-4o makes choices based on consistent criteria.

Score	Brief Reasoning Statement
1	Minimal change. The scene appears almost like a still image, with static elements remaining motionless and only minor changes in lighting or subtle movements of elements. No significant activity is noticeable.
2	Slight change. There is a small amount of movement or change in the elements of the scene, such as a few people or vehicles moving and minor changes in light or shadows. The overall variation is still minimal, with changes mostly being quantitative.
3	Moderate change. Multiple elements in the scene undergo changes, but the overall pace is slow. This includes gradual changes in daylight, moving clouds, growing plants, or occasional vehicle and pedestrian movements. The scene begins to show a transition from quantitative to qualitative change.
4	Significant change. The elements in the scene show obvious dynamic changes with a higher speed and frequency of variation. This includes noticeable changes in city traffic, crowd activities, or significant weather transitions. The scene displays a mix of quantitative and qualitative changes.
5	Dramatic change. Elements in the scene undergo continuous and rapid significant changes, creating a very rich visual effect. This includes events like sunrise and sunset, construction of buildings, and seasonal changes, making the variation process vivid and impactful. The scene exhibits clear qualitative change.

B.2 Further Description of Temporal Coherence Score

We present a detailed description of the algorithm for computing the Temporal Coherence Score. Specifically, we first process input video using the pre-trained model with grid size G and threshold T to get visibility of point p_{vis}. Then, we count the number of missing tracking points $m[i]$ in each frame, and the change in missed points between consecutive frames $\Delta m[i]$:

$$m[i] \leftarrow \frac{1}{N}\sum_{j=1}^{N}(1 - p_{\text{vis}}[i,j]) \tag{2}$$

$$\Delta m[i] \leftarrow |m[i+1] - m[i]| \tag{3}$$

where $N = G \times G$, i represents the position of the frame, j identifies different tracking points, and $p_{\text{vis}}[i, j]$ indicates the visibility of point j in frame i. To make the CHScore robust to temporally coherent disappearance of points, we first calculate the direction of camera/object movement based on the tracking points across all frames. Then, if the tracking point j of frame i disappears in the far direction, it is not included in the calculation of $m[i]$. Based on these, we then calculate the R_{missed}, which represents the average proportion of missed points per frame in the video. And the V_{missed}, which measures the variation in the number of missed points between consecutive frames, indicating frame-to-frame coherence:

$$R_{\text{missed}} = \frac{1}{F}\sum_{i=1}^{F} m[i] \tag{4}$$

$$V_{\text{missed}} = \sqrt{\frac{1}{F-1}\sum_{i=1}^{F-1}(\Delta m[i] - \bar{\Delta m})^2} \tag{5}$$

where $\Delta m[i] = m[i+1] - m[i]$, $\bar{\Delta m}$ is the mean of $\Delta m[i]$, F is the total number of frames and N is the number of points per frame. In addition, we need to calculate the R_{cut}, which indicates the ratio of frames that need to be cut to the total number of frames, reflecting the extent of video editing required. And the C_{missed}, which indicates the number of consecutive changes in missed points exceeding the threshold, indicating frequent large-scale instability in point tracking:

$$R_{\text{cut}} = \frac{|\{i : \Delta m[i] > T\}|}{F} \tag{6}$$

$$C_{\text{missed}} = \sum_{\substack{i=1 \\ \Delta m[i] > T}}^{F-1} \Delta m[i] \tag{7}$$

where T is the threshold for significant missed point variation, and $|\{i : \Delta m[i] > T\}|$ represents the number of frames with significant missed point variation. Then we calculate the M_{missed}, which measures the maximum continuous change in missed points, reflecting the most severe continuity breaks in the video, and finally get the Coherence Score (CHScore):

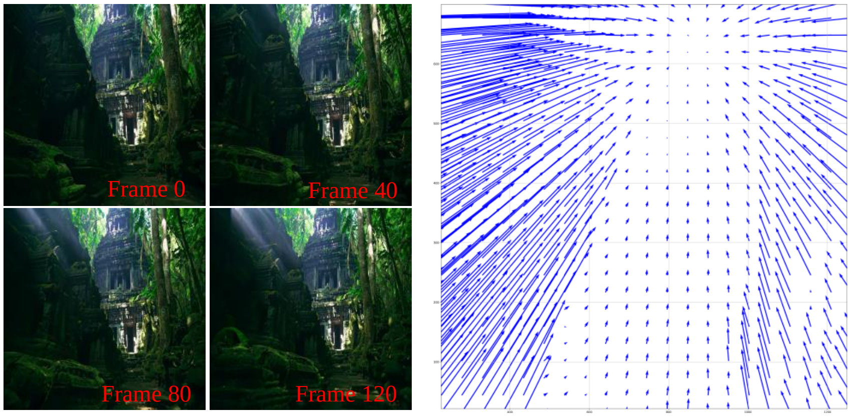
$$M_{\text{missed}} = \max(\Delta m) \tag{8}$$

$$\text{CHScore} = \frac{1}{\lambda_1 \hat{R}_{\text{missed}} + \lambda_2 \hat{V}_{\text{missed}} + \lambda_3 \hat{R}_{\text{cut}} + \lambda_4 \hat{C}_{\text{missed}} + \lambda_5 \hat{M}_{\text{missed}}} \tag{9}$$

where $\hat{X}$ represents the normalized variable, and λ_x denotes the corresponding weight coefficient. For the setting of λ_1 to λ_5, we follow the following principles. Specifically, R_{missed} is a global metric representing the model's overall performance across the entire video and holds the highest significance, with a weight of $\lambda_1 = 0.35$. V_{missed} measures the stability of missed points between frames, a critical aspect of video analysis, and is therefore assigned a weight of $\lambda_2 = 0.25$. R_{cut} indicates abnormal situations and carries a weight of $\lambda_3 = 0.15$. C_{missed}, similar in function to the R_{cut}, serves as a secondary indicator, also weighted at $\lambda_4 = 0.15$. Lastly, M_{missed} represents individual extreme cases and is assigned a lower weight of $\lambda_5 = 0.10$.

B.3 Details of Temporally Coherent Disappearance of Points

The 'Temporally coherent disappearance of points' describes the phenomenon where tracking points vanish over time due to movements such as camera movement or water flow, potentially causing these points to exit the camera's field of view. To prevent this from influencing the CHScore calculation, we initially identify the direction of change for various points within the video, as depicted in Figure 7. Subsequently, points that vanish proximate to this directional endpoint are excluded from the CHScore calculation.



(a) Video

(b) Points Direction

Figure 7: **The movement direction of the tracking points in the video.**

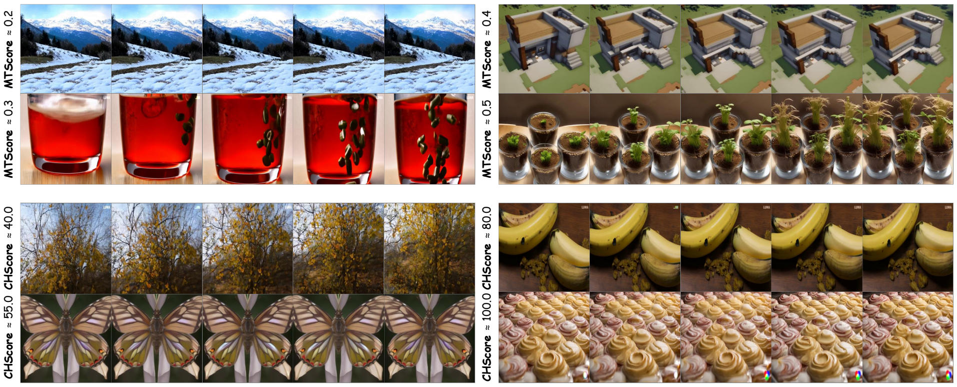


Figure 8: **Visual Reference for Varying Scores of MTScore and CHScore**. It is observed that higher scores correlate with increased metamorphic amplitude and coherence.

B.4 Visual Reference of the Different Scores of MTScore and CHScore

We also provide some samples of different scoring magnitudes for MTScore and CHScore, as shown in Figure 8. It can be seen that both scores are consistent with human perception. We strongly recommend checking out the Project Page, which provides more case studies on the metrics.

C More details about ChronoMaigc-Pro

C.1 Data Preprocessing

Due to the abundance of low-quality videos on video platforms, we filter out lower-quality videos based on metadata such as view counts, comments, and likes after acquiring the original videos, ultimately obtaining 66,226 original videos. Additionally, since our training data is sourced from video platforms (e.g., YouTube) where videos are designed to engage the audience, they inherently contain many transitions (significant changes in content during video playback). To address this issue, we follow the method described in Panda70M [16] to split the videos into multiple semantically consistent single-scene clips. Specifically, OpenCV [8] initially splits the video by analyzing pixel differences between adjacent frames. Let I_t be the image frame at time t; the difference between two adjacent frames can be computed as:

$$D_t = \sum_{i=1}^{H} \sum_{j=1}^{W} |I_t(i,j) - I_{t+1}(i,j)| \tag{10}$$

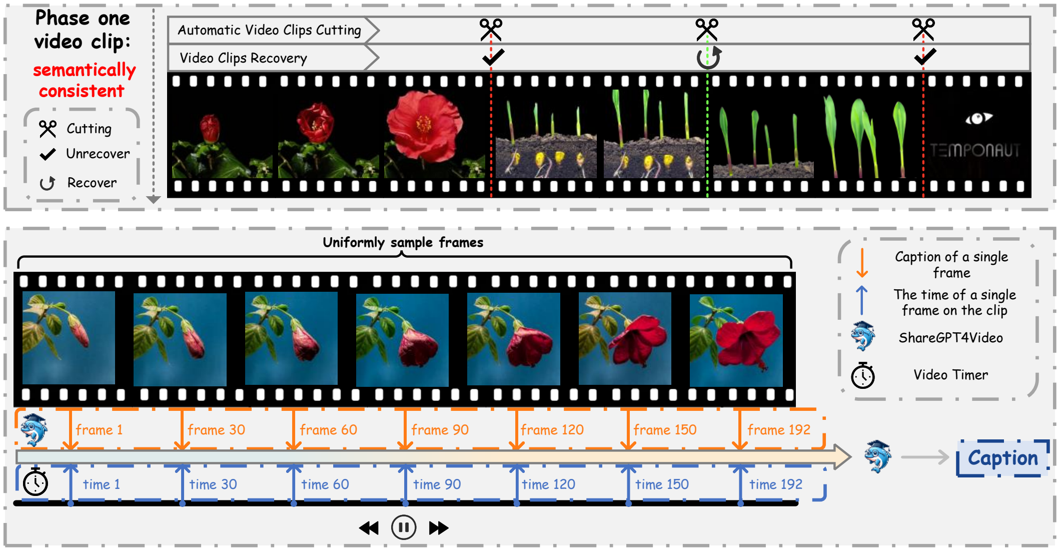


Figure 9: **The pipeline of constructing ChronoMagic-Pro.** ***(Top)*** We first use OpenCV [8] and ImageBind [22] to split the video and get semantically consistent single-scene video clips. ***(Bottom)*** Then, uniformly sample N frames and obtain captions for each using ShareGPT4Video [13]. And finally let ShareGPT4Video [13] summarize the video caption based on these captions and their frame positions.

where H and W are the height and width of the frame, and i and j represent pixel positions, respectively. Videos are split into clips where D_t exceeds a certain threshold τ. Then, the ImageBind model [22] recombines erroneously split clips by analyzing feature space differences between adjacent clips. Let $\phi(I_t)$ represent the feature vector of frame I_t obtained from the ImageBind model. The feature space difference between adjacent clips C_i and C_{i+1} can be computed as:

$$F_i = \left\|\phi(I_{t_i}) - \phi(I_{t_{i+1}})\right\|_2 \tag{11}$$

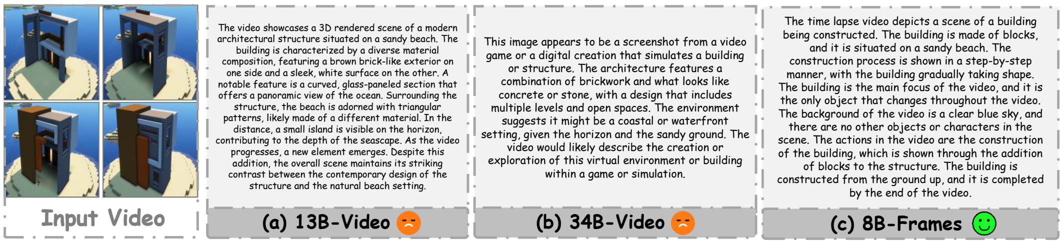
where t_i and t_{i+1} are the times of the last frame of C_i and the first frame of C_{i+1}, respectively. Clips are recombined where F_i is below a certain threshold η. This process results in semantically consistent single-scene video clips.

C.2 Time-Aware Annotation

After obtaining high-quality time-lapse video clips, it is crucial to add appropriate captions. The simplest approach is to input the video clips into a large multimodal model to generate text descriptions of the video content. However, our experiments found that the 8B [44], 13B [84], and 34B [38] models could not accurately describe the content of time-lapse videos, resulting in severe hallucinations, as shown in Figure 10. Therefore, we decided to follow the annotation strategy of MagicTime [88]. Unlike MagicTime, due to higher costs, we adopted an open-source model [13] instead of the closed-source GPT-4V [1]. As shown in Figure 10, we first uniformly sample N frames from each video segment, input these N frames into the multimodal large model to describe the content, and finally have the model summarize the final video captions based on the textual descriptions of N frames and the corresponding position of each frame in the video. To balance cost and effectiveness, we chose to use the 8B multimodal large model [13] instead of the 34B.

C.3 Distribution of the Generated Captions

To analyze the word distribution in our generated captions within ChronoMagic-Pro, we computed their frequency distributions. The results, shown in Figure 11, reveal a prevalence of terms related to time-lapse videos, including "change," "transition," and "progressing." Additionally, words from four primary categories are evident: biological (e.g., mealworm, flower, tree), human creation (e.g., building, painting, walking), meteorological (e.g., eclipse, cloud, sunrise), and physical (e.g., burning,



The video showcases a 3D rendered scene of a modern architectural structure situated on a sandy beach. The building is characterized by a diverse material composition, featuring a brown brick-like exterior on one side and a sleek, white surface on the other. A notable feature is a curved, glass-paneled section that offers a panoramic view of the ocean. Surrounding the structure, the beach is adorned with triangular patterns, likely made of a different material. In the distance, a small island is visible on the horizon, contributing to the depth of the seascape. As the video progresses, a new element emerges. Despite this addition, the overall scene maintains its striking contrast between the contemporary design of the structure and the natural beach setting.

(a) 13B-Video ☹

This image appears to be a screenshot from a video game or a digital creation that simulates a building or structure. The architecture features a combination of brickwork and what looks like concrete or stone, with a design that includes multiple levels and open spaces. The environment suggests it might be a coastal or waterfront setting, given the horizon and the sandy ground. The video would likely describe the creation or exploration of this virtual environment or building within a game or simulation.

(b) 34B-Video ☹

The time lapse video depicts a scene of a building being constructed. The building is made of blocks, and it is situated on a sandy beach. The construction process is shown in a step-by-step manner, with the building gradually taking shape. The building is the main focus of the video, and it is the only object that changes throughout the video. The background of the video is a clear blue sky, and there are no other objects or characters in the scene. The actions in the video are the construction of the building, which is shown through the addition of blocks to the structure. The building is constructed from the ground up, and it is completed by the end of the video.

(c) 8B-Frames ☺

Figure 10: **Ablation on different Captioning method.** Directly inputting the video into the model and having it describe the content is less effective than inputting keyframes into it.

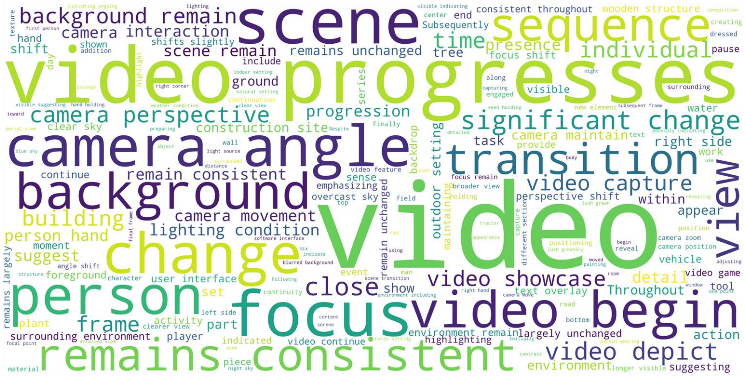


Figure 11: **The word clouds of the generated captions of ChronoMagic-Pro.** The dataset focuses on changes (gradually, progressing, increasing, etc.), processes spanning a large amount of time, such as flower blooming, ice melting, building construction, sunrise and sunset.

explosion). These terms underscore ChronoMagic-Pro's focus on large-scale metamorphic changes, persistent transformations, and substantial physical interactions.

C.4 Samples of the ChronoMagic-Pro

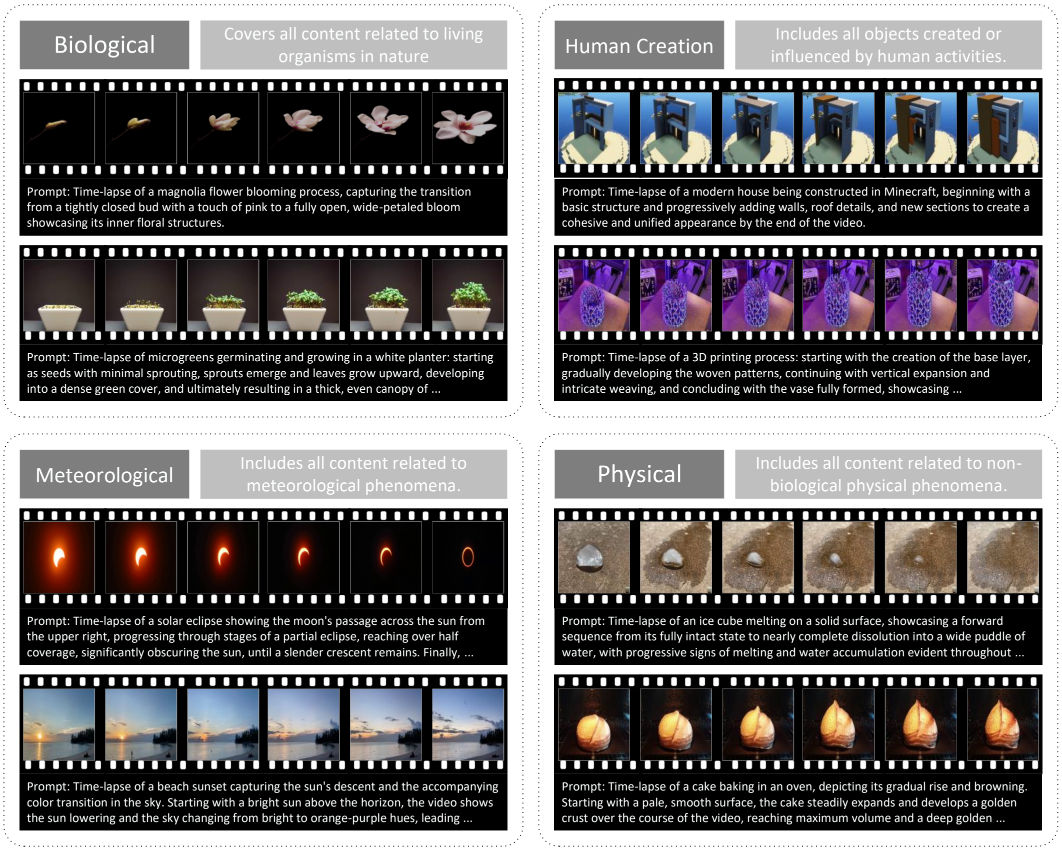
Figure 12 showcases a diverse array of samples from the ChronoMagic-Pro dataset, which features an extensive collection of time-lapse videos across several categories, including plants, buildings, ice, food, and various other objects and phenomena. Each video captures dynamic changes over time, providing rich visual information that surpasses the physical knowledge contained in many existing Text-to-Video (T2V) datasets. These samples illustrate the dataset's diversity and depth, encompassing biological, human creation, meteorological, and physical categories, designed to support advanced research in high-dynamic text-to-video generation and related fields. Additionally, the dataset includes both time-lapse videos with significant state changes (e.g., flowers blooming) and videos with smaller state changes (e.g., clouds floating).

C.5 Additional Statements

1. The aesthetic detector exhibits inherent biases, favoring artistic images, such as oil paintings and other art forms, over more realistic styles. Consequently, low aesthetic scores do not necessarily indicate poor-quality data. Retaining a small portion of such data can enhance the diversity of the videos. Thus, we include 13% of clips with low aesthetic scores in ChronoMagic-Pro.

2. There are two types of time-lapse videos: compressed and uncompressed. The former represents the entire process in a few seconds, while the latter can last for several minutes or even tens of minutes. ChronoMagic consists of compressed videos, whereas ChronoMagic-Pro includes both types to increase diversity. If the 60s+ videos, which account for 27% of ChronoMagic-Pro, are excluded, the average length would be only 12.36 seconds.

3. Since the data of ChronoMagic-Bench and ChronoMagic-Pro both include videos from YouTube, we deduplicate the data using video IDs. Additionally, we employed different annotation models


Biological
Covers all content related to living organisms in nature
Human Creation
Includes all objects created or influenced by human activities.
Meteorological
Includes all content related to meteorological phenomena.
Physical
Includes all content related to non-biological physical phenomena.

Prompt: Time-lapse of a magnolia flower blooming process, capturing the transition from a tightly closed bud with a touch of pink to a fully open, wide-petaled bloom showcasing its inner floral structures.

Prompt: Time-lapse of microgreens germinating and growing in a white planter: starting as seeds with minimal sprouting, sprouts emerge and leaves grow upward, developing into a dense green cover, and ultimately resulting in a thick, even canopy of ...

Prompt: Time-lapse of a modern house being constructed in Minecraft, beginning with a basic structure and progressively adding walls, roof details, and new sections to create a cohesive and unified appearance by the end of the video.

Prompt: Time-lapse of a 3D printing process: starting with the creation of the base layer, gradually developing the woven patterns, continuing with vertical expansion and intricate weaving, and concluding with the vase fully formed, showcasing ...

Prompt: Time-lapse of a solar eclipse showing the moon's passage across the sun from the upper right, progressing through stages of a partial eclipse, reaching over half coverage, significantly obscuring the sun, until a slender crescent remains. Finally, ...

Prompt: Time-lapse of a beach sunset capturing the sun's descent and the accompanying color transition in the sky. Starting with a bright sun above the horizon, the video shows the sun lowering and the sky changing from bright to orange-purple hues, leading ...

Prompt: Time-lapse of an ice cube melting on a solid surface, showcasing a forward sequence from its fully intact state to nearly complete dissolution into a wide puddle of water, with progressive signs of melting and water accumulation evident throughout ...

Prompt: Time-lapse of a cake baking in an oven, depicting its gradual rise and browning. Starting with a pale, smooth surface, the cake steadily expands and develops a golden crust over the course of the video, reaching maximum volume and a deep golden ...

Figure 12: **Samples from the ChronoMagic-Pro dataset.** The dataset consists of time-lapse videos, which exhibit more physical knowledge than existing T2V dataset.

(e.g., GPT-4o [1], ShareGPT4Video [13]) to label the benchmark and dataset, further reducing the risk of data leakage.

D More Details about Experiment

D.1 Details of Resource

We employ two types of GPUs: Nvidia H100 (x8) and Nvidia A800 (x8). All implementations are conducted based on the official code using the PyTorch framework.

D.2 Details of Evaluation Models

Since most T2V models do not support dynamic resolution or variable duration, it is not feasible to standardize these parameters. Therefore, we follow the official popular settings [47, 28, 66, 80] to maintain a degree of fairness. Moreover, both MTScore and CHScore mitigate the influence of resolution by employing a resizing strategy that adjusts the shorter edge and utilizes center cropping. MTScore further employs a fixed frame extraction method to ensure a consistent frame count, while the different terms of CHScore are insensitive to *num_frames*, thereby mitigating discrepancies due to varying frame numbers.

ModelScopeT2V. *Model Details.* ModelScopeT2V [73], featuring a U-Net architecture, extends the T2I model Stable Diffusion [62] by incorporating 1D temporal convolution and attention modules alongside the 2D modules for video modeling. Its training data consists primarily of image-text pairs (LAION [64]) and general video-text pairs (WebVid-10M [2] and MSR-VTT [83]), but it does not include the time-lapse videos discussed in this paper. *Implementation Setups.* We utilized the ModelScopeT2V code and model officially released on HuggingFace, maintaining the original parameter settings. We used a spatial resolution of 256×256 and a frame rate of 8 fps to generate a 2-second (16-frame) video.

ZeroScope. *Model Details.* ZeroScope [69] is a watermark-free U-Net-based video model built on ModelScopeT2V [73], capable of generating high-quality 16:9 compositions and smooth video outputs. The model is trained on 9,923 clips and 29,769 labeled frames (24 frames per clip, 576×320 resolution) derived from the original weights of ModelScopeT2V [73]. The official documentation does not specify the exact training data; we speculate that time-lapse videos were not included. *Implementation Setups.* We utilized the ZeroScope_v2_576w code and model officially released on HuggingFace, maintaining the original parameter settings. We used a spatial resolution of 576×320 and a frame rate of 8 fps to generate a 3-second (24-frame) video.

T2V-Zero. *Model Details.* Text2Video-Zero [30], featuring a U-Net architecture, is a zero-shot video generation method based on the T2I model Stable Diffusion [62]. It generates latent codes for all frames using rich motion dynamics and utilizes a self-attention mechanism to enable all frames to interact with the latent codes of the first frame. This process ultimately achieves high spatial and temporal consistency in the video through denoising. It does not require training data and, therefore, does not use time-lapse videos as training data. *Implementation Setups.* We utilized the officially released Text2Video-Zero code and model, maintaining the original parameter settings. Specifically, we used the dreamlike-photoreal-2.0 version of Stable Diffusion [62], with a spatial resolution of 512×512 and a frame rate of 8 fps, to generate a 2-second (16-frame) video.

LaVie. *Model Details. Model Details.* LaVie [76], featuring a U-Net architecture, is an extension of the T2I model Stable Diffusion [62]. It converts the T2I model into a T2V model by adding temporal dimension attention after the spatial modules and adopting an image-video joint training strategy. Its training data primarily consists of image-text pairs (LAION [64]) and general video-text pairs (WebVid-10M [2] and Vimeo25M [76]), but it does not include the time-lapse videos discussed in this paper. *Implementation Setups.* We used the officially released LaVie code and model. Although LaVie [76] provides options for frame interpolation and super-resolution after video generation, we did not use them to maintain fairness. We followed the original parameter settings, using a spatial resolution of 512×320 and a frame rate of 8 fps, to generate a 2-second (16-frame) video.

AnimateDiff. *Model Details.* AnimateDiff [23], featuring a U-Net architecture, is an extension of the T2I model Stable Diffusion [62]. It attaches a newly initialized motion modeling module to a frozen text-to-image model, then trains it on video clips to extract reasonable motion priors for video generation. Its training data primarily consists of general video-text pairs (WebVid-10M [2]), excluding the time-lapse videos discussed in this paper. *Implementation Setups.* We used the officially released AnimateDiffV3 code and model, maintaining the original parameter settings. We used a spatial resolution of 384×256 and a frame rate of 8 fps to generate a 2-second (16-frame) video.

MCM. *Model Details.* MCM [89], featuring a U-Net architecture, is a distillation video generation method based on the T2I model Stable Diffusion [62]. It propose motion consistency models (MCM) to improve video diffusion distillation by disentangling motion and appearance learning, addressing frame quality issues and training-inference discrepancies. Its training data primarily includes image-text pairs (LAION-aes [64]) and general video-text pairs (WebVid-2M [2]), but it does not include the time-lapse videos discussed in this paper. *Implementation Setups.* We used the officially released MCM-modelscopet2v-laion code and model, maintaining the original parameter settings. We used a spatial resolution of 256×256 and a frame rate of 7 fps to generate a 2-second (14-frame) video.

MagicTime. *Model Details.* MagicTime [88] is a U-Net-based metamorphic video generation model built on AnimateDiff [23]. It is capable of generating time-lapse videos with significant time spans and pronounced state changes, such as the entire process of a seed blooming or building construction. The model is trained using 2,265 metamorphic (time-lapse) clips and the original weights from AnimateDiffV3 [23]. Its training data primarily includes ChronoMagic [88], making it the only existing T2V model that uses time-lapse videos in the training process. *Implementation Setups.* We used the officially released MagicTime code and model, maintaining the original parameter settings. We used a spatial resolution of 512×512 and a frame rate of 8 fps to generate a 2-second (16-frame) video.

VideoCrafter2. *Model Details.* VideoCrafter2 [11], featuring a U-Net architecture, is similar to AnimateDiff [23], as both add temporal modules to Stable Diffusion [62] to achieve video generation. However, VideoCrafter2 differs by encoding fps as a condition into the model and implementing the I2V function. Its training data primarily includes image-text pairs (LAION-COCO [17], JDB [70]) and general video-text pairs (WebVid-10M [2]), but it does not include the time-lapse videos discussed in this paper. *Implementation Setups.* We used the officially released VideoCrafter2 code and model, maintaining the original parameter settings. We used a spatial resolution of 512×320 and a frame rate of 10 fps to generate a 2-second (20-frame) video.

Latte. *Model Details.* Latte [49] is a pioneer in open-source DiT-based T2V algorithms. It inherits the pure Transformer architecture of the T2I algorithm PixArt-α [12] and extends it by adding temporal modules after each spatial module, training from the original weights of PixArt-α [12] to achieve a DiT-based T2V algorithm. Its training data primarily includes general video-text pairs (Vimeo25M [76] and WebVid-10M [2]). Although it includes the time-lapse videos mentioned in this paper, they primarily consist of sky videos with fewer physical priors, making it unable to generate videos such as seed germination and flower blooming. *Implementation Setup.* We used the officially released LatteT2V code and model, maintaining the original parameter settings. We used a spatial resolution of 512×512 and a frame rate of 8 fps to generate a 2-second (16-frame) video.

OpenSoraPlan v1.1. *Model Details.* OpenSoraPlan v1.1 [43] is a high-quality video generation model based on Latte [49]. It replaces the Image VAE [31] with Video VAE (CausalVideoVAE [43]), similar to Sora [9], enabling the generation of videos up to approximately 21 seconds long and high-quality images. Its training data consists of videos and images scraped from open-source websites under the CC0 license, labeled using ShareGPT4Video [13] to create a high-quality self-built dataset. The official documentation does not specify the exact training data; we speculate that time-lapse videos were not used. *Implementation Setup.* We used the officially released OpenSoraPlan v1.1 code and model. Although it provides T2V models in three versions: 65 frames, 221 frames, and 513 frames, we chose the 65-frame version to ensure fairness by maintaining a similar video length to other models. We kept the original parameter settings, using a spatial resolution of 512×512 and a frame rate of 24 fps to generate a 3-second (65-frame) video.

OpenSora 1.1 & 1.2. *Model Details.* OpenSora 1.1 & 1.2 [95] is a high-quality DiT-based T2V model that introduces the ST-DiT-2 architecture, building on Latte [49]the former is based on the Diffusion Model and the latter is based on the Flow Model. It supports the generation of images or videos with any aspect ratio, different resolutions, and durations. Its training data consists of images and videos scraped from open-source websites and a labeled self-built dataset. The official documentation does not specify the exact training data; we speculate that time-lapse videos were not used. *Implementation Setup.* We used the officially released OpenSora 1.1 & 1.2 code and model. For OpenSora 1.1, we employed the stage-3 checkpoint, setting the spatial resolution to 512×512 and the frame rate to 24 fps, to generate a 2-second (48-frame) video. For OpenSora 1.2, we set the spatial resolution to 1280×720 and the frame rate to 24 fps, producing a 4-second (96-frame) video.

CogVideoX *Model Details.* CogVideoX [86] is a state-of-the-art text-to-video diffusion model that builds upon the success of large-scale DiT models. To enhance text-video alignment, CogVideoX utilizes an expert transformer with expert adaptive LayerNorm, facilitating deep fusion between modalities. The model implements 3D full attention to comprehensively model videos along both temporal and spatial dimensions, ensuring temporal consistency and capturing large-scale motions. Its training data consists of scraped videos and images, and custom refined Panda70M [16], COCO caption [45] and WebVid [2]. *Implementation Setup.* We use the officially released CogVideoX code and model. For our experiments, we set the spatial resolution to 720x480, generated 48 frames, and used a frame rate of 8 fps, resulting in a 6-second video.

EasyAnimate *Model Details.* EasyAnimate [82] is an advanced text-to-video generation model designed to create high-quality animated videos from textual prompts. It adopts U-ViT [3] architectures and slice-vae to avoid unstable training. Its training data consists of videos and images scraped from open-source websites, and open-source dataset 10M SAM [32] and 2M JourneyDB [70]. The official documentation does not specify that time-lapse videos were used. *Implementation Setup.* We utilized the officially released EasyAnimateV3 code and model. For our experiments, we used the 720P version of the model. As per the default setting, we set the spatial resolution to 1008x576, generated 96 frames, and used a frame rate of 24 fps, resulting in a 4-second video.

D.3 Further Verification Experiment on ChronoMagic-Pro

Notably, after fine-tuning in ChronoMagic-Pro, the enhancement in metamorphic amplitude endowed OpenSoraPlan [43] with the ability to generate time-lapse videos of significant state changes, such as

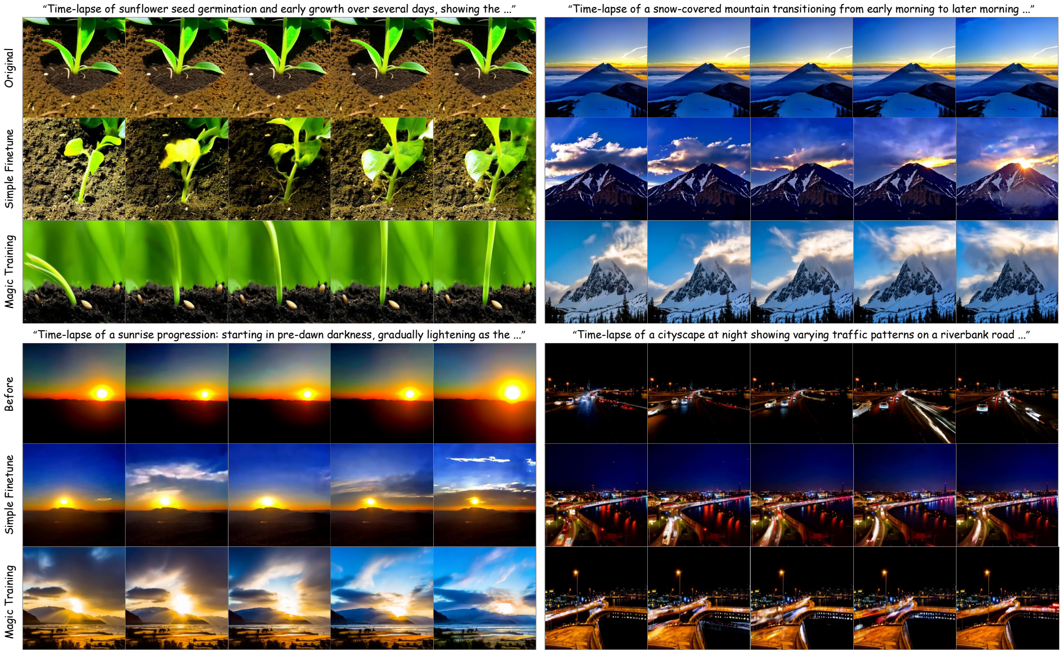

"Time-lapse of sunflower seed germination and early growth over several days, showing the ..."
"Time-lapse of a snow-covered mountain transitioning from early morning to later morning ..."
Original
Simple Finetune
Magic Training
"Time-lapse of a sunrise progression: starting in pre-dawn darkness, gradually lightening as the ..."
"Time-lapse of a cityscape at night showing varying traffic patterns on a riverbank road ..."
Before
Simple Finetune
Magic Training

Figure 13: **Qualitative comparison of OpenSoraPlan v1.1 [43] before and after fine-tuning using ChronoMagic-Pro 10K.** After fine-tuning, the changes in the generated videos are no longer limited to lighting and camera movement, but are extended to changes in the state of objects. Additionally, it ensures that the *visual quality*, *text relevance*, and *coherence* are maintained without loss. Moreover, the efficacy of simple fine-tuning is inferior to that achieved through the Magic Training Strategy[88].

blooming flowers and city traffic. We provide additional qualitative analysis, as shown in Figure 13. It is evident that, after fine-tuning, the generated videos can extend changes beyond mere lighting and camera movements to alterations in the state of objects, while ensuring that the visual quality, text relevance, and coherence remain uncompromised. This proves that ChronoMagic-Pro can support existing models in generating high-quality time-lapse videos with significant state changes, providing a new approach for future T2V model training. Moreover, our findings suggest that with appropriate fine-tuning, it is possible to correct the common tendency of video models to produce nearly static videos on arbitrary topics. This phenomenon has also been observed in MagicTime-DiT [88], despite utilizing only around 2,000 time-lapse videos. However, it is important to note that the Magic Training Strategy [88], originally designed for U-Net-based models, may not be as effective for DiT-based models. In this study, we employ this methods solely for verification experiments. Additionally, the efficacy of simple fine-tuning is inferior to that achieved through the Magic Training Strategy [88].

D.4 More Qualitative Evaluation on ChronoMagic-Bench

Due to space limitations, additional time-lapse videos generated by different baseline methods are shown in Figure 14. Similar to the results in the main text, most algorithms, except for MagicTime [88], fail to generate time-lapse videos with significant state changes, such as building construction. However, for time-lapse videos with smaller state changes, essentially faster-moving videos like city traffic changes, U-Net-based methods [73, 69, 30, 76, 23, 11, 88] exhibit much better visual quality, text relevance, and coherence compared to DiT-based methods [49, 43, 95]. This again demonstrates that U-Net-based methods are currently more stable and capable of producing satisfactory results with minimal inference. All videos generated by all models on ChronoMagic-Bench is publicly available on https://pku-yuangroup.github.io/ChronoMagic-Bench.

D.5 More Analysis of Closed-Source Models

We present and analyze the results from a qualitative perspective, as shown in Figure 15. The results are consistent with Table 4. For metamorphic amplitude, most methods can only generate simple time-lapse videos, such as traffic flow; only Dream Machine [48] can generate a moderately

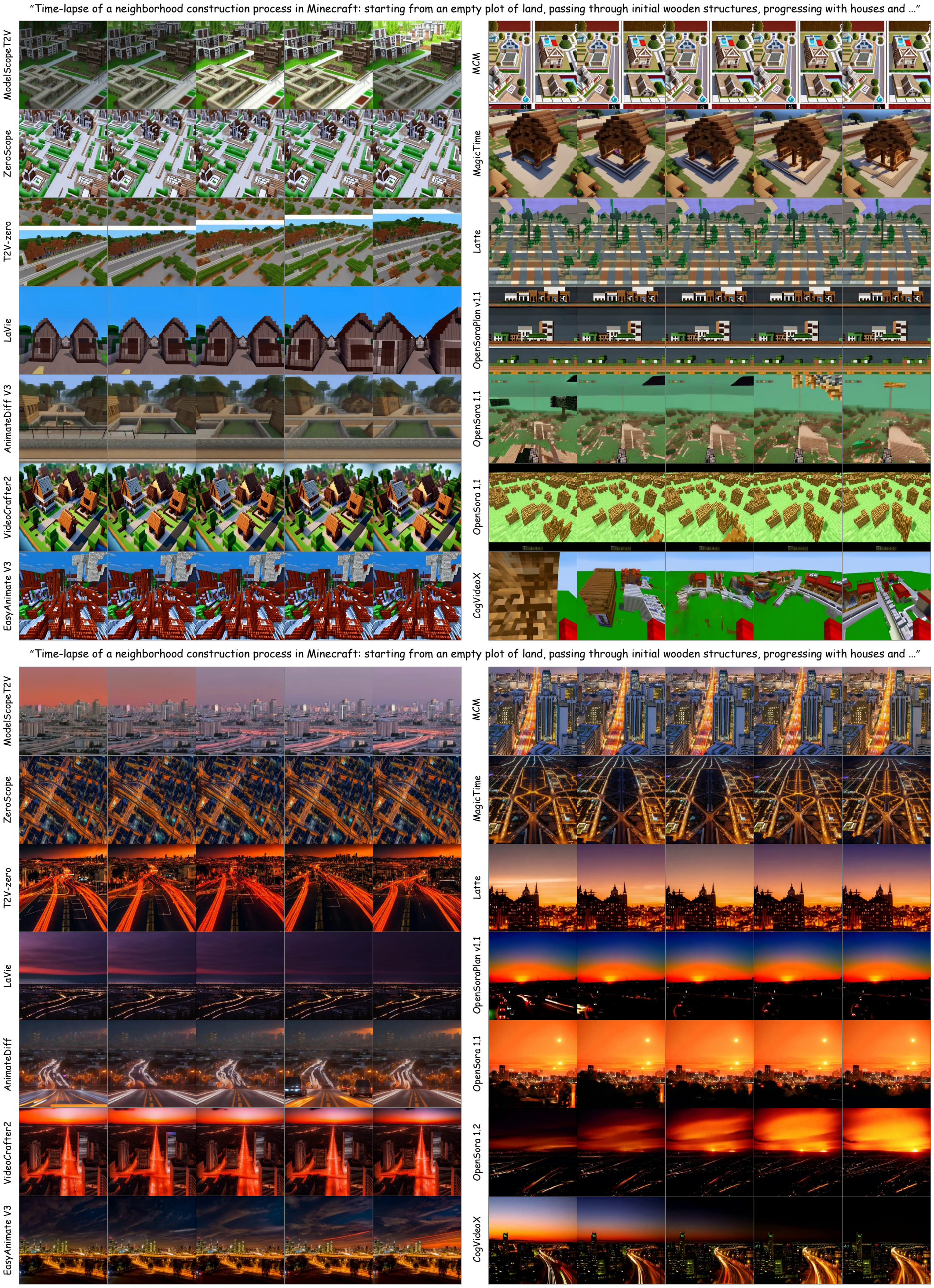

"Time-lapse of a neighborhood construction process in Minecraft: starting from an empty plot of land, passing through initial wooden structures, progressing with houses and ..."
ModelScopeT2V
ZeroScope
T2V-zero
LaVie
AnimateDiff V3
VideoCrafter2
EasyAnimate V3
MCM
MagicTime
Latte
OpenSoraPlan v1.1
OpenSora 1.1
OpenSora 1.1
CogVideoX
"Time-lapse of a neighborhood construction process in Minecraft: starting from an empty plot of land, passing through initial wooden structures, progressing with houses and ..."
ModelScopeT2V
ZeroScope
T2V-zero
LaVie
AnimateDiff
VideoCrafter2
EasyAnimate V3
MCM
MagicTime
Latte
OpenSoraPlan v1.1
OpenSora 1.1
OpenSora 1.2
CogVideoX

Figure 14: **More Qualitative Comparison with different T2V generation methods for the text-to-video task in ChronoMaigc-Bench.** Most methods struggle to follow the prompt to generate time-lapse videos with high physics prior content.

challenging full process of night-to-day transformation; no method can generate complex changes like plant growth or building construction. In terms of temporal coherence, the performance of

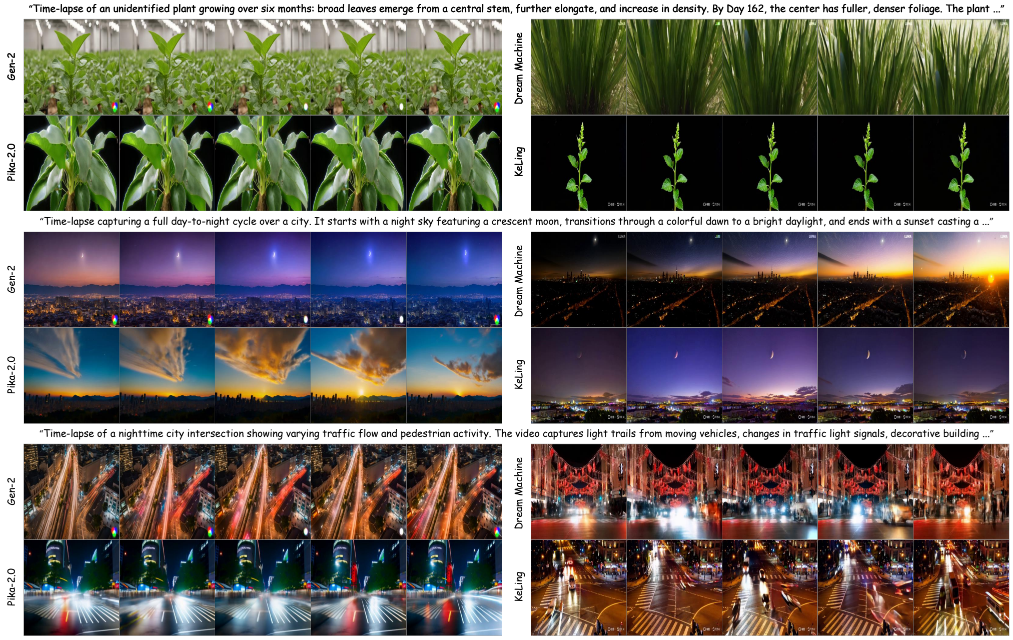

"Time-lapse of an unidentified plant growing over six months: broad leaves emerge from a central stem, further elongate, and increase in density. By Day 162, the center has fuller, denser foliage. The plant ..."
Gen-2
Dream Machine
Pika-2.0
KeLing
"Time-lapse capturing a full day-to-night cycle over a city. It starts with a night sky featuring a crescent moon, transitions through a colorful dawn to a bright daylight, and ends with a sunset casting a ..."
Gen-2
Dream Machine
Pika-2.0
KeLing
"Time-lapse of a nighttime city intersection showing varying traffic flow and pedestrian activity. The video captures light trails from moving vehicles, changes in traffic light signals, decorative building ..."
Gen-2
Dream Machine
Pika-2.0
KeLing

Figure 15: **Qualitative comparison with *Close-Source* generation methods for the text-to-video task in ChronoMaigc-Bench-150.** Most methods can only generate simple time-lapse videos such as traffic flows and starry skies, and are incapable of generating complex changes such as plant growth or building construction.

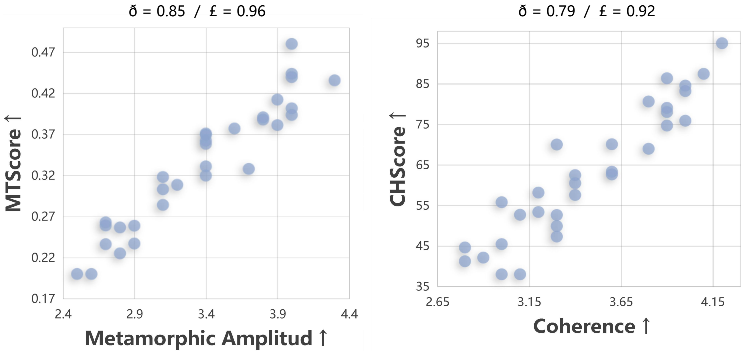
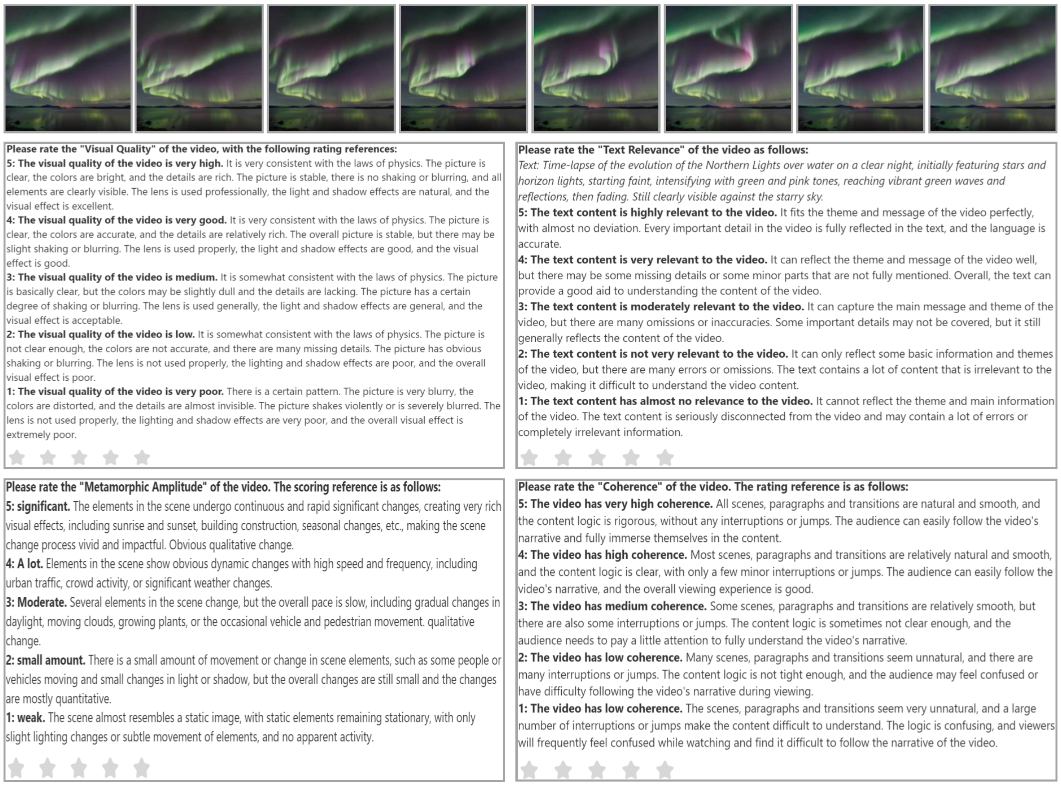


Figure 16: **Alignment between automatic metrics and human perception in terms of disaggregated data.** ð and £ represent Kendall↑ and Spearman↑ coefficients, respectively. ↑" denotes higher is better.

various closed-source models is comparable, with minor visible differences. Regarding visual quality, the DiT-based methods Dream Machine [48] and KeLing [35] outperform those based on U-Net, producing more realistic plants, more accurately saturated sky colors, and clearer traffic flow. In terms of text relevance, all methods adhere to the prompt's instructions to generate content relevant to the theme, except for Pika-1.0 [36], which mistakenly interprets day-to-night as night-to-day.

D.6 Additional Details of Human Evaluation

Pre-processing The questionnaire for human evaluators to rate the generated content was established following methodologies from prior studies [60, 72, 88, 66]. The evaluation focused on four primary aspects: *Visual Quality*, *Text Relevance*, *Metamorphic Amplitude*, and *Coherence*. For each criterion, we employed a five-point rating scale and provided scoring guidelines to ensure consistent user selections, thereby minimizing assessment bias. For detailed criteria, please refer to Figure 17. For detailed explanation of voters, the voter population predominantly comprises undergraduate, master's, and phd students from universities, along with a segment of the general public who are not



Please rate the "Visual Quality" of the video, with the following rating references:

5: The visual quality of the video is very high. It is very consistent with the laws of physics. The picture is clear, the colors are bright, and the details are rich. The picture is stable, there is no shaking or blurring, and all elements are clearly visible. The lens is used professionally, the light and shadow effects are natural, and the visual effect is excellent.

4: The visual quality of the video is very good. It is very consistent with the laws of physics. The picture is clear, the colors are accurate, and the details are relatively rich. The overall picture is stable, but there may be slight shaking or blurring. The lens is used properly, the light and shadow effects are good, and the visual effect is good.

3: The visual quality of the video is medium. It is somewhat consistent with the laws of physics. The picture is basically clear, but the colors may be slightly dull and the details are lacking. The picture has a certain degree of shaking or blurring. The lens is used generally, the light and shadow effects are general, and the visual effect is acceptable.

2: The visual quality of the video is low. It is somewhat consistent with the laws of physics. The picture is not clear enough, the colors are not accurate, and there are many missing details. The picture has obvious shaking or blurring. The lens is not used properly, the lighting and shadow effects are poor, and the overall visual effect is poor.

1: The visual quality of the video is very poor. There is a certain pattern. The picture is very blurry, the colors are distorted, and the details are almost invisible. The picture shakes violently or is severely blurred. The lens is not used properly, the lighting and shadow effects are very poor, and the overall visual effect is extremely poor.

☆ ☆ ☆ ☆ ☆

Please rate the "Text Relevance" of the video as follows:

Text: Time-lapse of the evolution of the Northern Lights over water on a clear night, initially featuring stars and horizon lights, starting faint, intensifying with green and pink tones, reaching vibrant green waves and reflections, then fading. Still clearly visible against the starry sky.

5: The text content is highly relevant to the video. It fits the theme and message of the video perfectly, with almost no deviation. Every important detail in the video is fully reflected in the text, and the language is accurate.

4: The text content is very relevant to the video. It can reflect the theme and message of the video well, but there may be some missing details or some minor parts that are not fully mentioned. Overall, the text can provide a good aid to understanding the content of the video.

3: The text content is moderately relevant to the video. It can capture the main message and theme of the video, but there are many omissions or inaccuracies. Some important details may not be covered, but it still generally reflects the content of the video.

2: The text content is not very relevant to the video. It can only reflect some basic information and themes of the video, but there are many errors or omissions. The text contains a lot of content that is irrelevant to the video, making it difficult to understand the video content.

1: The text content has almost no relevance to the video. It cannot reflect the theme and main information of the video. The text content is seriously disconnected from the video and may contain a lot of errors or completely irrelevant information.

☆ ☆ ☆ ☆ ☆

Please rate the "Metamorphic Amplitude" of the video. The scoring reference is as follows:

5: significant. The elements in the scene undergo continuous and rapid significant changes, creating very rich visual effects, including sunrise and sunset, building construction, seasonal changes, etc., making the scene change process vivid and impactful. Obvious qualitative change.

4: A lot. Elements in the scene show obvious dynamic changes with high speed and frequency, including urban traffic, crowd activity, or significant weather changes.

3: Moderate. Several elements in the scene change, but the overall pace is slow, including gradual changes in daylight, moving clouds, growing plants, or the occasional vehicle and pedestrian movement. qualitative change.

2: small amount. There is a small amount of movement or change in scene elements, such as some people or vehicles moving and small changes in light or shadow, but the overall changes are still small and the changes are mostly quantitative.

1: weak. The scene almost resembles a static image, with static elements remaining stationary, with only slight lighting changes or subtle movement of elements, and no apparent activity.

☆ ☆ ☆ ☆ ☆

Please rate the "Coherence" of the video. The rating reference is as follows:

5: The video has very high coherence. All scenes, paragraphs and transitions are natural and smooth, and the content logic is rigorous, without any interruptions or jumps. The audience can easily follow the video's narrative and fully immerse themselves in the content.

4: The video has high coherence. Most scenes, paragraphs and transitions are relatively natural and smooth, and the content logic is clear, with only a few minor interruptions or jumps. The audience can easily follow the video's narrative, and the overall viewing experience is good.

3: The video has medium coherence. Some scenes, paragraphs and transitions are relatively smooth, but there are also some interruptions or jumps. The content logic is sometimes not clear enough, and the audience needs to pay a little attention to fully understand the video's narrative.

2: The video has low coherence. Many scenes, paragraphs and transitions seem unnatural, and there are many interruptions or jumps. The content logic is not tight enough, and the audience may feel confused or have difficulty following the video's narrative during viewing.

1: The video has low coherence. The scenes, paragraphs and transitions seem very unnatural, and a large number of interruptions or jumps make the content difficult to understand. The logic is confusing, and viewers will frequently feel confused while watching and find it difficult to follow the narrative of the video.

☆ ☆ ☆ ☆ ☆

Figure 17: **Visualization of the Questionnaire for Human Evaluation.** We employ a five-point rating scale and provided scoring guidelines to ensure consistent selections by users, thereby minimizing assessment bias.

associated with this field. They come from various regions around the world, including China, USA, Singapore, etc., which ensures that the participants have universality. This composition guarantees the precision and diversity of human evaluations.

Pose-processing Given the use of a simple five-point evaluation scale, we remove outliers from the responses as follows:

- Restricted each IP address to prevent duplicates and required users to log in to their accounts before voting, ensuring that each person could only submit once.
- Determined the validity of data based on the time taken to complete the questionnaire. Given that completing a questionnaire typically takes 10 to 20 minutes, we excluded samples where the response time was less than 10 minutes.
- Randomized the order in which different videos were presented to avoid cognitive biases among voters.
- Required a sliding verification at submission to ensure that all questionnaires were completed manually and not by bots.
- Discarded any questionnaire where 50% of the ratings were extreme values, i.e., the sum of 5-point and 1-point options exceeded 50%.

Additional Evaluation In addition to the main text, Figure 6 analyzes the video metrics aggregated by the model. We also provide a human evaluation of disaggregated data (that is, where each point represents a video), which consists of 32 videos randomly selected from all the questionnaires. The results are shown in Figure 16. It can be seen that the proposed MTScore and CHScore are consistent with human perception in terms of disaggregated data.

E More details about 75 subcategories in ChronoMaigc-Bench

Due to space limitations, we provide detailed descriptions of the 75 search terms used in ChronoMagic-Bench below (*each term includes the phrase "time-lapse"*), all of which pertain to time-lapse. Because of search engine limitations, some precise search terms may not yield optimal results. Therefore, to collect search terms more comprehensively, some overlap may exist between broader terms like "plant" and precise terms like "flower".

Biological:

- *Animal.* Captures the movements, behaviors, and interactions of various animals over an extended period. This includes everything from the daily activities of pets to the complex behaviors of wild animals in their natural habitats.
- *Spider Web.* Showcases the intricate process of spiders spinning their webs. It highlights the changes the web undergoes over time.
- *Butterfly.* Focuses on the life cycle of butterflies, particularly the metamorphosis from caterpillar to chrysalis to adult butterfly. It includes the intricate process of pupation and emergence.
- *Hatching.* Documents the hatching process of various eggs, including those of birds, reptiles, and insects. This category captures the moment of emergence and the initial activities of the newborns.
- *Flower Dying.* Captures the end-of-life process of flowers, showing how they wilt and decay over time.
- *Mealworm.* Showcases the behavior of mealworms, including their feeding habits.
- *Plant Growing.* This broad category includes time-lapse videos of various plants as they grow from seeds to mature plants. It encompasses root development, stem elongation, and the emergence of leaves and flowers.
- *Ripening.* Documents the ripening process of fruits and vegetables, showing the changes in color, texture, and overall appearance as they become ready for consumption.
- *Leaves.* Focuses on the growth, movement, and changes of leaves on plants. This includes the unfolding of new leaves, changes in color, and responses to environmental factors.
- *Seed.* Captures the germination and initial growth stages of seeds, from the first signs of sprouting to the establishment of seedlings. It focuses on the early and often delicate stages of plant development.
- *Blooming.* Showcases the process of flowers blooming, capturing the gradual opening of petals and the transformation from buds to full blossoms.
- *Mushroom.* Captures the rapid growth and development of mushrooms, from the initial emergence of the mycelium to the full development of the fruiting body.

Human Creation:

- *3D Printing.* Captures the process of 3D printing objects. These videos show the additive manufacturing process layer by layer, from the initial base to the final, complete object.
- *Painting.* Showcases the process of creating a painting, from the initial sketch to the final strokes.
- *Laser Engraving.* Show the process of laser engraving on various materials, such as the process of pattern formation.
- *Building.* Documents the construction of various structures, including residential, commercial, and industrial buildings. This category highlights the step-by-step development from foundation to completion.
- *Minecraft Build.* Captures the construction of complex structures and landscapes within the game Minecraft.
- *Demolition.* Captures the process of demolishing buildings and structures.

- *Fireworks.* Captures the display of fireworks, showcasing the entire process from the launch of the explosive into the sky to its transformation into bursts of color and patterns in the night sky.
- *People.* Focuses on the activities and movements of people in various settings, including streets, parks, and public spaces.
- *Sport.* Captures sporting events and activities, highlighting the movement of athletes, the progression of games, and the energy of the crowd.
- *City.* Focuses on the dynamic activities within a city, including urban development, traffic flow, and daily life. These videos often showcase the bustling and ever-changing nature of urban environments.
- *Factory.* Highlights the operations within a factory, including assembly lines, manufacturing processes, and the movement of goods.
- *Market.* Documents the activities within a market, including the setting up of stalls, movement of people, and trading of goods.
- *Office.* Captures the daily activities within an office environment, including the ebb and flow of workers, meetings, and the general hustle and bustle of office life.
- *Restaurant.* Documents the activities within a restaurant, including food preparation, service, and customer interactions.
- *Road.* Capture the traffic flow, and changes in road conditions over time.
- *Station.* Focuses on the activities within transportation stations, such as train stations, bus terminals, and airports. These videos capture the flow of passengers, arrivals, departures, and the hustle and bustle of travel hubs..
- *Traffic.* Captures the movement of vehicles on roads and highways, including the traffic flow, congestion, and the changing pace of vehicular movement throughout the day.
- *Walking.* Focuses on people walking in various environments, such as city streets, parks, and malls.
- *Parking.* Captures the movement of vehicles in parking lots or garages, including the flow of cars as they enter, park, and exit.

Meteorological:

- *Day to Night.* Show the transitions from daylight to nighttime, capturing the gradual shift in light and atmosphere as day turns to night.
- *Night to Day.* Shows the transitions from nighttime to daylight, showing the gradual change in lighting and environment as night turns to day.
- *Day.* Captures the progression of daylight hours, highlighting changes in light intensity, shadows, and weather conditions.
- *Night.* Shows the sequences of nighttime scenes, often capturing the movement of stars, phases of the moon, and nocturnal activities.
- *Cloud.* Shows the formation, movement, and dissipation of clouds, providing a dynamic view of the ever-changing sky.
- *Lunar Eclipse.* Shows the gradual movement of the moon through the Earth's shadow and the resulting changes in appearance during a lunar eclipse.
- *Rainbow.* Captures the formation, duration, and fading of rainbows, providing a colorful display over time.
- *Sky.* Captures a variety of atmospheric phenomena such as cloud movements, sunrises, sunsets, and weather changes over time.
- *Snowstorm.* Shows the accumulation of snow and the changing conditions during and after a snowstorm.

- *Storm.* Highlights the intensity and movement of storm clouds and lightning during various types of storms.
- *Sunrise.* Captures the gradual increase in light and the awakening of the environment during sunrise.
- *Sunset.* Showcases the beautiful colors and gradual fading of light as the day ends during sunset.
- *Aurora.* Captures the dynamic changes and movement of the Northern and Southern Lights, showcasing the evolving natural light displays over time.
- *Tide.* Illustrates the rise and fall of sea levels and their impact on coastal landscapes over time.
- *Wind.* Captures the effects of wind on landscapes, including the movement of vegetation, dust storms, and changing cloud patterns over time.
- *Seasons.* Shows the dramatic changes across different seasons, highlighting the transformation of landscapes throughout the year.
- *Nature.* Captures various natural scenes, including the growth of plants, changes in landscapes, and wildlife activity.
- *Beach.* Illustrate the changes in tides, waves, and shifting weather conditions throughout the day.
- *Desert.* Shows the dramatic changes in light, temperature, and atmosphere in desert landscapes over time.
- *Forest.* Illustrates changes in foliage, light patterns, and wildlife activity in forests throughout the day or seasons.
- *Grassland.* Highlight the subtle yet significant changes in vegetation and weather in grasslands over time.
- *Lake.* Captures reflections, water level changes, and the transformation of surrounding landscapes.
- *Mountain.* Showcases changes in light, weather, and cloud movement around mountainous peaks over time.
- *Ocean.* Highlights the continuous motion of waves, tides, and the impact of weather on ocean scenes over time.
- *Plain.* Shows the transformation of open landscapes due to changing light and weather conditions over time.
- *River.* Illustrates the flow of water, changes in water levels, and the transformation of surrounding landscapes over time.
- *Valley.* Highlights changes in light, weather, and seasonal transformations in valley areas over time.

Physical:

- *Baking.* Shows the transformation of dough or batter as it rises and turns into baked goods, highlighting changes in color, texture, and volume over time.
- *Cooking.* Shows the various stages of food preparation and cooking, highlighting changes in texture, color, and form.
- *Candle Burning.* Illustrates the gradual melting and burning of a candle, including changes in the wax and the flickering flame.
- *Tea Diffusing.* Illustrates how tea leaves release their color and flavor into hot water, showing the gradual diffusion process and changes in the liquid.
- *Corrosion.* Captures the slow process of materials deteriorating due to chemical reactions with their environment, often resulting in rust or other forms of decay.

- *Decompose.* Shows organic materials breaking down over time, illustrating the process of decomposition and the changes in form and structure.
- *Fruit Rotting.* Illustrates the gradual decay and breakdown of fruit, showing changes in color, texture, and structure as it rots.
- *Explosion.* Captures the rapid and dramatic release of energy, showing the sudden change in materials and the environment.
- *Burning.* Captures the process of combustion, showing how materials ignite, burn, and reduce to ash or other residues.
- *Gasification.* Shows the process of a solid or liquid turning into gas, highlighting the changes in state and movement of particles.
- *Ice Melting.* Captures the transition of ice from solid to liquid, showing the gradual melting process and changes in shape and volume.
- *Ink Diffusing.* Illustrates how ink spreads and disperses in a liquid, showing the dynamic patterns and changes in concentration over time.
- *Melting.* Shows the process of a solid turning into a liquid, highlighting changes in form and consistency as the material melts.
- *Rusting.* Captures the slow formation of rust on metal surfaces, showing the chemical changes and resulting texture and color changes.
- *Water Freezing.* Shows the transition of water from liquid to solid, capturing the formation of ice and changes in volume and structure.

F Ethics Statement

Potential Harms Caused by the Research Process. The video data utilized by ChronoMagic-Bench is sourced from free content available on four platforms: Pexels (CC0), MixKit (CC0), PixaBay (CC0), and YouTube (CC BY 4.0). Conversely, ChronoMagic-Pro exclusively employs videos from YouTube (CC BY 4.0). The licensing types of these videos are clearly indicated on their respective platforms. The CC0 license (Creative Commons Zero) designates content as public domain, allowing unrestricted use without the need for additional permissions or licenses. Videos from the YouTube platform adhere to the CC BY 4.0 license (Creative Commons Attribution 4.0); consequently, we have included video IDs and author information in the metadata to prevent any potential contractual disputes. The video content consists entirely of time-lapse footage, and we detect and discard NSFW content based on the video caption. For videos involving identifiable individuals, we accelerate the blurring process to ensure the security of personally identifiable information. The collected videos are organized into four major categories (comprising 75 subcategories), with contributors hailing from various countries and regions worldwide. This diversity ensures that ChronoMagic-Bench and ChronoMagic-Pro possess ample representativeness. The Open-Sora-Plan model [43], fine-tuned using our dataset, exhibited no significant content bias.

Data collection was facilitated by the dedicated efforts of numerous contributors, including the authors of this paper and those who participated in the human evaluation. We regard an individual's hourly wage or compensation as personal information, which, due to privacy considerations, cannot be disclosed. Nonetheless, we can confirm that all participants received appropriate compensation in compliance with the legal requirements of their respective countries or regions. The privacy information of all participants is protected, so there is no additional risk to the them.

Societal Impact and Potential Harmful Consequences. The objective of ChronoMagic-Bench is to identify the limitations of current text-to-video generation models in producing time-lapse videos and to develop the ChronoMagic-Pro dataset to advance the field. Although time-lapse video generation models offer substantial potential to support and enhance human creativity, it is crucial to consider broader societal implications during their development:

First and foremost, environmental issues cannot be overlooked. As text-to-video generation technology advances, the demand for computational resources escalates. Large-scale data processing, model testing, and training generally depend on energy-intensive data centers, which significantly contribute to carbon emissions. For instance, this study utilized the energy-intensive H100 for experiments. If not addressed, the widespread adoption of this technology could further exacerbate climate change. Consequently, researchers and developers should focus on optimizing algorithms to reduce energy consumption.

Secondly, the generation of false content has raised significant social and ethical concerns. After appropriate fine-tuning using the ChronoMagic-Pro/ProH dataset, text-to-video generation models are capable of producing not only metamorphic videos with extended time spans and high levels of realism but also high-quality general videos. These generative models could be misused to create deceptive videos, potentially misleading the public or disseminating misinformation, particularly on fast-paced and widely influential platforms such as social media. To prevent such misuse, it is crucial to consider the implementation of content authenticity verification mechanisms and the establishment of robust legal and ethical frameworks during the development and deployment of these technologies.

Lastly, the issue of dataset bias may result in skewed and inequitable outcomes in model generation. Although the video content in the ChronoMagic-Bench and ChronoMagic-Pro datasets is sourced globally, the captions are exclusively in English. This single-language choice may impair the model's ability to accurately interpret and generate videos across diverse cultural contexts and non-English language environments. Moreover, this bias could exacerbate existing language inequalities by disregarding the needs of non-English-speaking users. Therefore, future dataset construction should incorporate multilingual support to ensure broader adaptability and fairness in models on a global scale.

Impact Mitigation Measures. We take full responsibility for the licensing, distribution, and maintenance of our ChronoMaigc-Bench and ChronoMagic-Pro/ProH. Our datasets and benchmark are released under a CC-BY-4.0 license, and our code under an Apache license. We have clearly stated on our homepage that all data is for academic research only to prevent misuse or improper use. And we provide the email address for YouTube authors to contact and remove invalid videos in time. All the metadata are hosted on GitHub and HuggingFace at the following URLs: https://github.com/PKU-YuanGroup/ChronoMagic-Bench and https://huggingface.co/collections/BestWishYsh.