ChronoMagic-Bench: A Benchmark for Metamorphic Evaluation ✨ Spotlight ✨ of Time-lapse Text-to-Video Generation

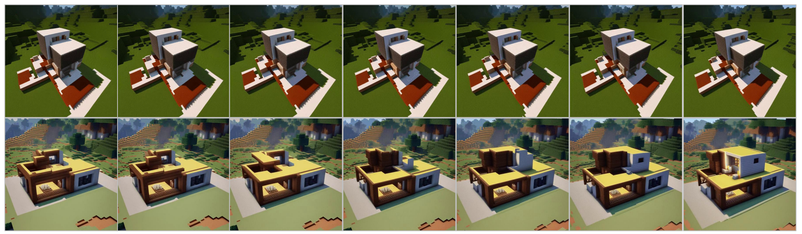
Shenghai Yuan, Jinfa Huang, Yongqi Xu, YaoYang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, Li Yuan

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Contributions

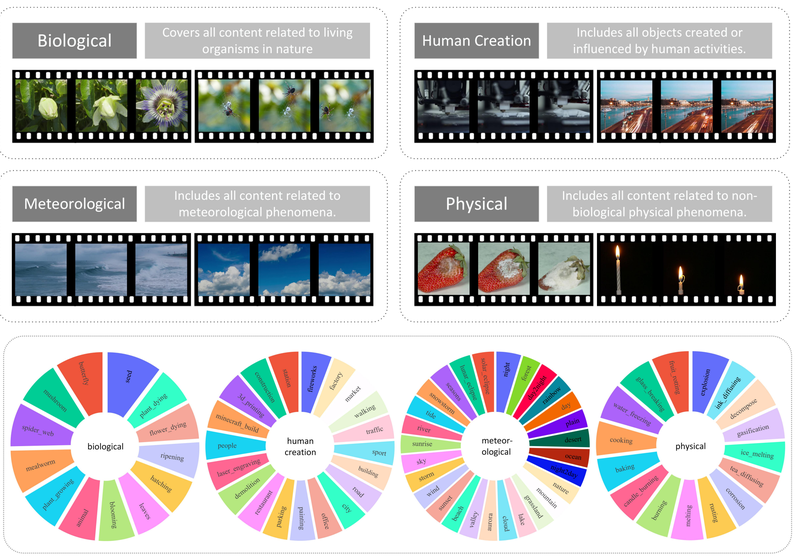
- **New T2V Benchmark.** A new T2V benchmark focusing on metamorphic amplitude, temporal coherence, visual quality and text relevance.
- **New Automatic Metrics.** Develop MTScore and CHScore for assessing metamorphic attributes and temporal coherence.
- **Large-Scale Time-lapse Video-Text Dataset.** Create ChronoMagic-Pro, a dataset with 460k high-quality 720p time-lapse videos and detailed captions. (contains more physics than general videos)



upper: video generated by most of T2V models. (e.g., OpenSora, EasyAnimate)
lower: only a little can generate the complete of time-lapse. (e.g., MagicTime)

Categories Construction

- We construct a search database containing 75 categories of time-lapse videos, and divide it into 4 major categories.



Automatic Metrics

Assessing Metamorphic

MTScore: we designed *N* retrieval sentences, and use a video retrieval model to calculate the probabilities of *n* metamorphic and *m* general videos.

GPT4o-MTScore: we set a 5-point evaluation standard and questionnaire, then ask GPT-4o to rate the score.

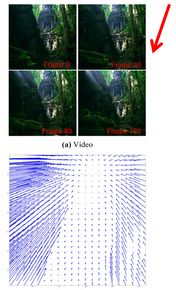
Calculation of Metamorphic Score:

$$S_c = \frac{\sum_{i=1}^{n} P_i^{\text{meta}}}{\sum_{i=1}^{n} P_i^{\text{meta}} + \sum_{i=1}^{m} P_i^{\text{gen}}}$$

Assessing Temporal Coherence

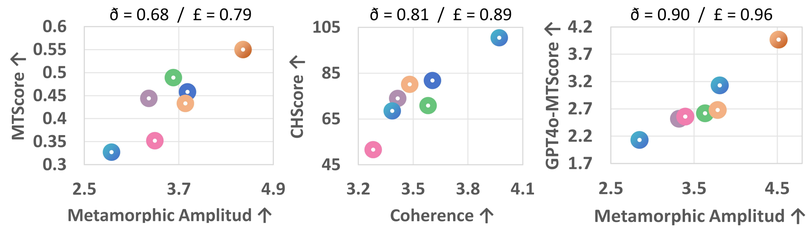
Algorithm 1 Calculation of Coherence Score

Input: Video, pre-trained model with grid size G and threshold T
Output: Coherence score
Process input video using pre-trained model with grid size G and threshold T to get p_{vis}
for each frame i **do**
  count the number of missing tracking points in each frame (except the time vanishing point)
  $m[i] \leftarrow \frac{1}{N}\sum_{j=1}^{N}(1 - p_{\text{vis}}[0, i, j])$
end for
for each frame i **do**
  $\Delta m[i] \leftarrow |m[i+1] - m[i]|$
  if $\Delta m[i] > T$ **then**
    frame i will be added to the set frames_to_be_cut
    $C_{\text{missed}} \leftarrow C_{\text{missed}} + \Delta m[i]$
  end if
end for
$R_{\text{cut}} \leftarrow \frac{\text{len(frames\_to\_be\_cut)}}{\text{frames}}$
$R_{\text{missed}} \leftarrow \frac{1}{\text{frames}}\sum_{i=1}^{\text{frames}} m[i]$
$V_{\text{missed}} \leftarrow \text{std}(\Delta m)$
$M_{\text{missed}} \leftarrow \max(\Delta m)$
C_sum $\leftarrow \lambda_1 \hat{R}_{\text{missed}} + \lambda_2 \hat{V}_{\text{missed}} + \lambda_3 \hat{R}_{\text{cut}} + \lambda_4 \hat{C}_{\text{missed}} + \lambda_5 \hat{M}_{\text{missed}}$
Coherence_score $\leftarrow \frac{1}{\text{C\_sum}}$

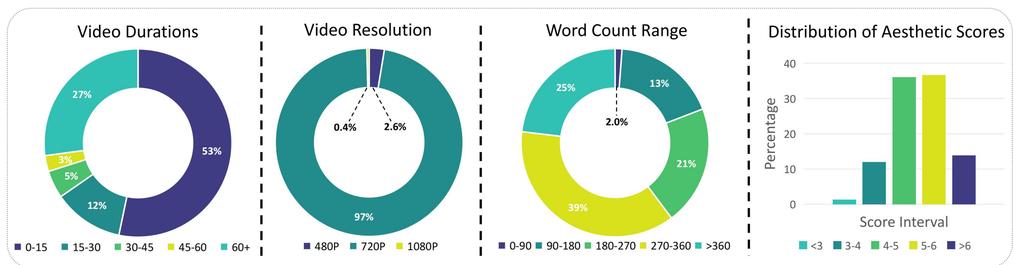


Human Evaluation

The proposed metric is well aligned with human perception



Data Statistic



Dataset	# Categories	Video clips	Resolution	Type	Average length	Video duration (h)
MSR-VTT [78]	General	10K	240p	Video-Text	15.0s	40
WebVid-10M [2]	General	10M	360p	Video-Text	18.72s	52K
InternVid [72]	General	234M	720p	Video-Text	11.90s	760.3K
Panda-70M [16]	General	70M	720p	Video-Text	8.50s	166.8K
HD-VG-130M [70]	General	130M	720p	Video-Text	4.93s	178K
Time-Lapse-D [76]	Time-lapse	2K	360p	Video	-	-
Sky Time-Lapse [80]	Time-lapse	17K	1080p	Video	-	-
ChronoMagic [83]	Time-lapse	2K	720p	Video-Text	11.4s	7
ChronoMagic-Pro	Time-lapse	460K	720p	Video-Text	234s	30K

- We construct the first large-scale time-lapse video dataset by collecting time-lapse videos based on the search terms, which contains more physics than general videos.

Results

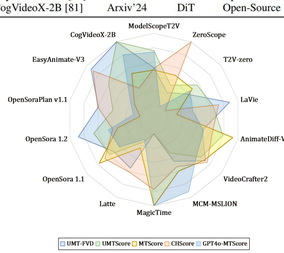
Backbone	Type	Visual Quality	Text Relevance	Metamorphic Amplitude	Temporal Coherence
UCF-101	General	✔	✔	✘	✘
Make-a-Video-Eval	General	✔	✔	✘	✘
MSR-VTT	General	✔	✔	✘	✘
FETV	General	✔	✔	✘	✔
VBench	General	✔	✔	✘	✔
T2VScore	General	✔	✔	✘	✘
ChronoMagic-Bench	Time-lapse	✔	✔	✔	✔

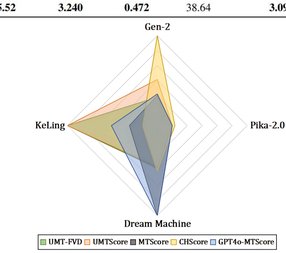
- Our bench emphasizes generating videos with high persistence and strong variation (e.g., high physical prior content.)

ChronoMagic-Bench	Venue	Backbone	UMT-FVD↓	UMTScore↑	MTScore↑	CHScore↑	GPT4o-MTScore↑
ModelScopeT2V [68]	Arxiv'23	U-Net	194.77	2.909	0.401	61.07	2.86
ZeroScope [64]	CVPR'23	U-Net	227.02	2.350	0.400	**99.67**	2.09
T2V-zero [28]	ICCV'23	U-Net	209.66	2.661	0.400	20.78	2.55
LaVie [71]	Arxiv'23	U-Net	**166.97**	2.763	0.346	77.89	2.46
AnimateDiff V3 [22]	ICLR'24	U-Net	197.89	**2.944**	0.467	70.85	2.62
VideoCrafter2 [11]	Arxiv'24	U-Net	178.45	2.753	0.433	80.10	2.68
MCM-MSLION [84]	Arxiv'24	U-Net	202.08	2.33	0.417	62.60	3.04
MagicTime [83]	Arxiv'24	U-Net	257.56	1.916	**0.478**	81.82	**3.13**
Latte [47]	Arxiv'24	DiT	192.12	2.111	0.363	68.68	2.20
OpenSora 1.1 [90]	Github'24	DiT	195.43	2.678	**0.444**	73.98	2.52
OpenSora 1.2 [90]	Github'24	DiT	166.92	2.781	0.375	51.60	2.56
OpenSoraPlan v1.1 [41]	Github'24	DiT	188.53	2.421	0.327	68.52	2.19
EasyAnimate V3 [77]	Arxiv'24	DiT	164.30	2.713	0.349	**90.54**	2.32
CogVideoX-2B [81]	Arxiv'24	DiT	**159.31**	**3.225**	0.404	43.15	**2.92**
OpenSoraPlan v1.1†	Ours	DiT	185.72	2.753	0.341	49.85	3.03
OpenSoraPlan v1.1‡	Ours	DiT	**180.11**	**2.864**	**0.346**	**70.12**	**3.05**

- We evaluated several open/closed-source video generation models, providing new insights for T2V model selection.

ChronoMagic-Bench-150	Venue	Backbone	Status	UMT-FVD↓	UMTScore↑	MTScore↑	CHScore↑	GPT4o-MTScore↑
Gen-2 [58]	Runway	U-Net	Close-Source	218.99	**2.400**	0.373	**125.25**	2.62
Pika-2.0 [34]	PikaLab	U-Net	Close-Source	223.05	2.317	0.347	75.98	2.48
Dream Machine [46]	LUMA	DiT	Close-Source	214.91	2.387	**0.474**	95.97	**3.11**
KeLing [33]	Kwai	DiT	Close-Source	**202.32**	2.517	0.369	74.20	2.74
ModelScopeT2V [68]	Arxiv'23	U-Net	Open-Source	230.74	2.783	0.409	61.01	3.01
ZeroScope [64]	CVPR'23	U-Net	Open-Source	260.61	2.232	0.403	**94.67**	2.29
T2V-zero [28]	ICCV'23	U-Net	Open-Source	250.22	2.559	0.399	18.54	2.62
LaVie [71]	Arxiv'23	U-Net	Open-Source	**210.39**	2.714	0.350	81.32	2.50
AnimateDiff V3 [22]	ICLR'24	U-Net	Open-Source	239.31	**2.837**	0.470	70.36	2.62
VideoCrafter2 [11]	CVPR'23	U-Net	Open-Source	214.06	2.763	0.437	75.90	2.87
MCM-MSLION [84]	Arxiv'24	U-Net	Open-Source	244.49	2.282	0.422	58.08	**3.06**
MagicTime [83]	Arxiv'24	U-Net	Open-Source	294.72	1.763	**0.479**	77.98	3.05
Latte [47]	Arxiv'24	DiT	Open-Source	232.29	2.122	0.366	72.57	2.42
OpenSora 1.1 [90]	Github'24	DiT	Open-Source	241.09	2.676	0.448	75.94	2.57
OpenSora 1.2 [90]	Github'24	DiT	Open-Source	210.93	2.681	0.383	51.87	2.50
OpenSoraPlan v1.1 [41]	Github'24	DiT	Open-Source	228.70	2.459	0.331	61.50	2.21
EasyAnimate V3 [77]	Arxiv'24	DiT	Open-Source	202.03	2.733	0.352	**88.48**	2.33
CogVideoX-2B [81]	Arxiv'24	DiT	Open-Source	**195.52**	**3.240**	**0.472**	38.64	**3.09**





Personal Paper LeaderBoard Code