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

✨ NeurIPS D&B 2024 Spotlight ✨

Shenghai Yuan[1], Jinfa Huang[1,3], Yongqi Xu[1], YaoYang Liu[1], Shaofeng Zhang[4], Yujun Shi[5], Ruijie Zhu[6], Xinhua Cheng[1,2], Jiebo Luo[3], Li Yuan[1,†]

[1]Peking University, [2]Rabbitpre Intelligence, [3]University of Rochester, [4]Shanghai Jiao Tong University, [5]National University of Singapore, [6]University of California Santa Cruz

Motivation

Existing T2V generation evaluation is lack of metamorphic benchmark

- existing T2V models have not adequately encoded physical knowledge of the real world, thus generated videos tend to have limited motion and poor variations.

upper: video generated by most of T2V models. (e.g., OpenSora, CogVideoX)

lower: only a little can generate the complete of time-lapse. (e.g., MagicTime)

Motivation

PEKING UNIVERSITY 北京大学

Existing T2V benchmark is lack of reliable metrics for physical assessment

- a common practice is to report aesthetic quality and textual adherence, but ignores how to assess how much physical priors are encoded in the model.

Benchmark	Type	Visual Quality	Text Relevance	Metamorphic Amplitude
UCF-101 [63]	General	✓	✓	✗
Make-a-Video-Eval [61]	General	✓	✓	✗
MSR-VTT [78]	General	✓	✓	✗
FETV [45]	General	✓	✓	✗
VBench [26]	General	✓	✓	✗
T2VScore [74]	General	✓	✓	✗
ChronoMagic-Bench (Ours)	Time-lapse	✓	✓	✓

Benchmark Construction

Prompt Categorization

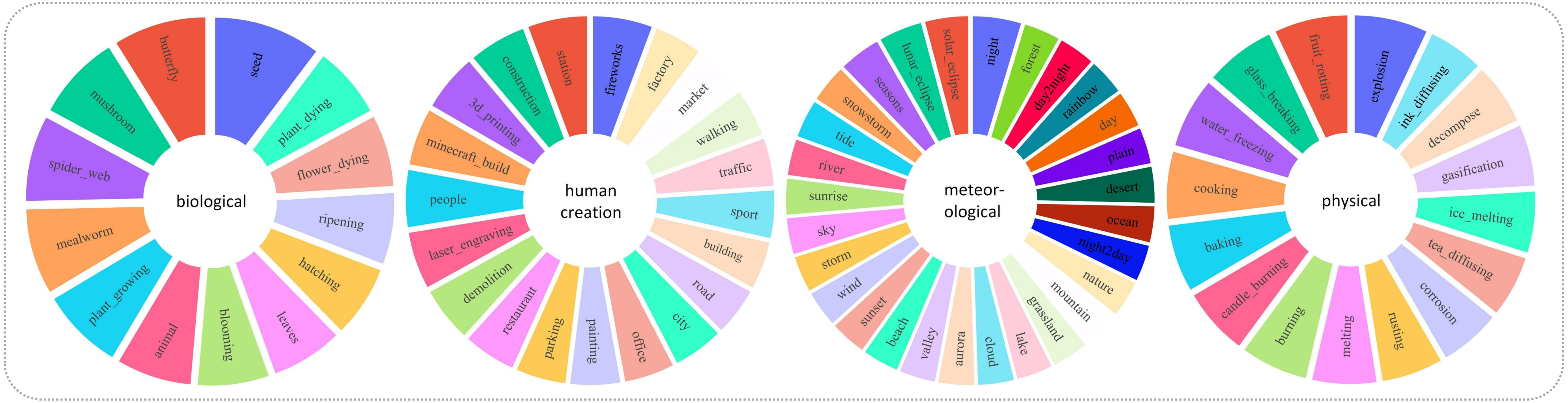
- Step1: Hand-crafted rules for automatic categorization
- Step2: Manual selection and revision
- Step3: Crawl real-world videos from Internet
- Step4: Obtain annotations using GPT-4o

Automatic Metric Design

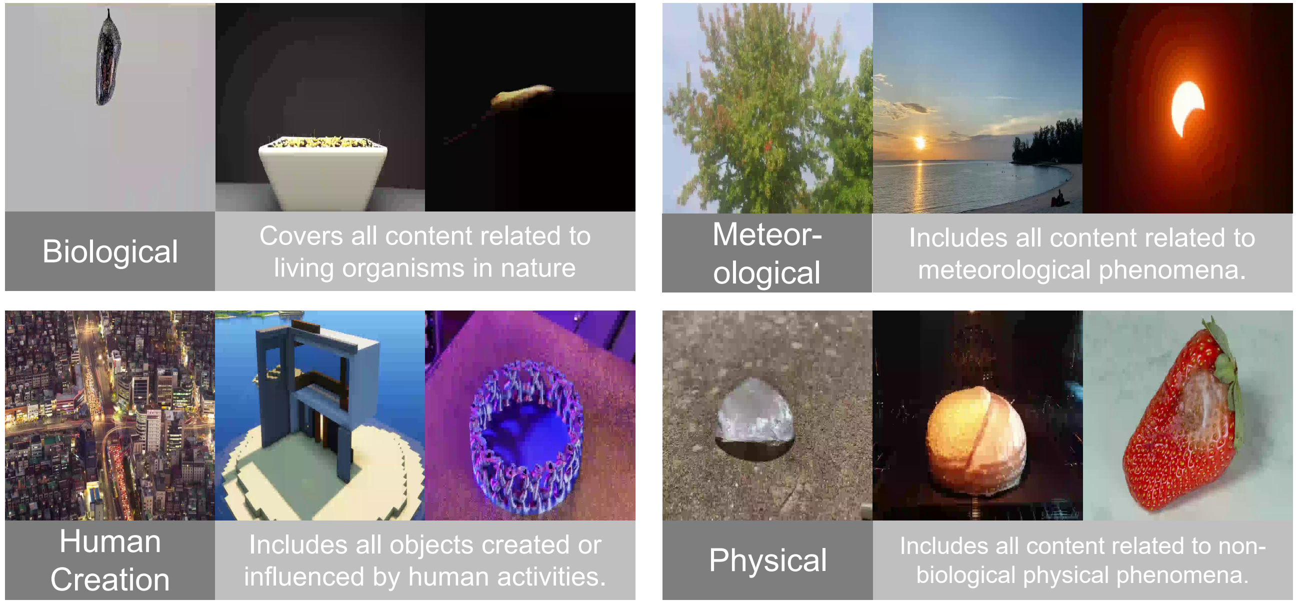
- MTScore: for coarse-grained metamorphic assessment
- GPT4o-MTScore: for fine-grained assessment
- CHScore: evaluate the aesthetics of the time-lapse process

Prompt Categorization

4 types of time-lapse videos: biological, human creation, meteorological, and physical phenomena, which are further divided into **75** subcategories.

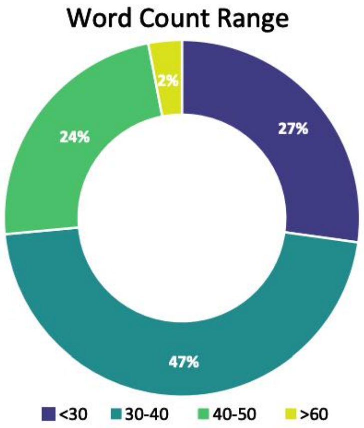


Chronomagic-Bench: Data Analysis

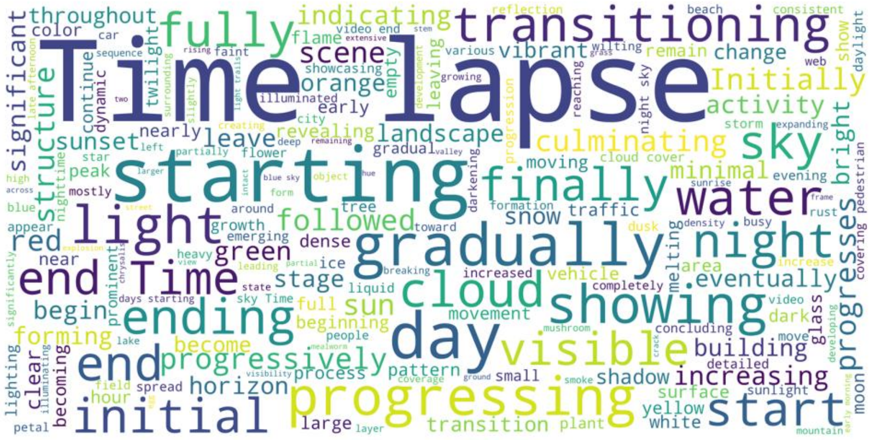


ChronoMagic-Bench introduces

1,649 prompts and real-world videos.



(a)



(b)

Assessing Metamorphic: MTScore & GPT4o-MTScore

- **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.

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

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.

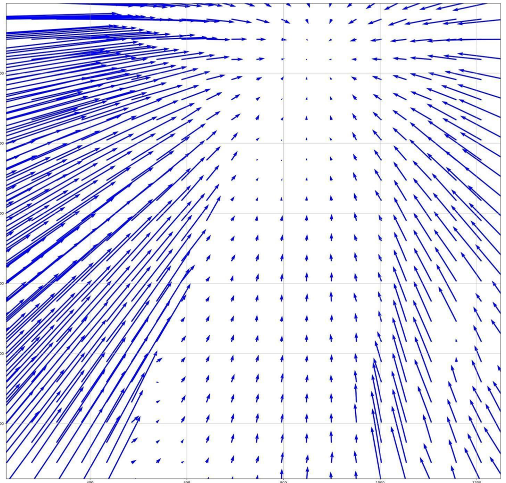
Score	Brief Reasoning Statement
1	Minimal change. The scene appears almost like a still image, with static elements remaining motionless and only minor changes in lighting or subtle movements of elements. No significant activity is noticeable.
2	Slight change. There is a small amount of movement or change 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.

Assessing Temporal Coherence: Coherence Score

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
$R_{\text{cut}} \leftarrow \frac{\text{len(frames\_to\_be\_cut)}}{\text{frames}}$
$R_{\text{missed}} \leftarrow \frac{1}{\text{frames}} \sum_{i=1}^{\text{frames}} m[i]$
$V_{\text{missed}} \leftarrow \text{std}(\Delta m)$
$M_{\text{missed}} \leftarrow \max(\Delta m)$
$\text{C\_sum} \leftarrow \lambda_1 \hat{R}_{\text{missed}} + \lambda_2 \hat{V}_{\text{missed}} + \lambda_3 \hat{R}_{\text{cut}} + \lambda_4 \hat{C}_{\text{missed}} + \lambda_5 \hat{M}_{\text{missed}}$
$\text{Coherence\_score} \leftarrow \frac{1}{\text{C\_sum}}$

(a) Video



(b) Points Direction

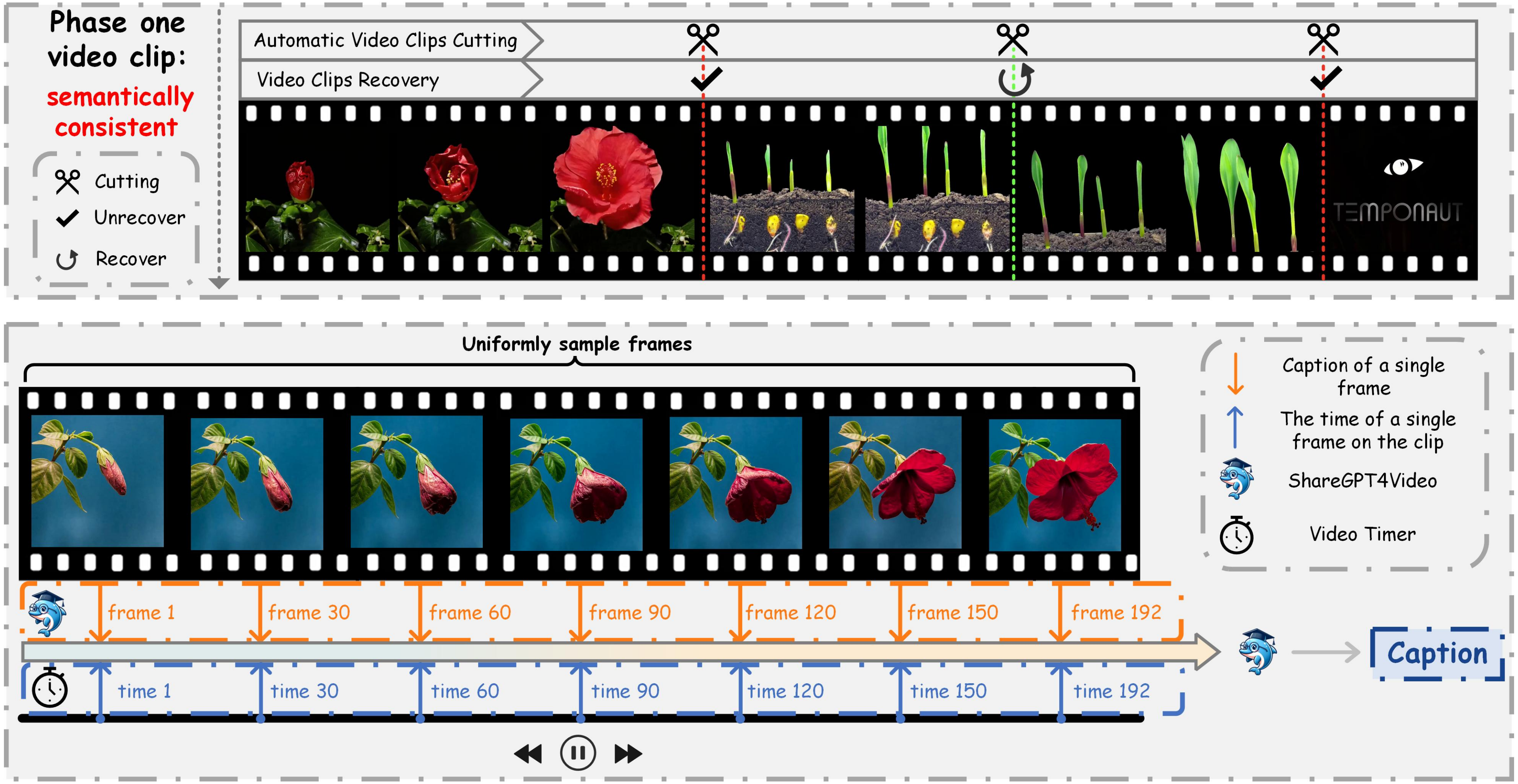
[1] Karaev, Nikita, et al. "Cotracker: It is better to track together." ECCV 2024.

ChronoMagic-Pro

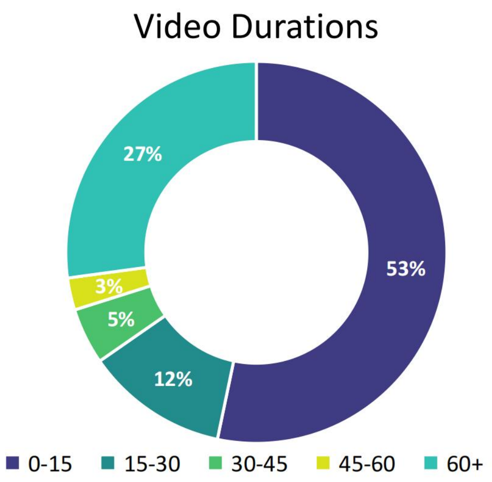
- We construct the first large-scale time-lapse video dataset by collecting time-lapse videos based on the search terms, which contains more physics than general videos.

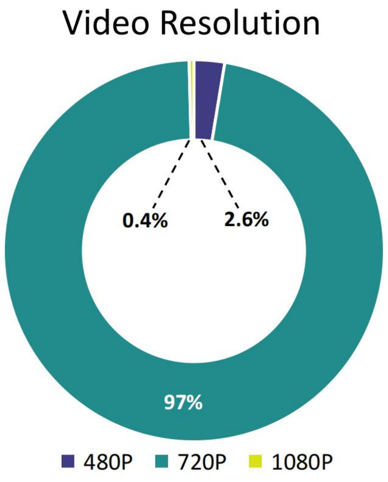
Dataset	# Categories	Video clips	Resolution	Type	Average length	Video duration (h)
MSR-VTT [78]	General	10K	240p	Video-Text	15.0s	40
WebVid-10M [2]	General	10M	360p	Video-Text	18.72s	52K
InternVid [72]	General	234M	720p	Video-Text	11.90s	760.3K
Panda-70M [16]	General	70M	720p	Video-Text	8.50s	166.8K
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

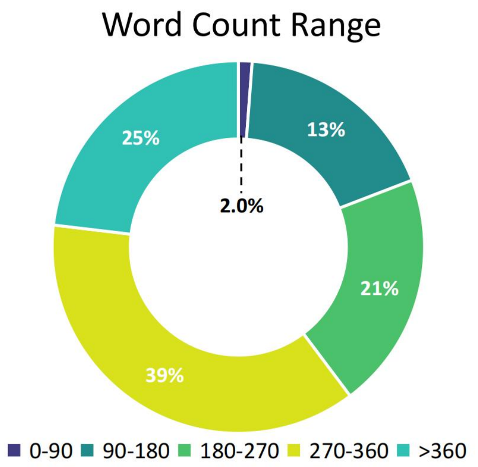
Pipeline of Constructing ChronoMagic-Pro

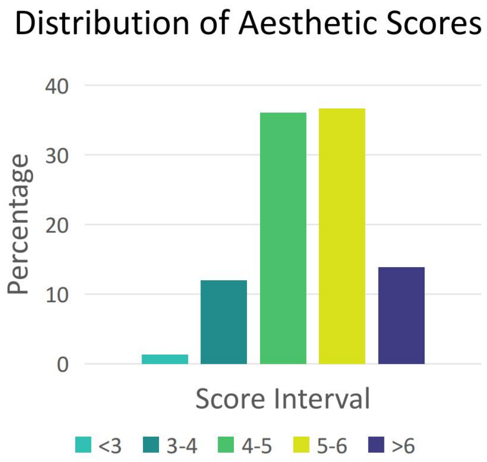


ChronoMagic-Pro: Dataset Statistic









460k high-quality pairs of 720p time-lapse videos and detailed captions. Each caption ensures high physical content and large metamorphic amplitude.

Main Results of ChronoMagic-Bench

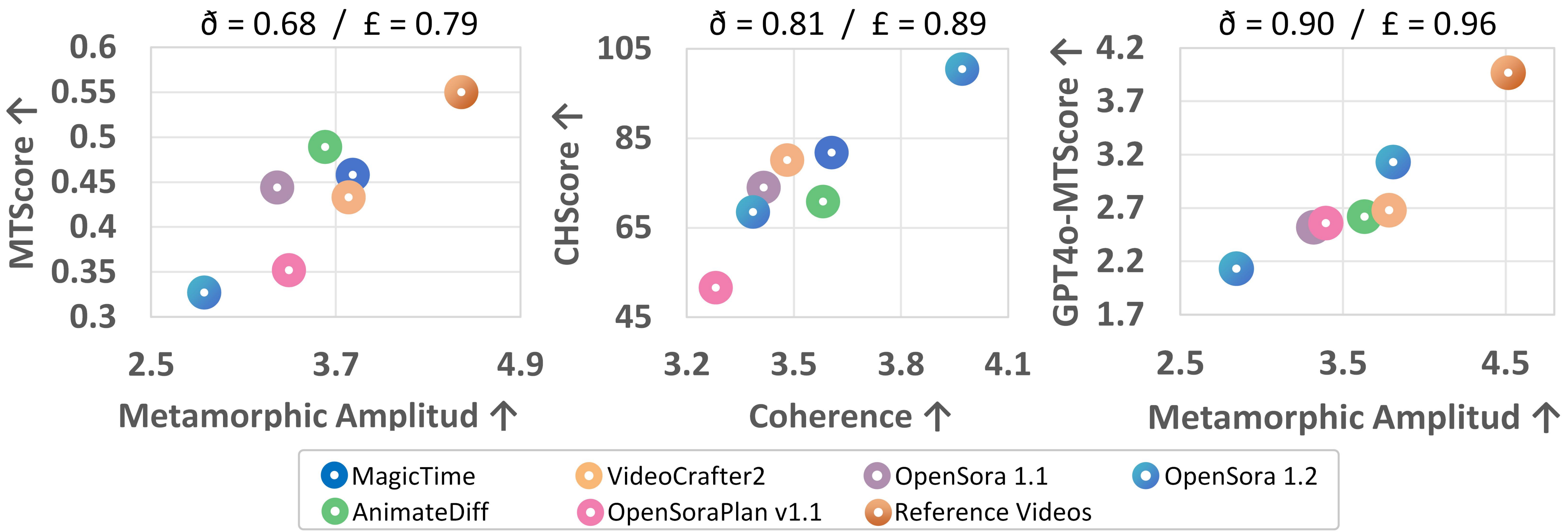
Method	Venue	Backbone	UMT-FVD↓	UMTScore↑	MTScore↑	CHScore↑	GPT4o-MTScore↑
ModelScopeT2V [68]	Arxiv'23	U-Net	194.77	2.909	0.401	61.07	2.86
ZeroScope [64]	CVPR'23	U-Net	227.02	2.350	0.400	**99.67**	2.09
T2V-zero [28]	ICCV'23	U-Net	209.66	2.661	0.400	20.78	2.55
LaVie [71]	Arxiv'23	U-Net	**166.97**	2.763	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**

Main Results of ChronoMagic-Bench-150

- 4 latest closed source T2V models and 14 open source T2V models, providing useful insights for users to choose suitable T2V models.

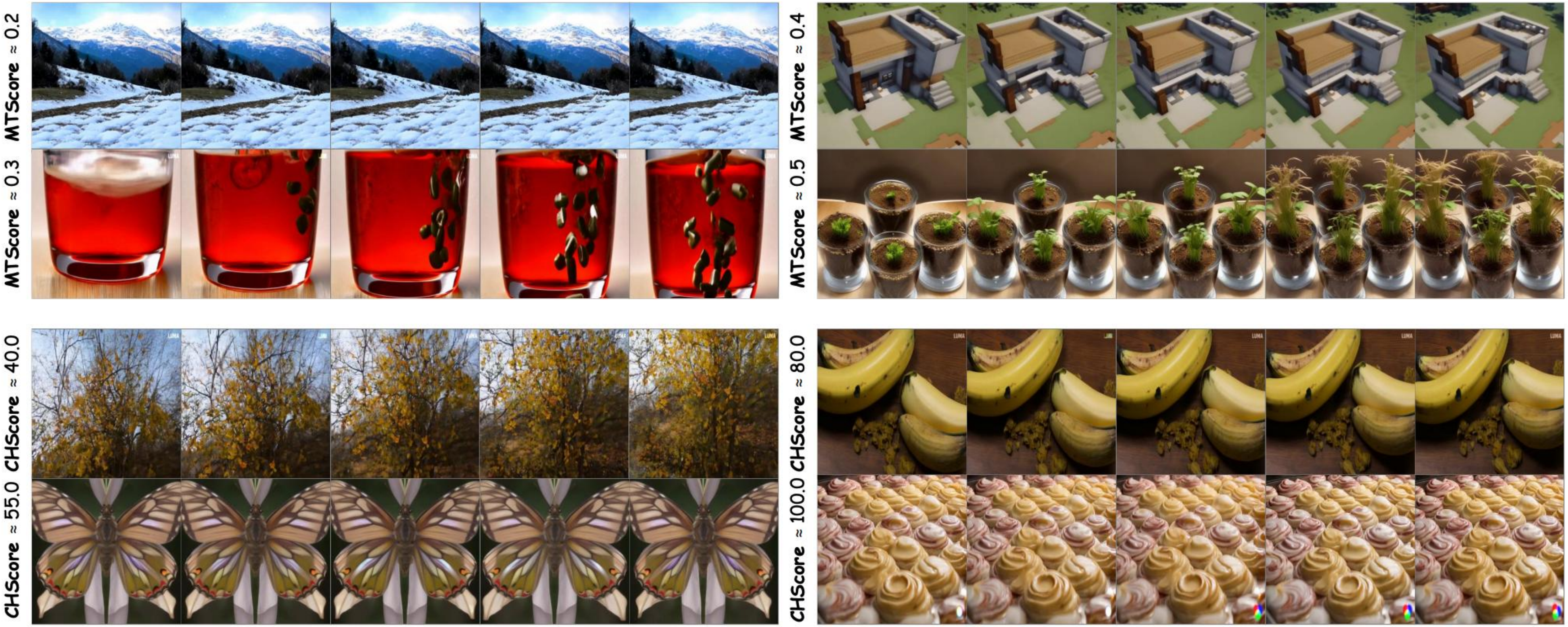
Method	Venue	Backbone	Status	UMT-FVD↓	UMTScore↑	MTScore↑	CHScore↑	GPT4o-MTScore↑
Gen-2 [63]	Runway	U-Net	Close-Source	218.99	**2.400**	0.373	**125.25**	2.62
Pika-1.0 [36]	PikaLab	U-Net	Close-Source	223.05	2.317	0.347	75.98	2.48
Dream Machine [48]	LUMA	DiT	Close-Source	214.91	2.387	**0.474**	95.97	**3.11**
KeLing [35]	Kwai	DiT	Close-Source	**202.32**	2.517	0.369	74.20	2.74
ModelScopeT2V [73]	Arxiv'23	U-Net	Open-Source	230.74	2.783	0.409	61.01	3.01
ZeroScope [69]	CVPR'23	U-Net	Open-Source	260.61	2.232	0.403	**94.67**	2.29
T2V-zero [30]	ICCV'23	U-Net	Open-Source	250.22	2.559	0.399	18.54	2.62
LaVie [76]	Arxiv'23	U-Net	Open-Source	**210.39**	2.714	0.350	81.32	2.50
AnimateDiff V3 [23]	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 [89]	Arxiv'24	U-Net	Open-Source	244.49	2.282	0.422	58.08	**3.06**
MagicTime [88]	Arxiv'24	U-Net	Open-Source	294.72	1.763	**0.479**	77.98	3.05
Latte [49]	Arxiv'24	DiT	Open-Source	232.29	2.122	0.366	72.57	2.42
OpenSora 1.1 [95]	Github'24	DiT	Open-Source	241.09	2.676	0.448	75.94	2.57
OpenSora 1.2 [95]	Github'24	DiT	Open-Source	210.93	2.681	0.383	51.87	2.50
OpenSoraPlan v1.1 [43]	Github'24	DiT	Open-Source	228.70	2.459	0.331	61.50	2.21
EasyAnimate V3 [82]	Arxiv'24	DiT	Open-Source	202.03	2.733	0.352	**88.48**	2.33
CogVideoX-2B [86]	Arxiv'24	DiT	Open-Source	**195.52**	**3.240**	**0.472**	38.64	**3.09**

Reasonableness of Automatic Evaluation Metrics



The proposed metrics are well aligned with human perception.

Visual Reference for Varying Scores



Qualitative Analysis of Our Benchmark

Gen-2

MagicTime

CogVideo2B

KeLing

EasyAnimateV3

OpenSora1.2

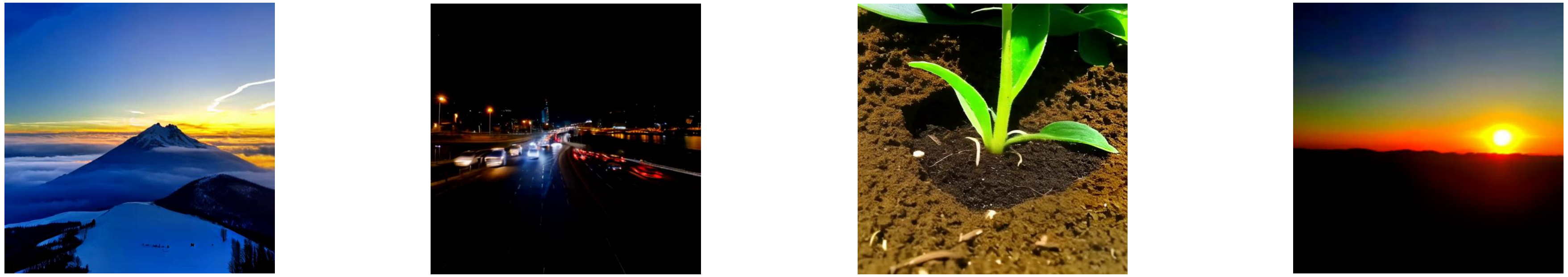
LUMA

Pika 1.0

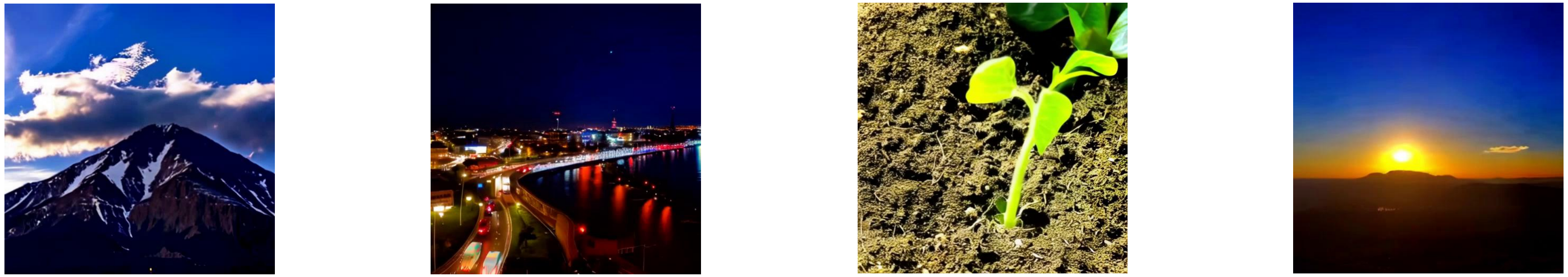
OpenSoraPlan v1.1

Dataset Verification (with ChronoMagic-Pro 10K)

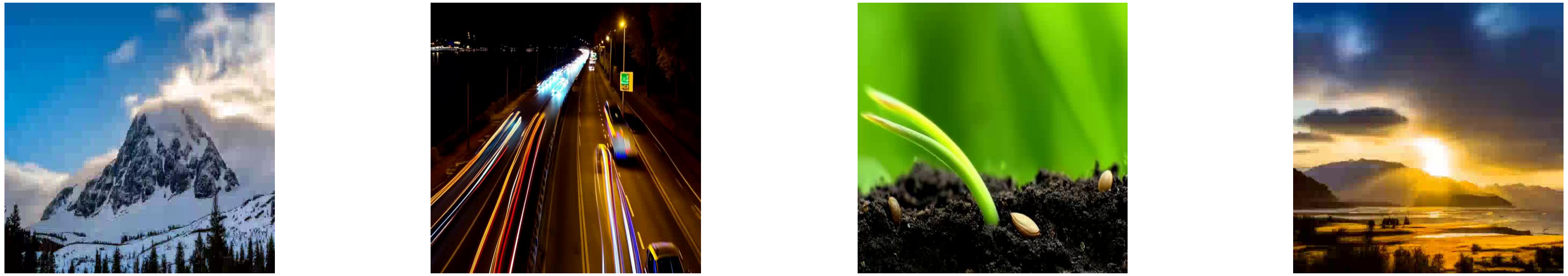
Original



Simple Finetune



Magic Training



Conclusion

- **New T2V Benchmark.** We introduce ChronoMagic-Bench for comprehensive evaluation of T2V models, focusing on visual quality, text relevance, metamorphic amplitude, and temporal coherence.
- **New Automatic Metrics.** We develop MTScore and CHScore, which align better with human judgment than existing metrics, for assessing metamorphic attributes and temporal coherence.
- **New Insights for T2V Model Selection.** Our evaluations using ChronoMagic-Bench provide crucial insights into the strengths and weaknesses of various T2V models.
- **Large-Scale Time-lapse Video-Text Dataset.** We create ChronoMagic-Pro, a dataset with 460k high-quality 720p time-lapse videos and detailed captions, promoting advances in T2V research.

Follow up (From paper to product)

Allegro: Open the Black Box of Commercial-Level Video Generation Model

Yuan Zhou, Qiuyue Wang, Yuxuan Cai, Huan Yang*
Rhymes AI

2 Data Curation

Data curation is the primary task in building video generation models, permeating the entire training process. Existing publicly available datasets, such as WebVid [Bain et al., 2021], Panda-70M [Chen et al., 2024b], HD-VILA [Xue et al., 2022], HD-VG [Wang et al., 2023] and OpenVid-1M [Nan et al., 2024], have provided solid foundation for data sourcing and acquisition, offering diverse and extensive video data. However, with the sheer volume of data now available, significant challenges arise in terms of processing efficiency, data redundancy, and ensuring high-quality inputs for model training.

[2] Zhou, Yuan, et al. "Allegro: Open the Black Box of Commercial-Level Video Generation Model." *arXiv preprint 2024.*

3 Curating `OpenVid-1M`

This section outlines the date processing steps as detailed in Table 1. `OpenVid-1M` is curated from ChronoMagic, CelebvHQ [26], Open-Sora-plan [3] and Panda[3]. Since Panda is much larger than the other datasets, here we primarily describe the filtering details on our downloaded Panda-50M.

[3] Nan, Kepan, et al. "Openvid-1m: A large-scale high-quality dataset for text-to-video generation." arXiv preprint 2024.

Table 3: Evaluation results of CogVideoX-5B and CogVideoX-2B.

Models	Human Action	Scene	Dynamic Degree	Multiple Objects	Appear. Style	Dynamic Quality	GPT4o-MT Score
T2V-Turbo	95.2	**55.58**	49.17	54.65	24.42	–	
AnimateDiff	92.6	50.19	40.83	36.88	22.42	–	2.62
VideoCrafter-2.0	95.0	55.29	42.50	40.66	**25.13**	43.6	2.68
OpenSora V1.2	85.8	42.47	47.22	58.41	23.89	63.7	2.52
Show-1	95.6	47.03	44.44	45.47	23.06	57.7	–
Gen-2	89.2	48.91	18.89	55.47	19.34	43.6	2.62
Pika	88.0	44.80	37.22	46.69	21.89	52.1	2.48
LaVie-2	96.4	49.59	31.11	64.88	25.09	–	2.46
CogVideoX-2B	88.0	39.94	**63.33**	53.70	23.67	57.7	3.09
CogVideoX-5B	**96.8**	55.44	62.22	**70.95**	24.44	**69.5**	**3.36**

[4] Yang, Zhuoyi, et al. "Cogvideox: Text-to-video diffusion models with an expert transformer." arXiv preprint 2024.

Follow up (As a benchmark for T2V post-alignment)

PEKING UNIVERSITY

SePPO: Semi-Policy Preference Optimization for Diffusion Alignment

Daoan Zhang[1] *, Guangchen Lan[2] *, Dong-Jun Han[3], Wenlin Yao[4], Xiaoman Pan[4], Hongming Zhang[4], Mingxiao Li[4], Pengcheng Chen[5], Yu Dong[4], Christopher Brinton[2], Jiebo Luo[1]

[1] University of Rochester, [2] Purdue University, [3] Yonsei University, [4] Tencent AI Lab, [5] University of Washington

[5] Zhang, Daoan, et al. "SePPO: Semi-Policy Preference Optimization for Diffusion Alignment." arXiv preprint 2024.

Table 3: Metric Scores on the ChronoMagic-Bench-150 Dataset. ↓ indicates the lower the better, and ↑ indicates the higher the better.

	FID ↓	LPIPS ↓	SSIM ↑	PSNR ↑	FVD ↓
AnimateDiff	134.86	0.68	0.16	9.18	1608.41
SFT	129.14	0.65	0.17	9.25	1415.68
SePPO	**115.32**	**0.61**	**0.20**	**9.36**	**1300.97**

align-anything Public

Watch 8 | Fork 31 | Starred 194

main | 1 Branch | Tags | Go to file | Add file | Code

htlou feat: support SFT for Emu3 models (#85) 642ca1e · 2 weeks ago 55 Commits

.github	fix: fix examples, deps, and typos (#72)	last month
align_anything	feat: support SFT for Emu3 models (#85)	2 weeks ago
assets	docs: Update Readme (#71)	last month

About

Align Anything: Training All-modality Model with Feedback

align-anything.readthedocs.io

chameleon, multimodal, dpo, large-language-models, rlhf, vision-language-model

[6] https://github.com/PKU-Alignment/align-anything

Evaluation

We support evaluation datasets for `Text -> Text`, `Text+Image -> Text` and `Text -> Image`.

Modality	Supported Benchmarks
`t2t`	ARC, BBH, Belebele, CMMLU, GSM8K, HumanEval, MMLU, MMLU-Pro, MT-Bench, PAWS-X, RACE, TruthfulQA
`ti2t`	A-OKVQA, LLaVA-Bench(COCO), LLaVA-Bench(wild), MathVista, MM-SafetyBench, MMBench, MME, MMMU, MMStar, MMVet, POPE, ScienceQA, SPA-VL, TextVQA, VizWizVQA
`tv2t`	MVBench, Video-MME
`ta2t`	AIR-Bench
`t2i`	ImageReward, HPSv2, COCO-30k(FID)
`t2v`	ChronoMagic-Bench
`t2a`	AudioCaps(FAD)

Paper

LeaderBoard

Code

Thank you!

[7] https://github.com/PKU-YuanGroup/ChronoMagic-Bench