

320	Contents	
321	A Supplementary Video	10
322	B Method Details	10
323	B.1 Temporal Sensitivity Analysis	10
324	B.2 ArrowRL	11
325	B.3 AoTBench	13
326	C Additional Results	13
327	C.1 Quantitative Results	13
328	C.2 Qualitative Results	15
329	D Limitations	15
330	E Societal Impacts	16

331 A Supplementary Video

332 We invite readers to view the supplementary video attached for a visual demonstration of our work’s
333 overview and additional qualitative examples. Understanding how video content differs between
334 forward and reverse playback—a key aspect of our study—is most effectively conveyed through
335 video. Therefore, the supplementary video will offer readers a more intuitive grasp of how ArrowRL
336 successfully enhances temporal sensitivity in LMM responses and how AoTBench effectively probes
337 this crucial capability.

338 B Method Details

339 B.1 Temporal Sensitivity Analysis

340 The temporal sensitivity analysis, visualized in Fig. 2 of the main paper, is conducted using
341 LLaVA-OV-7B [29] across several evaluation benchmarks: EgoSchema [39], LongVideoBench [60],
342 MVBench [31], NExT-QA [62], PerceptionTest [11], TemporalBench [6], VITATECS [32], and
343 VideoMME [17]. To ensure consistent and fair evaluation, we utilize *lmms-eval*² and a standardized
344 setting of 16 input frames across three conditions: forward, reversed and shuffled video. We specifi-
345 cally select deterministic MCQ tasks to obviate the need for potentially unreliable third-party LLM
346 evaluators and mitigate evaluation ambiguity.

347 Our subsequent analyses, including the development of AoTBench, focus on short videos. This
348 approach allows for a targeted investigation of AoT awareness, separating it from complexities
349 specific to long video processing, especially since current LMMs already struggle with shorter
350 temporal sequences. Table 3 presents the forward, reversed, and shuffled frame performance (i.e.,
351 MCQ accuracy) for three selected LMMs (LLaVA-OV-7B [29], LLaVA-Video-7B [76], and Qwen2.5-
352 VL-7B [3]) on several VQA benchmarks: VITATECS [32], TemporalBench [6], NExT-QA [62],
353 PerceptionTest [43], VideoMME [17], Vinoground [71], TempCompass [38] and TVBench [12].
354 These results, which supplement the TDS-based benchmark sensitivity analysis in Fig. 3 of the main
355 paper, further illustrates the great variance in temporal order sensitivity across benchmarks: some,
356 such as Vinoground, TempCompass, and TVBench, demonstrate a stronger ability to probe this,
357 whereas others, like VITATECS, show extreme insensitivity.

²<https://github.com/EvolvingLMMs-Lab/lmms-eval>

Table 3: Impact of video frame order manipulation (forward, reversed, shuffled) on MCQ Accuracy (%) for three LMMs across various VQA benchmarks. S: short, V: video, T: text.

Benchmark	LLaVA-OV-7B			LLaVA-Video-7B			Qwen2.5-VL-7B		
	forward	reverse	shuffled	forward	reverse	shuffled	forward	reverse	shuffled
VITATECS	85.27	83.00	84.98	87.74	85.87	87.28	82.95	81.33	81.54
TemporalBench (S)	61.92	57.53	59.04	62.85	58.43	59.57	68.36	66.24	67.23
NeXT-QA	78.29	77.31	77.85	81.97	80.27	80.80	81.47	79.80	79.80
PerceptionTest	57.15	55.72	56.52	67.63	65.07	65.57	68.81	64.86	65.89
VideoMME (S)	70.89	69.22	68.89	76.33	72.89	73.67	72.33	71.00	69.56
Vinoground (V)	58.00	56.90	52.30	56.90	54.20	49.90	56.00	53.40	48.60
Vinoground (T)	68.20	35.80	53.80	65.90	36.80	51.80	61.10	38.70	49.50
TempCompass	69.78	53.34	62.16	72.09	55.29	63.24	73.88	57.15	65.21
TVBench	49.11	33.86	40.51	53.19	34.73	42.77	54.69	33.66	42.34

B.2 ArrowRL

Post-training Data To enhance AoT understanding, our training data incorporates three core tasks. For MCQ-based sequence direction classification, we use selected videos from UCF101 [50]. For video captioning, we leverage the RTime dataset [14] with a varied set of 16 prompts. For open-ended QA, we employ original questions from LLaVA-NeXT-178K [76]. The prompts used for MCQ and captioning tasks are detailed below. For the MCQ-based sequence direction classification task, we follow [71] to concatenate the forward video and its reversed version—separated by a 2-second black frame—into a single video input, as LMMs typically process one video stream at a time.

Training Prompt I (MCQ-based Sequence Direction Classification)

This video contains two segments showing the same action, separated by a 2-second black frame. One segment is played forwards, and the other is played in reverse. Which video segment is played in reverse?
A. The first segment (before the black frame)
B. The second segment (after the black frame)
Answer with the option’s letter from the given choices directly.

Training Prompt II (Video Captioning)

Describe the following video in detail.
Provide a detailed description of the given video.
Give an elaborate explanation of the video you see.
Share a comprehensive rundown of the presented video.
Offer a thorough analysis of the video.
Explain the various aspects of the video before you.
Clarify the contents of the displayed video with great detail.
Characterize the video using a well-detailed description.
Break down the elements of the video in a detailed manner.
Walk through the important details of the video.
Portray the video with a rich, descriptive narrative.
Narrate the contents of the video with precision.
Analyze the video in a comprehensive and detailed manner.
Illustrate the video through a descriptive explanation.
Examine the video closely and share its details.
Write an exhaustive depiction of the given video.

Reward Calculation Fig. 8 demonstrates ArrowRL’s reward calculation using a VQA example that necessitates causal-temporal reasoning. Here, the fidelity reward (r_i^{fid}) alone might be insufficient; for instance, candidates o_1 and o_3 both exhibit high similarity to the target o^* . The reverse reward

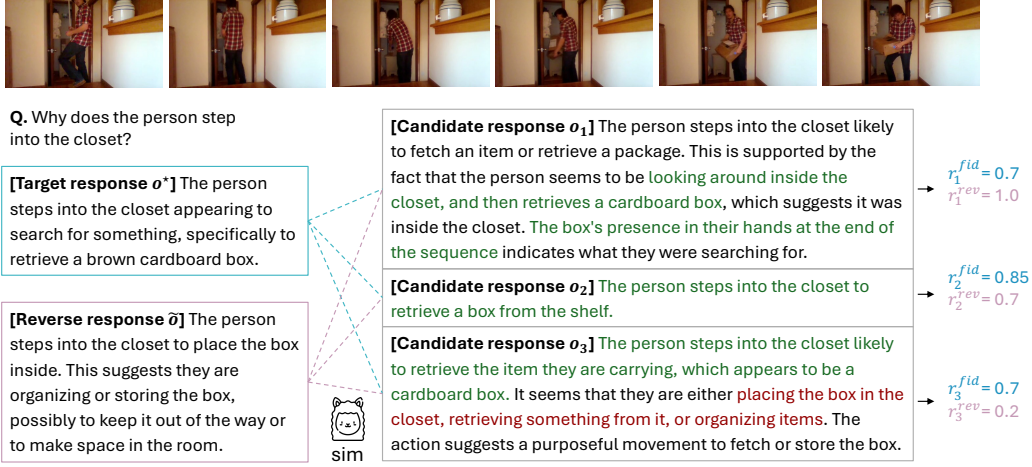


Figure 8: An illustration of the reward calculation process for ArrowRL, using one VQA example from LLaVA-Video-178K [76]. An auxiliary LLM is employed to compute similarity scores between responses. While the fidelity reward r_i^{fid} ensures candidate responses o_i align with the target o^* , the reverse reward r_i^{rev} cultivates AoT sensitivity by using dissimilarity from the reverse response \tilde{o} as its signal. Consequently, temporally correct and sensitive responses that diverge from \tilde{o} (like o_1) are favored, while those that misunderstand AoT (like o_3 , highlighted red) are penalized via a lower reward.

distinguishes them by leveraging the semantic difference between forward and reverse video plays. A response \tilde{o} to the reversed video (e.g., describing “organizing or storing the box”) is used as a negative reference. The reverse reward, $r_i^{rev} = 1 - \text{Similarity}(o_i, \tilde{o})$, then penalizes forward-video responses like o_3 if they incorrectly align with the reverse response \tilde{o} (as indicated by red highlighting), thereby favoring temporally aware responses like o_1 that accurately reflect the forward video’s AoT.

Implementation As discussed in Sec. 3.2 of the main paper, $\text{Similarity}(\cdot, \cdot)$ return a similarity score between 0 and 1 and is implemented as follows. For MCQ tasks, it uses deterministic value checking (1.0 for a correct match, 0.0 otherwise); For open-ended QA and captioning, we employ Llama-3.1-70B-Instruct [20] as an LLM judge, which is prompted to output a semantic similarity score within the $[0, 1]$ range, using the prompts below. The previously defined language query l , target response o^* , candidate response o_i and reverse response \tilde{o} are referenced here.

LLM-based Similarity Calculation

[Open-ended QA prompt]

Please compare the following two answers for the question below and rate their similarity on a scale of 0 to 1.

Question: l

Answer 1: o_i

Answer 2: \tilde{o}

Output only a single numeric value between 0 and 1 (no additional text or explanation).

[Captioning Prompt]

Compare the following video caption with the ground truth caption and rate their similarity on a scale of 0 to 1.

Generated caption: o_i

Ground truth caption: o^*

Output only a single numeric value between 0 and 1 (no additional text or explanation).

382

During training, candidate responses o_i are generated with a temperature of 1.0 to encourage exploration. The reverse-conditioned response \tilde{o} is generated deterministically (temperature set to 0).

ArrowRL training for each model involves 2000 RL steps over approximately 3 days on 6 NVIDIA GH200 GPUs.

To ensure a fair comparison, inference settings for base LMMs and their ArrowRL-enhanced are identical, differing only by model checkpoint. Our default input configuration is 16 frames for LLAVA-OV-7B, and 1 FPS (with a maximum of 16 frames) for Qwen2-VL-7B and Qwen2.5-VL-7B. Benchmark-specific adjustments include processing up to 32 frames (sampled at 1 FPS for Qwen models) for TVBench (due to video length) and reporting Vinoground at 4FPS for Qwen models to align with [71] (further frame rate analysis in Fig. 10). For TempCompass, adhering to [73], we report only on its deterministic subtasks (multi-choice QA, yes/no QA, caption matching; 5536 samples).

B.3 AoTBench

To evaluate AoT perception in LMMs, we construct AoTBench. The dataset composition is detailed in Table 4, with illustrative examples in Fig. 9. For the first and second T2V task, we concatenate the forward, a two-second black video and reversed video segments into a single input, and use the same prompt as Training Prompt I above, but with shuffled MCQ options to test model generalization.

Furthermore, Table 5 quantifies the increased AoT sensitivity of our selected VQA subset, showing significantly higher TDS values compared to the original benchmark.

Table 4: AoTBench Dataset Breakdown. The benchmark comprises three distinct tasks, derived from a diverse suite of video sources, designed to assess AoT awareness of LMMs.

Task	Video Source	# VQA
Sequence Direction Classification	ReverseFilm [58]	144
	UCF101 [50]	500
Directional Caption Matching (V2T)	RTime [14]	1,992
Directional Caption Matching (T2V)	RTime [14]	1,992
AoT-sensitive VQA	VITATECS [32], TemporalBench [6], NExT-QA [62], PerceptionTest [43], VideoMME [17], Vinoground [71], TempCompass [38], TVBench [12]	1,800

Table 5: Comparing Temporal Divergence Score (TDS) averaged for all samples vs. top 200 selected high-TDS samples across nine existing VQA benchmarks. The selection process yields a 1,800-sample subset with substantially increased average TDS, specifically designed to challenge LMM temporal perception.

	VITATECS	TemporalBench (S)	NExT-QA	PerceptionTest	VideoMME (S)	Vinoground (V)	Vinoground (T)	TempCompass	TVBench
All	0.039	0.062	0.073	0.083	0.113	0.212	0.408	0.461	0.549
Selected	0.738	1.258	1.757	3.185	0.690	0.512	1.287	4.617	3.504

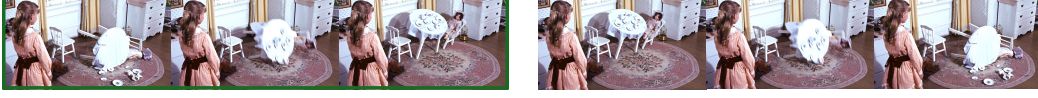
C Additional Results

C.1 Quantitative Results

Further Results Analysis From Table 1 in the main paper, we see that ArrowRL appears to leverage inherent base model strengths for improvements. For instance, the visually adept Qwen2-VL-7B (evidenced by its base performance on T2V caption matching) sees a +21.0% gain with ArrowRL on the visually-focused sequence direction classification task. Meanwhile, Qwen2.5-VL-7B, which exhibits greater language proficiency (reflected in its base V2T caption matching scores), achieves its largest gains (+9.2%) on the more language-centric AoT-VQA task after ArrowRL enhancement.

AoTBench

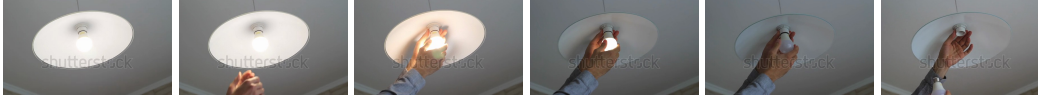
[Sequence Directionality Classification] Two video segments are presented: one plays forward, the other in reverse.
Which video segment is played in reverse?



[Directional Caption Matching – T2V] Two video segments are presented: one plays forward, the other in reverse.
Which video segment matches the caption “someone taking personal care items out of a clear plastic bag”?



[Directional Caption Matching – V2T] Which caption best describes the video?
- A bulb is installed to the white round lamp shade by a person and is lighted up.
- A person removes the luminous bulb from the white round lamp shade with his hands.



[AoT-sensitive VQA] What change is occurring to the 3D house model?
- being constructed
- being dismantled
- being renovated



Figure 9: Visual overview of the three core task components in AoTBench, designed to evaluate different facets of AoT understanding in LMMs.

Expanded Ablation Study Table 6 provides full ablation study results, expanding upon Table 2 in the main paper. Our results validate the design of the reverse reward: its overall contribution (row 2) and dynamic weighting (row 3) are crucial for optimal performance. Note that the reverse reward is not directly applied during training of MCQ-based sequence direction classification; its influence is observed on captioning matching and AoT-VQA tasks (columns 3-5). In addition, the limitations of standard SFT become evident in the direction classification task. As we modify the inference template to be different from training (i.e., by shuffling options), SFT goes through a notable degradation, showcasing its poor generalization under such perturbations, unlike ArrowRL. These ablations collectively demonstrate the importance of ArrowRL’s components for enhancing AoT perception.

Table 6: Ablation study of ArrowRL on AoTBench using Qwen2.5-VL-7B as the base LMM. The effectiveness of ArrowRL’s components is demonstrated: (i) the reverse reward is crucial (removing it in row 2 drops performance below base); (ii) dynamic thresholding for this reward greatly boosts gains (row 3); (iii) SFT fails to generalize on direction classification tasks (row 4).

Model	Direc. Cls.		Cap. Match		AoT-VQA	Average
	<i>RFilm</i>	<i>UCF</i>	T2V	V2T		
Qwen2.5-VL-7B [3]	50.0	51.6	53.4	66.6	49.6	56.2
+ ArrowRL ($\alpha = 0$)	51.4 \uparrow	59.2 \uparrow	49.8 \downarrow	63.7 \downarrow	52.3 \uparrow	55.6 \downarrow
+ ArrowRL ($\gamma = 1$)	54.9 \uparrow	60.0 \uparrow	50.8 \downarrow	68.4 \uparrow	55.4 \uparrow	58.3 \uparrow
+ SFT	49.3 \downarrow	46.2 \downarrow	53.9 \uparrow	66.1 \downarrow	55.3 \uparrow	57.4 \uparrow
+ ArrowRL	51.4 \uparrow	54.8 \uparrow	55.6 \uparrow	69.6 \uparrow	58.8 \uparrow	60.7 \uparrow

Frame Rate Analysis Fig. 10 presents our inference frame analysis on Vinoground, comparing ArrowRL-enhanced Qwen2.5-VL-7B against its base model across different frame settings (1-4 FPS). Crucially, although ArrowRL training utilizes a fixed 16 frames per input (for efficiency), the

performance gains provided by ArrowRL remain consistent across varying temporal granularities at inference, showcasing its generalization beyond the specific training setup, and giving evidence that the base models overlook temporal detail even with higher framerates.

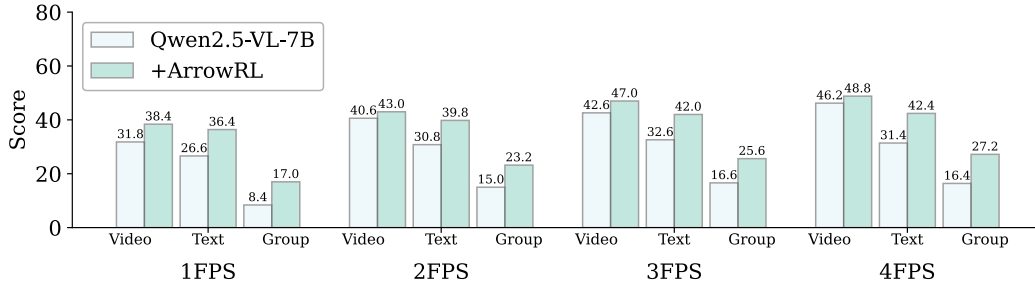


Figure 10: Impact of inference frame rate analysis on Vinoground. The ArrowRL’s performance gains remain consistent across different frame settings (1-4FPS), showcasing its generalizability.

C.2 Qualitative Results

More Qualitative Results Supplementing Fig. 7 in the main paper, Fig. 11 presents additional qualitative examples that reinforce the effectiveness of ArrowRL. These comparisons highlight how base LMMs often overlook temporal progression or direction, often relying on static cues or language biases. Conversely, our ArrowRL-enhanced models exhibit improved AoT sensitivity, leading to more accurate and temporally coherent responses across these challenging scenarios.

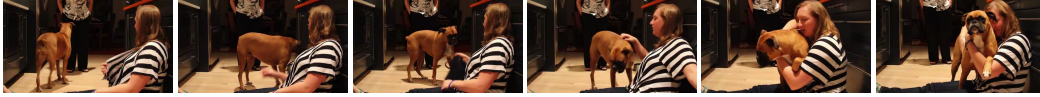
Failure Cases Fig. 12 presents one failure case on AoTBench, where neither the base LMM nor the ArrowRL-enhanced version answers correctly. The failure stems from the uniform sampling of 16 frames (6 visualized here) not capturing the key visual moments where a man in a dark suit holds a ring—a crucial cue for inferring he is about to get married. Such failures suggest potential avenues for future improvement, like advanced keyframe selection methods over uniform sampling, or the incorporation of auxiliary modalities such as audio, which in this case contains helpful cues.

D Limitations

Our construction of AoTBench relies on using selected LMMs as evaluators, meaning its sensitivity is inherently dependent on the initial AoT perception capabilities of these models. Potentially challenging samples might be missed if current evaluator models universally fail to exhibit sensitivity (i.e., yield low TDS) despite underlying temporal relevance. Nevertheless, to provide a quantitative check of AoTBench, we manually verify a small subset and find our TDS-based selection of temporally challenging examples aligns with human judgment 44 out of 50 times, suggesting reasonable concordance where models do possess baseline sensitivity.

Additionally, the reward calculation for ArrowRL utilizes an auxiliary LLM to generate similarity scores; this dependence on an LLM judge, while common [77, 21, 63], can introduce a degree of uncertainty or potential bias into the reward signal. Nevertheless, we found this approach viable for our purposes, as the LLM performs a relatively easy, straightforward text-to-text similarity assessment. We empirically find that (also illustrated in Fig. 5 of the main paper and Fig. 8) the resulting similarity scores, which underpin our reward calculation, are consistent and appear reasonable.

Furthermore, our current investigation focuses on short videos, a scope chosen due to the significant temporal challenges already evident in these scenarios for leading LMMs. We view this work as a initial step towards improving AoT understanding. Future directions include extending these methods to long videos (potentially incorporating keyframe selection) and exploring integration with temporal reasoning tasks that necessitate explicit reasoning traces (e.g. Chain-of-Thought).



Q. Why does the dog walk toward the lady after the lady reaches her hand toward it?

Qwen2-VL-7B: **to catch the food**

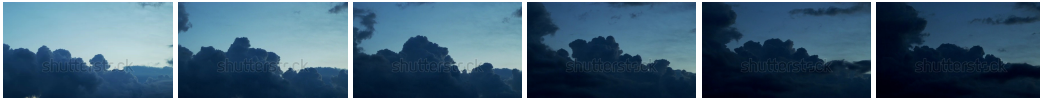
Qwen2-VL-7B + ArrowRL: **let the lady hug it**



Q. Which caption best describes this video?

Qwen2.5-VL-7B: A woman is speaking while she is handling an air filter with both hands. Using her right hand to hold in place the filter device, the left hand **pulls the filter out of its place in the device**. Both hands then pick up the outer shield of the filter from the table.

Qwen2.5-VL-7B + ArrowRL: A woman is speaking while she is handling an an air filter with both hands. Using her right hand to hold in place the filter device, the left hand **pushes the filter into its place in the device**. Both hands then pick up the outer shield of the filter from the table.



Q. Provide a one-sentence description of the video, focusing solely on the primary actions taking place.

LLaVA-OV-7B: The video captures the serene and dynamic movement of clouds against a tranquil blue sky, creating a peaceful and picturesque scene.

LLaVA-OV-7B + ArrowRL: The video features a series of images capturing the movement of clouds in the sky, with the clouds **transitioning from lighter to darker shades as the light diminishes**.

Figure 11: Additional qualitative examples comparing base LMMs with their ArrowRL-enhanced counterparts. ArrowRL enables models to succeed on AoT-sensitive VQA and produce temporally coherent captions, while base LMMs often struggle with understanding temporal progression.



Q. According to the video, who is about to get married?

A. The man in the dark suit.

Qwen2.5-VL-7B: It is unclear

Qwen2.5-VL-7B + ArrowRL (ours): It is unclear.

Figure 12: A failure case from AoTBench. Uniformly sampling 16 frames (6 visualized) fails to capture the critical visual moments (i.e., the man in the dark suit holding a wedding ring).

456 E Societal Impacts

457 This work focuses on improving a fundamental aspect of temporal perception (AoT sensitivity) in
 458 LMMs. Positive impacts stem from creating more reliable and rational AI systems. Better AoT
 459 perception can lead to LMMs with more accurate internal world models, improving their utility
 460 in tasks requiring understanding of processes, procedures, or event timelines. This could benefit
 461 applications like education, assistive technology and robotics. The primary risks are associated with
 462 the general capabilities of advanced LMMs rather than AoT sensitivity specifically. Any improvement
 463 could potentially be misused if integrated into systems for generating disinformation or invasive
 464 surveillance, though our method doesn't directly enable these.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. Test of time: Instilling video-language models with a sense of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2503–2516, 2023.
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibong Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [4] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2917–2927, 2022.
- [5] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*, 2024.
- [6] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.
- [7] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.
- [8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.
- [9] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.
- [10] Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Lu Qiu, Ying Shan, and Xihui Liu. Exploring the effect of reinforcement learning on video understanding: Insights from seed-bench-r1. *arXiv preprint arXiv:2503.24376*, 2025.
- [11] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025.
- [12] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024.
- [13] Michael Dorkenwald, Fanyi Xiao, Biagio Brattoli, Joseph Tighe, and Davide Modolo. Scvrl: Shuffled contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4132–4141, 2022.
- [14] Yang Du, Yuqi Liu, and Qin Jin. Reversed in time: A novel temporal-emphasized benchmark for cross-modal video-text retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5260–5269, 2024.
- [15] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1801–1810, 2019.

- [16] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.
- [17] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- [18] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.
- [19] Amir Ghodrati, Efstratios Gavves, and Cees GM Snoek. Video time: Properties, encoders and evaluation. *arXiv preprint arXiv:1807.06980*, 2018.
- [20] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [21] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.
- [22] Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.
- [23] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.
- [24] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhit-ing Hu. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*, 2023.
- [25] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018.
- [26] David Layzer. The arrow of time. *Scientific American*, 233(6):56–69, 1975.
- [27] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 667–676, 2017.
- [28] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022.
- [29] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [30] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024.
- [31] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.
- [32] Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, pages 331–348. Springer, 2024.

- [33] Yun Li, Zhe Liu, Yajing Kong, Guangrui Li, Jiyuan Zhang, Chao Bian, Feng Liu, Lina Yao, and Zhenbang Sun. Exploring the role of explicit temporal modeling in multimodal large language models for video understanding. *arXiv preprint arXiv:2501.16786*, 2025.
- [34] Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodal large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, 2024.
- [35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [37] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.
- [38] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.
- [39] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.
- [40] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. *arXiv preprint arXiv:2308.10901*, 2023.
- [41] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 527–544. Springer, 2016.
- [42] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [43] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023.
- [44] Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T Freeman. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2035–2042, 2014.
- [45] Will Price and Dima Damen. Retro-actions: Learning ‘close’ by time-reversing ‘open’ videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.
- [46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmlR, 2021.
- [47] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

- [48] Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 535–544, 2021.
- [49] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [50] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [51] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.
- [52] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- [53] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.
- [54] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024.
- [55] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [56] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohalluciner: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024.
- [57] Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. Paxion: Patching action knowledge in video-language foundation models. *Advances in Neural Information Processing Systems*, 36:20729–20749, 2023.
- [58] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8052–8060, 2018.
- [59] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [60] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.
- [61] Junbin Xiao, Nanxin Huang, Hangyu Qin, Dongyang Li, Yicong Li, Fengbin Zhu, Zhulin Tao, Jianxing Yu, Liang Lin, Tat-Seng Chua, et al. Videoqa in the era of llms: An empirical study. *International Journal of Computer Vision*, pages 1–24, 2025.
- [62] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.
- [63] Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2reward: Reward shaping with language models for reinforcement learning. *arXiv preprint arXiv:2309.11489*, 2023.

- [64] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.
- [65] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024.
- [66] Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36:53688–53710, 2023.
- [67] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- [68] En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Zining Zhu, Haoran Wei, Jianjian Sun, Zheng Ge, Xiangyu Zhang, Jingyu Wang, et al. Unhackable temporal rewarding for scalable video mllms. *arXiv preprint arXiv:2502.12081*, 2025.
- [69] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.
- [70] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024.
- [71] Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing llms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024.
- [72] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.
- [73] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.
- [74] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024.
- [75] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*, 2024.
- [76] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.
- [77] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.
- [78] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [79] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

701 [80] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch,
702 Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of
703 video understanding in large multimodal models. *arXiv preprint arXiv:2412.10360*, 2024.