Video-as-Answer: Predict and Generate Next Video Event with Joint-GRPO

Supplementary Material

This Appendix is organized as follows:

- Section A provides the dataset details.
- Section B offers an intuitive illustration of the Joint-GRPO reward design, along with training details.
- Section C describes the implementation details of models.
- Section D reports additional experimental results.

A. Details of VANS-Data-100K

Table 1 demonstrates the composition and key statistics of our VANS-Data-100K dataset. It contains a total of 100K samples, with 30K dedicated to procedural tasks and 70K to predictive scenarios. These are sourced from a diverse set of publicly available video datasets and Internet to ensure broad coverage of real-world dynamics and instructional content.

In terms of video characteristics, the input videos have an average duration of 9.43 seconds, providing sufficient context for event reasoning. The corresponding target videos, which depict the predicted next event, average 3.76 seconds in length, ensuring concise and focused demonstrations.

Table 1. Statistics of VANS-Data-100K dataset.

Component	Size/Duration
Data Composition	
Procedural (Total: 30K)	
YouCook2 [9]	9K
COIN [7]	21K
Predictive (Total: 70K)	
Video-Holmes [2]	10K
ActivityNet [1]	20K
V1-33K [5]	10K
YouTube Videos	30K
Video Duration (Avg.)	
Input Video	9.43s
Target Video	3.76s

B. Details of Joint-GRPO

B.1. Reward Design

Figure 1 provides an intuitive illustration of how the individual reward components work in concert within our two-stage training process of Joint-GRPO.

In Stage 1 (VLM Tuning), we examine the role of the text fidelity reward (r_{t1}) and the video fidelity reward (r_{v1}). For the provided example, if only r_{t1} is used, Sample 2 receives a high score comparable to the Ground Truth (GT), as both captions correctly describe the action. However, Sample 2 exhibits poor visual consistency. Conversely, if only r_{v1} is used, both Sample 1 and Sample 2 receive similarly low scores, failing to reflect that Sample 1 is semantically worse due to an incorrect action prediction. Only when both r_{t1} and r_{v1} are combined does the composite reward correctly rank the samples, successfully identifying the GT as the best.

In Stage 2 (VDM Adaptation), we analyze the video fidelity reward (r_{v2}) and the semantic consistency reward (r_{c2}). Relying solely on r_{v2} results in Sample 1 receiving a score similar to the GT, even though Sample 1 depicts an incorrect semantic action (it should show two people pointing guns at each other). Using only r_{c2} causes Sample 2 to be scored similarly to the GT, despite its poor visual consistency. The joint reward effectively combines these signals to prioritize samples that are correct in both semantics and visual quality.

The final combined reward in each stage is a sum of the normalized individual rewards. All weighting coefficients (λ) are set to 1, assigning equal importance to each objective. This design ensures a balanced optimization towards captions that are both semantically accurate and visually plausible (Stage 1), and videos that are both high-quality and semantically faithful (Stage 2).

B.2. Training Process

Figure 2 illustrates the training dynamics of Joint-GRPO. In Stage 1 (VLM Tuning), the format reward (r_f) in Figure 2(a) quickly saturates, indicating rapid adoption of the instruction template. Both text fidelity (r_{t1}, Figure 2b) and video fidelity (r_{v1}, Figure 2c) rewards show progressive improvement, reflecting the VLM's learning to generate captions that are both semantically accurate and visually plausible. The combined reward (Figure 2d) stabilizes after approximately 600 steps, demonstrating effective optimization. Concurrently, the increasing thinking length (Figure 2e) indicates more detailed reasoning chains.

In Stage 2 (VDM Adaptation), both video fidelity (r_{v2}, Figure 2f) and semantic alignment (r_{c2}, Figure 2g) rewards improve consistently, with convergence occurring after about 1000 steps. This demonstrates the VDM's successful adaptation to generate videos that preserve visual consistency while faithfully rendering the semantically-grounded captions from Stage 1. The total reward (Figure 2h) also reaches a stable level, confirming effective cross-modal alignment.

Collectively, these training curves validate the effectiveness of our Joint-GRPO design, demonstrating coordinated improvement across both stages.

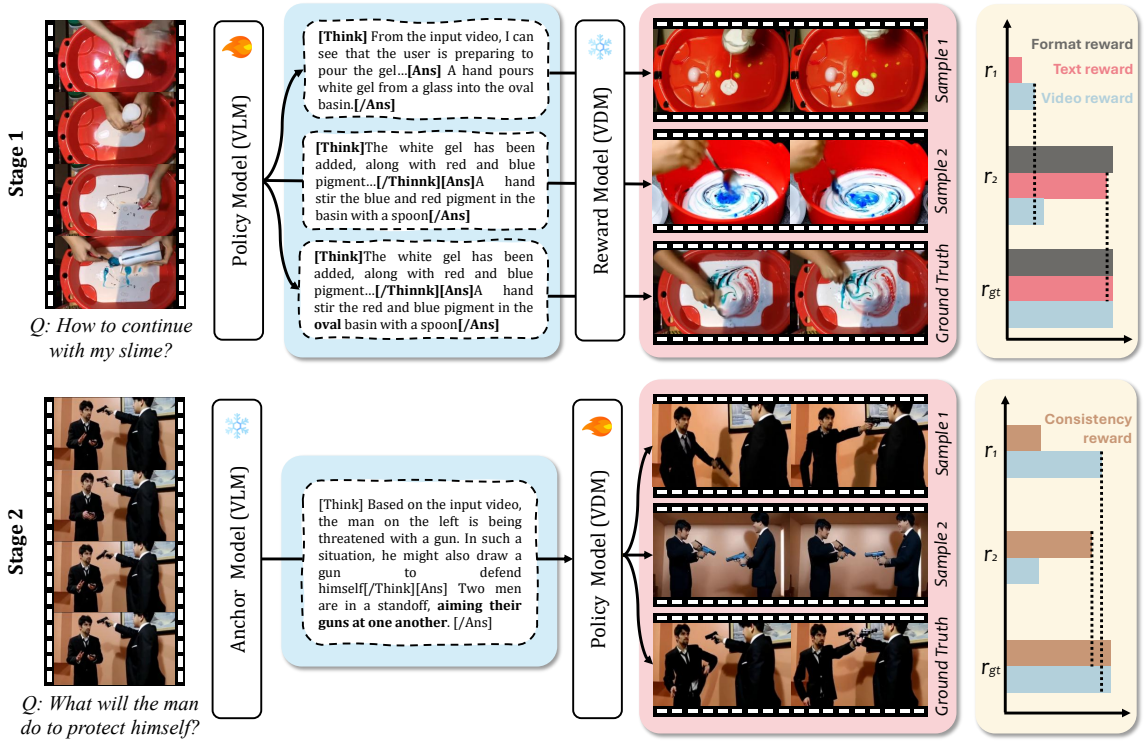
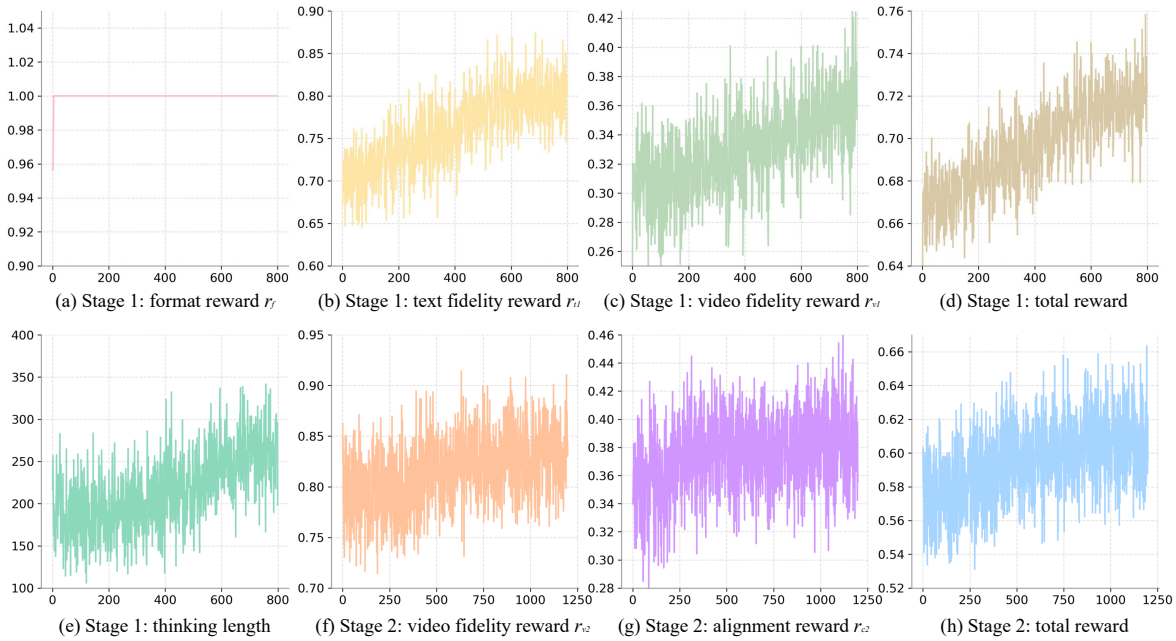


Figure 1. Illustration of our Joint-GRPO reward design. Top: For a Stage-1 case, we simulate three reasoning samples during GRPO training. The text-only reward fails to penalize Sample 2's visual inconsistency, while the video-only reward fails to penalize Sample 1's semantic error. Bottom: For a Stage-2 case, the video-only reward fails to penalize Sample 1's semantic inaccuracy, while the consistency-only reward fails to penalize Sample 2's visual inconsistency.



(a) Stage 1: format reward r_f (b) Stage 1: text fidelity reward r_{t1} (c) Stage 1: video fidelity reward r_{v1} (d) Stage 1: total reward

(e) Stage 1: thinking length (f) Stage 2: video fidelity reward r_{v2} (g) Stage 2: alignment reward r_{c2} (h) Stage 2: total reward

Figure 2. Training curves of Joint-GRPO: (a) format reward (r_f) in Stage 1; (b) text fidelity reward (r_{t1}) in Stage 1; (c) video fidelity reward (r_{v1}) in Stage 1; (d) total reward in Stage 1; (e) thinking length evolution in Stage 1; (f) video fidelity reward (r_{v2}) in Stage 2; (g) semantic alignment reward (r_{c2}) in Stage 2; (h) total reward in Stage 2.

C. Implementation Details

C.1. Training of VANS

We initialize VANS with Qwen2.5-VL-3B as the VLM and Wan-2.1-1.3B as the VDM. The VDM is configured to use $n = 6$ reference frames.

In the SFT stage, the VLM is trained for 10K steps using LoRA [3] (rank=8, alpha=32) with a learning rate of 5×10^{-5}, while the VDM is fully fine-tuned for 20K steps across all DiT blocks with the same learning rate of 5×10^{-5}.

For Joint-GRPO post-training, Stage 1 is optimized for 800 steps with a learning rate of 5×10^{-5}. In Stage 2, to ensure the quality of anchor captions, we filter out those with ROUGE-L scores below 0.6 before proceeding with VDM adaptation. Stage 2 is then trained for 1K steps with the same learning rate of 5×10^{-5}. We equip the VLM with LoRA (rank=8, alpha=32). For the VDM, we adopt the method from [4] to convert a deterministic Ordinary Differential Equation (ODE) into an equivalent Stochastic Differential Equation (SDE) to enable GRPO training. We set the KL coefficient $\beta = 0.004$, clip range to 1×10^{-3}, and sample group size to 8 per prompt.

C.2. Evaluation

Evaluation Protocol. All methods are evaluated under a unified protocol. For Video-GPT, we provide only the input video and utilize its native video continuation capability. For VANS and other baseline methods, we provide the input video along with the corresponding question and the following system prompt:

> You will be given a video. Your task is to predict the next event based on the input video and the user's instructions. Please begin by providing your detailed reasoning between the [Think][/Think] tags, followed by your detailed description of the next event within the [Ans][/Ans] tags.

Input Adaptation. To accommodate different model architectures, we adapt the video input accordingly: for models that can directly process video input (e.g., Gemini), we provide the original video; for other models (e.g., Qwen), we set the input video fps = 1 for their video processors.

Output Specification. All methods are required to generate a video answer with a resolution of 352×640 and a length of 33 frames to ensure consistent and fair comparison.

Metric Computation. The CLIP-Score for video consistency (CLIP-V) and semantic consistency (CLIP-T) is computed using a ViT-B/32 model. Specifically, each frame of the generated video is compared with the corresponding frame in the ground-truth video, and the scores are averaged across all frames.

Table 2. Results on procedural VNEP. The comparison with fine-tuned baselines (*) shows that our architectural design, rather than data advantage, is the primary source of improvement.

Model	BELU@4↑	ROUGE-L↑	FVD↓	CLIP-V↑	CLIP-T↑
Qwen-Wan	0.0013	0.1530	148.75	0.6619	0.2448
Qwen*-Wan	0.0233	0.2812	140.32	0.6790	0.2466
Qwen*-Wan*	0.0233	0.2812	140.07	0.6795	0.2470
Gemini-Wan	0.0215	0.2802	120.34	0.6898	0.2547
VANS (SFT)	0.0233	0.2812	85.34	0.7655	0.3202
VANS (Joint-GRPO)	**0.0987**	**0.3631**	**78.32**	**0.8021**	**0.3824**

D. Additional Results

D.1. Inference Time

The inference time of VANS is comparable to other cascaded pipelines, requiring approximately 4 seconds for caption generation and 35 seconds for video generation using the official VAN library. In contrast, unified models exhibit longer inference times: Omni-Video requires approximately 50 seconds, while VideoGPT needs about 60 seconds for complete generation.

D.2. Comparison with Fine-tuned Baseline

To analyze the source of performance improvements in VANS, we compare it with fine-tuned baselines. As shown in Table 2, the results indicate three main observations: data quality provides a foundation, architectural modification contributes to noticeable gains, and Joint-GRPO provides the decisive enhancement that pushes performance to the state-of-the-art level.

Data Quality as the Foundation. When fine-tuned on our VANS-Data-100K for 10K steps (denoted as Qwen*), the model achieves reasoning capability competitive with Gemini-2.5-Flash in a zero-shot setting (ROUGE-L: 0.2812 vs. 0.2802). This confirms that our high-quality dataset enables smaller models to learn sophisticated reasoning.

Architectural Modification Contributes to Gains. Fine-tuning both components of the Qwen-Wan pipeline (denoted as Qwen*-Wan*) yields limited video metric improvements over the base fine-tuned VLM (denoted as Qwen*-Wan). In contrast, VANS (SFT) with the same text input achieves better video results: FVD decreases from 140.07 to 85.34 and CLIP-V increases from 0.6795 to 0.7655, suggesting the proposed VAE reference feature aids visual consistency.

Joint-GRPO Delivers the Decisive Enhancement. The most striking improvement comes from Joint-GRPO, which elevates VANS to unprecedented performance levels across all metrics. Compared to VANS (SFT), Joint-GRPO boosts ROUGE-L from 0.2812 to 0.3631 (29.1% relative improvement) and CLIP-T from 0.3202 to 0.3824 (19.4% relative improvement), while further reducing FVD to 78.32. These results unequivocally demonstrate that Joint-GRPO is the

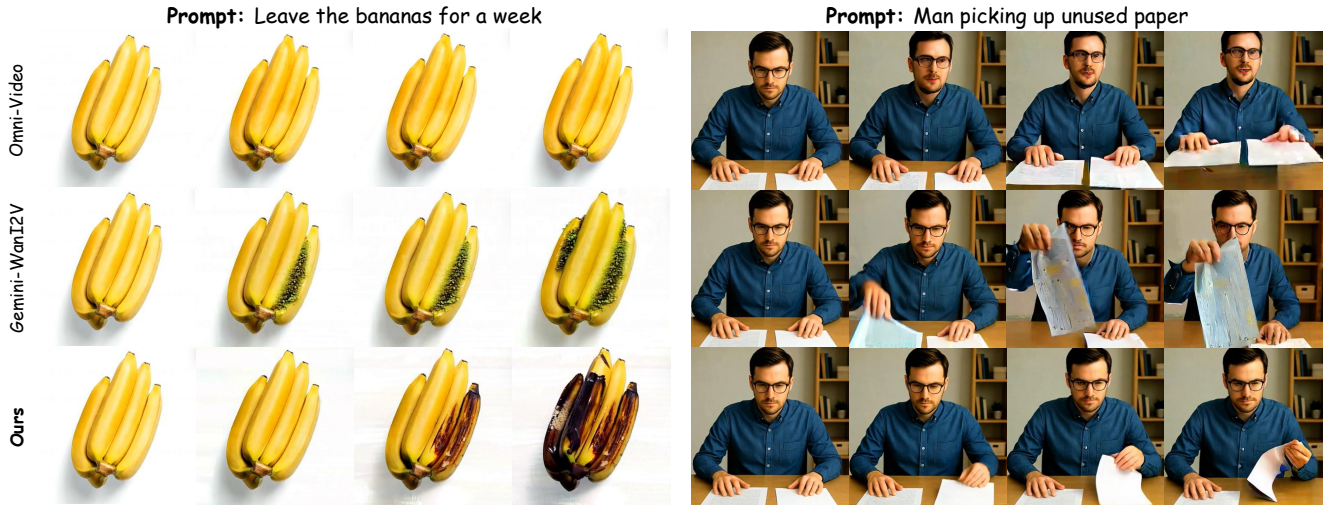


Figure 3. Visual comparison results on UI2V-Bench.

Figure 4. Multi-Future Prediction Results.

most critical component for achieving state-of-the-art performance, effectively aligning both textual and visual outputs with human preferences.

D.3. Generalization

Multi-Future Prediction. The established NEP task typically assumes a single, causal progression from the input context. In contrast, our VANS demonstrates a key generalization capability: multi-future prediction. This allows the model to generate semantically distinct and contextually appropriate video answers based on different hypothetical questions applied to the same input video, moving beyond deterministic continuation.

As shown in Figure 4, when presented with a scene of a woman reacting to a hot object, VANS can generate fundamentally different yet plausible outcomes conditioned on the scenario: in a realistic everyday context, it predicts a natural reaction of "coughing"; whereas in a stylized cinematic context, it visualizes a dramatic effect of "smoke exhaling from the mouth". This flexibility stems from our model's ability to ground its predictions in both the visual evidence and the diverse textual hypotheses provided, effectively exploring multiple potential futures from a single starting point.

Reasoning Image-to-Video Generation. VANS generalizes effectively to the reasoning image-to-video (I2V) task by treating a single image as a static video clip. This generalization capability is attributed to the model's training on mixed datasets including Koala-36M [6] for I2V tasks. Figure 3 demonstrates examples from UI2V-Bench [8], when given an image of a banana and the instruction "leave the banana for a week," our model accurately predicts the temporal evolution, generating a video where the banana skin darkens. In contrast, other strong baselines struggle to capture this causal-physical transformation correctly. This demonstrates the robustness of our approach in understanding static visual contexts and reasoning about their potential dynamic futures.

D.4. Human Evaluation

To complement automatic metrics, we conduct a human evaluation to assess the subjective quality of generated video answers. We recruit 30 evaluators (mean age = 25 years; all hold at least a bachelor's degree, including 20 postgraduate/PhD students and 10 full-time professionals). Each evaluator is presented with 20 randomly selected examples

Table 3. Human evaluation results (scale: 1-5). Our VANS with Joint-GRPO achieves the highest scores across all criteria.

Model	Semantic Correctness	Visual Consistency	Overall
Video-GPT	1.5	3.6	1.5
Omni-Video	2.1	3.2	2.2
Gemini-FilmWeaver	3.9	3.1	3.5
VANS (SFT)	3.8	3.9	3.7
VANS (Joint-GRPO)	**4.7**	**4.6**	**4.8**

(10 procedural and 10 predictive) and rates the results on three dimensions: semantic correctness, visual consistency, and overall satisfaction.

The results in Table 3 reveal that VANS (SFT) achieves semantic correctness comparable to the strong baseline Gemini-FilmWeaver, while demonstrating superior visual consistency. Furthermore, VANS with Joint-GRPO receives the highest ratings across all three criteria, indicating that our full approach yields video answers that are not only semantically and visually accurate but also subjectively more satisfactory to human observers.

D.5. Video Results

All video results corresponding to the figures in this paper, along with additional examples, are provided in our project page: https://video-as-answer.github.io.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the ieee conference on computer vision and pattern recognition, pages 961–970, 2015. 1

[2] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? arXiv preprint arXiv:2505.21374, 2025. 1

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. ICLR, 1(2):3, 2022. 3

[4] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. arXiv preprint arXiv:2505.05470, 2025. 3

[5] Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. Fostering video reasoning via next-event prediction. arXiv preprint arXiv:2505.22457, 2025. 1

[6] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content, 2024. 4

[7] Tang Yansong, Ding Dajun, Rao Yongming, Zheng Yu, Zhang Danyang, Zhao Lili, Lu Jiwen, and Zhou Jie. Coin: A large-scale dataset for comprehensive instructional video analysis. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 1

[8] Ailing Zhang, Lina Lei, Dehong Kong, Zhixin Wang, Jiaqi Xu, Fenglong Song, Chun-Le Guo, Chang Liu, Fan Li, and Jie Chen. Ui2v-bench: An understanding-based image-to-video generation benchmark. arXiv preprint arXiv:2509.24427, 2025. 4

[9] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In Proceedings of the AAAI conference on artificial intelligence, 2018. 1