RETRIEVAL, REFINEMENT, AND RANKING FOR TEXT-TO-VIDEO GENERATION VIA PROMPT OPTIMIZATION AND TEST-TIME SCALING

Zillur Rahman *
Algoverse AI

Alex Sheng
Algoverse AI

Cristian Meo
Algoverse AI

ABSTRACT

While large-scale datasets have driven significant progress in Text-to-Video (T2V) generative models, these models remain highly sensitive to input prompts, demonstrating that prompt design is critical to generation quality. Current methods for improving video output often fall short: they either depend on complex, post-editing models, risking the introduction of artifacts, or require expensive fine-tuning of the core generator, which severely limits both scalability and accessibility. In this work, we introduce 3R, a novel RAG based prompt optimization framework. 3R utilizes the power of current state-of-the-art T2V diffusion model and vision language model. It can be used with any T2V model without any kind of model training. The framework leverages three key strategies: RAG-based modifiers extraction for enriched contextual grounding, diffusion-based Preference Optimization for aligning outputs with human preferences, and temporal frame interpolation for producing temporally consistent visual contents. Together, these components enable more accurate, efficient, and contextually aligned text-to-video generation. Experimental results demonstrate the efficacy of 3R in enhancing the static fidelity and dynamic coherence of generated videos, underscoring the importance of optimizing user prompts.

1 INTRODUCTION

Due to exciting advancements in diffusion-based generative models and large-scale training procedures, modern text-to-video (T2V) models have achieved impressive capabilities for using natural language prompts to generate photorealistic video content (Peebles & Xie, 2023) (Ramesh et al., 2022).

Despite rapid uptake in natural language processing (NLP) and subsequent innovations in text-to-image (T2I) generation Podell et al. (2023); Esser et al. (2024), and improving image aesthetics (Chen et al., 2024a), their impact on video quality is limited (Hao et al., 2023). Besides, the application of test-time optimization (Zhang et al., 2025b) in text-to-video settings remains in early exploratory stages, with meaningful open challenges (Gu et al., 2025). This presents valuable opportunities to address problems in T2V like prompt adherence, visual quality, physical plausibility, and temporal coherence.

To generate videos from texts, users provide a short text prompt to the video generation model. Recent works show that a long detailed text prompt generates better quality videos than the short user provided prompt (Hao et al., 2023; Yang et al., 2025b). This underscores the importance of enhancing the user prompt before feeding it to a T2V model. The short user prompts do not contain detailed contextual information required to generate vivid visual content. Moreover, videos generated from the same prompt differ in quality due to the stochastic nature of the diffusion models. So generating multiple videos from one prompt and selecting the one that better fits the user prompt with better visual quality could be effective.

To address these, in this paper, we explore avenues combining retrieval, refinement, and ranking within this emerging paradigm of inference-time compute algorithms to improve video generation

*Correspondance to: zillur.mle@gmail.com

quality in T2V settings. We study a black-box problem definition that is designed for direct plug-and-play applicability to off-the-shelf T2V models in real-world settings.

Our contributions can be summarized as follows:

- We propose Retrieval-Refinement-Ranking (3R), a retrieval based training free prompt optimization framework for T2V generation.
- We propose an initial prompt refinement module that creates a detailed context rich description aligning with the user prompt.
- We validate the effectiveness of 3R on EvalCrafter benchmark where it achieves SOTA results among open-source models.

2 Related Works

Text-to-Video Models Text-to-video generation models (OpenAI, 2024; Rombach et al., 2021; Wang et al., 2024; Zhang et al., 2025a) have seen rapid advances in both model capabilities and practical accessibility. T2V models receive input prompts consisting of natural language descriptions, and comprehend described scenes, actions, objects and generate visual contents. T2V models are being used in generating animations (Chen et al., 2023), movies (Zhao et al., 2025), commercials, etc.

Prompt Optimization Frameworks IPO (Yang et al., 2025a) introduces an iterative optimization algorithm to align video foundation models with human preferences. It creates a human preference dataset and trains a critique model with that dataset. An iterative optimization loop is used to align a base T2V model with human preference, and thus improving subject consistency, motion smoothness, and aesthetic quality. CCC (Gu et al., 2025) introduces a simple vision language model for text to video generation. Each candidate video is queried multiple times to get a list of issues and a content score is computed from the number of common issues. Based on those issues, initial prompts are refined to generate better results. In (Gao et al., 2025), authors introduced RAPO: a RAG based prompt optimization model for text to video generation. A dataset is used to extract relevant modifiers to augment the user prompt. Then a fine-tuned Llama3 model (Grattafiori et al., 2024) is used to refactor the augmented prompt into the format of training prompts. Finally, another fine-tuned LLama3 is used to select the better prompt between the refactored prompt and a refined user prompt. Google publishes VISTA (Long et al., 2025), one of the most computationally expensive models where pairwise video comparison is used to select candidate videos that are evaluated by multi-modal language models for their visual, audio, and context quality. Then, LLMs review the issues and refine the original prompts to generate videos again. The entire model runs for maximum 5 iterations and each iteration can have maximum 30 videos, making it 150 videos per prompt.

3 Method

This section discusses the key design choices of 3R. The overall pipeline is illustrated in Fig. 1 and a pseudo algorithm is illustrated in Appendix A.1.

Modifiers Using synthetic data augmentation to rearrange knowledge for more data-efficient learning has been proven an effective pathway to mitigate the challenge of adapting a pre-trained model to a small corpus of domain-specific documents (Yang et al., 2024). Its main purpose is to overcome the model's context limitations, enabling effective context construction for diverse user queries. Given an original user prompt intent I, first, scene modifiers p_j are extracted from a relation database $\mathcal{D}$ using a pre-trained sentence transformer. Then using cosine similarity score, we select scenes from the relation graph if it is above a threshold τ. Each scene comes with its list of subject, action and environment modifiers. We select top-k modifiers for each scene.

$$P_{ret} = \{p_j \in \mathcal{D} \mid \text{sim}(\phi(I), \phi(p_j)) > \tau\} \tag{1}$$

We iteratively merge each modifier with I using a pre-trained LLM M_{LLM} in a few-shot manner. Existing method like RAPO (Gao et al., 2025) uses a large comma separated list of all the modifiers as an initial description and prompts the LLM to merge each modifier. However, this process sometimes generates incorrect and misleading results since some modifiers have little to no relevance to

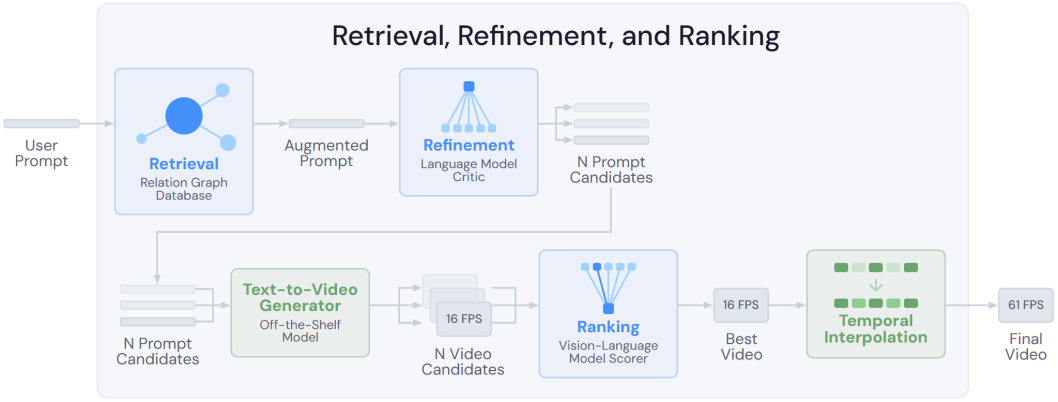

Figure 1: **Overview of 3R pipeline.** A short user prompt I is used to extract a few relevant subject, scene, actions modifiers from a relation database $\mathcal{D}$. Then M_{LLM} is used to merge those modifiers iteratively to the original user prompt to get detaild prompt P_m, and R_{LLM} checks P_m for any contradictory or missing information from the original prompt I, and generate N refined prompts. The refined prompts are fed to a T2V base model $\mathcal{G}$ to generate initial videos for each prompt. Next, a video selection model selects the best candidate based on a question answering test, and a temporal interpolation network enhances temporal consistency of the final video.

the I. To mitigate this issue, we initialize the description with only I. We show such an example in Appendix Table 3.

$$P_m = M_{LLM}(I \mid P_{ret}) \tag{2}$$

Refine Descriptions After merging is completed, we get a detailed description P_m of each user prompt. To eliminate any contradictory, misleading information or add any useful missing information, we prompt an LLM to further refine the description. This LLM aims to refine P_m based on information such as characters, actions, attributes like color, counts from I. This step is crucial for quality video generation. In our experiment section, we demonstrate the importance of the initial prompt. If the information in the initial prompt is not coherent, the generated video will not represent user intents. We use R_{LLM} in a few-shot manner and generate N distinct detailed prompts, maintaining the original user intent. The goal is to generate multiple videos that may have different positive and negative aspects so that we can choose the best candidate as the final video. The prompt for this step is illustrated in Appendix C.1.

$$\{P_n\}_{n=1}^{N} = R_{LLM}(I \mid P_m) \tag{3}$$

We explain the video selection and enhancement sections in details in Appendix A.2.

4 Experiments

Experimental Setup We use the EvalCrafter (Liu et al., 2023) benchmark for quantitative performance evaluation. This comprehensive text2video evaluation benchmark has 17 raw dimensions such as clip score, motion score, face consistency score, etc. These raw metrics are aggregated into 4 categories: Text-Video Alignment, Visual Quality, Motion Quality and Temporal Consistency. We compare our approach to 4 other models that reported their performance on EvalCrafter benchmark: Lavie (Wang et al., 2024) with original short prompts, IPO (Yang et al., 2025a), Show-1 (Zhang et al., 2025a) and Videocrafter2 (Chen et al., 2024b). We reproduced the IPO results, and for others, we use the results and metrics reported in the EvalCrafter benchmark leader-board. We report the implementation details in Appendix B.1.

Table 1: Results on EvalCrafter benchmark. The **first** and second best results in each column are highlighted in the corresponding colors. 3R achieves the best total result, and either best or second best results in most of the individual metrics.

Model	Total Score	Motion Quality	Text-Video Alignment	Visual Quality	Temporal Consistency
Show-1	229	53.74	62.07	52.19	60.83
LaVie	234	52.83	68.49	57.99	54.23
IPO	234	53.39	54.62	62.56	**63.40**
Videocrafter2	243	**54.82**	63.16	**63.98**	61.46
3R	**245**	54.72	**68.73**	58.79	62.65

4.1 Results

Table 1 reports the results of 3R in comparison with Show-1, LaVie, IPO, and Videocrafter2 on the four benchmark metrics of EvalCrafter. 3R approach achieves the highest total score, demonstrating the effectiveness of our inference-time approach for improving text-to-video performance. Compared with the LaVie text-to-video base model without additional inference-time processing, our approach (implemented with LaVie as the base model) demonstrates a consistently higher score on all four EvalCrafter metrics, showing general and direct performance lifts across multiple facets of video generation output quality contributed by the addition of our inference-time approach. This result would be consistent with the assumption that increasing compute at inference time can be used to improve output quality with an unchanged base model. We report the qualitative performance of 3R in Appendix B.2. Fig. 2 and Fig. 3 demonstrate 3R's superior text-video alignment performance in complex prompts such as mushroom growing out of a human head. 3R can also visualize fictional characters such as Pikachu Jedi and understand the meaning of close-up or zoom-in better.

Importance of the Initial Prompt Augmentation. As shown in Table 4 (row 2), incorporating RAG-based prompt augmentation significantly improves performance. The total score increases by +7, motion quality improves by +2, and temporal consistency improves by +4, with only a slight decrease of 1 point in text-video alignment. These results underscore the importance of high quality initial prompts, suggesting that the limitations of the base model are often rooted in underspecification in user prompts rather than architectural incapacity.

Effectiveness of Increasing Test-Time Compute. Our results in Table 4 demonstrate that increasing test-time compute is a highly effective training-free strategy to close the performance gap between base models and state-of-the-art video generators. By shifting the burden from model parameters to inference-time logic, specifically through LLM-based prompt refinement, multiple-candidate sampling for video selection, and temporal interpolation, we observed a cumulative increase in the total score from 234 to 245. We report the details of other experiments in Appendix B.3.

5 Conclusion

In this paper, we propose 3R, a novel framework for prompt optimization to improve the quality of T2V generated videos. We show that inference-time augmentation of a text-to-video model with retrieval, refinement, and ranking elements leads to performance gains in aggregate scores combining video generation metrics like motion quality, text-video alignment, visual quality, and temporal consistency. Our results contribute to a better understanding of the different inference-time pathways for improving output quality when using text-to-video models in a black-box setting.

Despite these gains, the 3R pipeline introduces increased inference latency due to multiple-candidate sampling and dense temporal interpolation. Furthermore, our ablation study highlights a critical "feedback bottleneck": contemporary vision-language models (VLMs) often provide overcorrective or semantically drifted critiques. Future research will explore more efficient sampling strategies and video-critique architectures that provide more grounded feedback, potentially enabling a truly iterative "generate-and-verify" loop that avoids the pitfalls of current VLM overcorrection.

REFERENCES

Chieh-Yun Chen, Chiang Tseng, Li-Wu Tsao, and Hong-Han Shuai. A cat is a cat (not a dog!): Unraveling information mix-ups in text-to-image encoders through causal analysis and embedding optimization, 2024a. URL https://arxiv.org/abs/2410.00321.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024b. URL https://arxiv.org/abs/2401.09047.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction, 2023. URL https://arxiv.org/abs/2310.20700.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Bingjie Gao, Xinyu Gao, Xiaoxue Wu, Yujie Zhou, Yu Qiao, Li Niu, Xinyuan Chen, and Yaohui Wang. The devil is in the prompts: Retrieval-augmented prompt optimization for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3173–3183, June 2025.

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.

Jing Gu, Ashwin Nagarajan, Tejas Polu, Kaizhi Zheng, Ruijian Zha, Jie Yang, and Xin Eric Wang. CCC: Enhancing video generation via structured MLLM feedback. In *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! at ICML 2025*, 2025. URL https://openreview.net/forum?id=B4eaJfAbCP.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation, 2023. URL https://arxiv.org/abs/2212.09611.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023.

Do Xuan Long, Xingchen Wan, Hootan Nakhost, Chen-Yu Lee, Tomas Pfister, and Sercan Ö. Arık. Vista: A test-time self-improving video generation agent, 2025. URL https://arxiv.org/abs/2510.15831.

OpenAI. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, February 2024. Accessed: 2026-01-06.

OpenAI et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Sungheon Park, Minjung Son, Seokhwan Jang, Young Chun Ahn, Ji-Yeon Kim, and Nahyup Kang. Temporal interpolation is all you need for dynamic neural radiance fields, 2023. URL https://arxiv.org/abs/2302.09311.

William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL https://arxiv.org/abs/2212.09748.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. URL https://arxiv.org/abs/2112.10752.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Ximeng Sun, Ryan Szeto, and Jason J. Corso. A temporally-aware interpolation network for video frame inpainting, 2018. URL https://arxiv.org/abs/1803.07218.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *IJCV*, 2024.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation, 2024. URL https://arxiv.org/abs/2412.21059.

Xiaomeng Yang, Zhiyu Tan, and Hao Li. Ipo: Iterative preference optimization for text-to-video generation, 2025a. URL https://arxiv.org/abs/2502.02088.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025b. URL https://arxiv.org/abs/2408.06072.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. Synthetic continued pretraining, 2024. URL https://arxiv.org/abs/2409.07431.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation, 2025a. URL https://arxiv.org/abs/2309.15818.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma. A survey on test-time scaling in large language models: What, how, where, and how well?, 2025b. URL https://arxiv.org/abs/2503.24235.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence, 2025. URL https://arxiv.org/abs/2407.16655.

A 3R MODEL

A.1 ALGORITHM

Algorithm 1 3R Methodology

procedure GENERATEVIDEO($I, \mathcal{D}, N$) ▷ I: User Intent, $\mathcal{D}$: Database, N: Candidates
  /* Step 1: Retrieval */
  $\boldsymbol{e}_I \leftarrow \phi(I)$ ▷ Encode user intent using embedding function ϕ
  $P_{ret}\{(p_j \in \mathcal{D} \mid \text{cosine\_similarity}(\boldsymbol{e}_I, \phi(p_j)) > \tau\}$
  /* Step 2: Refinement & Merging */
  $M_{LLM} \leftarrow \text{LLM\_Reasoning}(I, P_{ret})$ ▷ Merge user intent with retrieved knowledge
  $\{P_1, \ldots, P_N\} \leftarrow \text{GenerateVariants}(R_{LLM}, N)$ ▷ Sample N refined prompts
  /* Step 3: Generation */
  for $n \leftarrow 1$ **to** N **do**
    $V_n \leftarrow \mathcal{G}(P_n)$ ▷ Generate candidate video using T2V model $\mathcal{G}$
  end for
  /* Step 4: Ranking */
  for $n \leftarrow 1$ **to** N **do**
    $\text{TotalScore}_n \leftarrow 0$
    for $i \leftarrow 1$ **to** 29 **do** ▷ Evaluate 29 weighted VQA questions
      $s_{i,n} \leftarrow f_{vqa}(V_n, Q_i)$
      $\text{TotalScore}_n \leftarrow \text{TotalScore}_n + (\omega_i \times s_{i,n})$
    end for
  end for
  $V^* \leftarrow V_{\arg\max}(\text{TotalScore})$ ▷ Select best candidate
  /* Step 5: Enhancement */
  $V_{final} \leftarrow \mathcal{E}(V^*)$ ▷ Apply super-resolution/smoothing $\mathcal{E}$
  return V_{final}
end procedure

A.2 3R METHOD

Video Generation Each prompt in our approach is passed to a T2V base model $\mathcal{G}$. T2V model is treated as a black box that is only assumed to take a natural language input prompt and return a generated video output. This setup preserves practical applicability, as our approach does not require access beyond black-box text-to-video queries. By adhering to this framing, our algorithm is applicable to any off-the-shelf T2V model, regardless of whether they are open-source models or proprietary inference APIs.

$$\{V_n\}_{n=1}^{N} = \{\mathcal{G}(P_n) \mid P_n \in \{P_1, \ldots, P_N\}\} \tag{4}$$

Video Selection We adopt a video selection model f_{vqa} that evaluates each generated video by asking a set of questions Q_i covering text-video alignment, motion smoothness, and visual quality. Each question is associated with a learned weight w_i that reflects how strongly it correlates with human video preferences. Some highly weighted questions are illustrated in Table 2. The highest weights are assigned to questions related to prompt alignment (e.g., whether the video satisfies all requirements of the text), physical realism (e.g., whether object motion is realistic), and fine detail quality. In contrast, questions pertaining to subjective aesthetics (e.g., whether lighting is beautiful) receive much smaller weights. Consequently, the reward model places greater emphasis on semantic correctness and physical plausibility, allowing it to reliably select the best candidate among a set of generated videos.

$$V^* = \arg\max_{V_n} \mathcal{S}(V_n) \quad \text{where} \quad \mathcal{S}(V_n) = \sum_{i=1}^{29} w_i \cdot f_{vqa}(V_n, Q_i) \tag{5}$$

Table 2: Top 5 weighted questions used in VisionReward-Video scoring out of 29 questions. Higher scores are assigned to text-video alignment questions.

Rank	Weight	Question
1	1.1418	Does the video not completely fail to meet the requirements stated in the text "[prompt]"?
2	0.9544	Does the video meet all the requirements stated in the text "[prompt]"?
3	0.4390	Is the object's movement completely realistic?
4	0.4293	Are the details very refined?
5	0.3942	Is the video content part of the physical world?

Table 3: Comparison of Prompt Optimization Outputs. RAPO contains information such as 'eating hay' or 'shooting competition' that has no relevance to the user intent 'a tranquil tableau of barn' while 3R filters out irrelevant modifiers and generates relevant ones like 'wooden structure' and 'peaceful countryside setting'.

Category	Content
User Prompt	A tranquil tableau of barn.
Retrieved Modifiers	a lively and energetic performance, walking inside, eating hay, being kept in the barn for some purpose, appears to be a farmer, a lively and joyful scene, wearing pink outfits, in a barn, preparing for a shooting competition, dancing in a barn, laying on the ground, a barn, opening the door of a barn, inside a barn
RAPO	A lively and energetic performance is taking place inside a barn. A farmer, wearing pink outfits, is walking inside, eating hay, and appears to be preparing for a shooting competition. The scene is lively and joyful, and the animals seem to be laying on the ground. The farmer opens the door of the barn and invites the audience to witness the performance, which includes dancing in the barn.
3R	In a peaceful countryside setting, a barn sits quietly surrounded by lush green fields. The wooden structure, weathered yet sturdy, is bathed in the warm glow of the afternoon sun.

Video Enhancement Previous work (Park et al., 2023; Sun et al., 2018) has used temporal interpolation techniques to learn meaningful representations for dynamic scenes. In our approach, our goal is to specifically improve motion smoothness and enrich temporal detail by extracting features through a temporal interpolation network $\mathcal{E}$. This network first duplicates current frames to match the number of target frames with randomly initialized gaussian noise. Then it uses a pre-trained UNet diffusion model (Ronneberger et al., 2015) to denoise and results in interpolated frames. Besides, it uses the user prompt to guide the interpolation process to ensure proper temporal coherence and alignment with the user intent.

$$V_{final} = \mathcal{E}(V^*) \tag{6}$$

B Experiments

B.1 Implementation Details

We use the relation graph from RAPO (Gao et al., 2025) to extract relevant and useful modifiers. As the sentence transformer, we use all-MiniLML6-v2 to get the embeddings of sentences (Wang et al., 2020). To merge the retrieved modifiers with the user prompt, we use Mistral model (Jiang et al., 2023) with cosine similarity threshold $\tau = 0.5$ between an user prompt and a modifier. In our final model, we use GPT4o (OpenAI et al., 2024) as the prompt refiner. We create 4 prompt candidates with variations by keeping the original user intent intact. As our final base text2video generation model, we use Lavie (Wang et al., 2024). We choose Lavie because it is faster than other diffusion models like Wan (Wan et al., 2025) and generates quality video using minimal resources. In our Nvidia H200 GPU, it takes around 5s to generate one video. As for the video selection model,

we use Vision-Reward Model (Xu et al., 2024). It generates scores for all 4 candidate videos using multimodal visual question answering technique. The selected video is used as input to the temporal interpolation network (Wang et al., 2024) that increases the number of frames to 61. For the VLM Critique ablation study, we use GPT4o (OpenAI et al., 2024). We run the overall pipeline two times with two random seeds and report the average results in Table 1.

B.2 Qualitative Results

In this section, we illustrate a few challenging examples from the EvalCrafter benchmark. We compare the qualitative performance of 3R with Lavie base model and IPO.

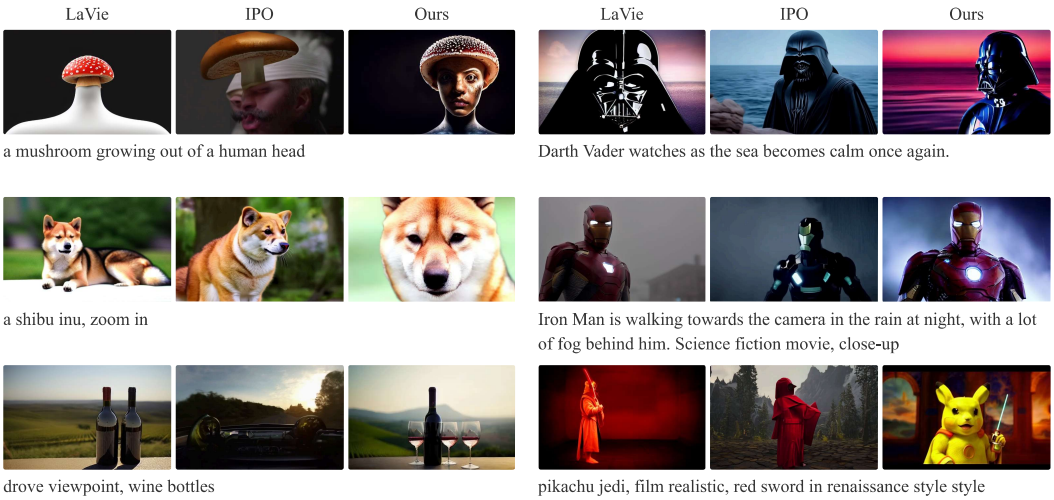

Figure 2: **Qualitative comparison of Lavie, IPO, and 3R in two common video generation failure modes.** The left side shows prompts and video frames representing challenges in semantic alignment such as mushroom growing out of human head or zoom-in and the right side shows prompts and video frames representing challenges in addressing fictional references such as Darth Vedar or Pikachu Jedi.

B.3 Ablation Study

We explain the results of other research questions in details here.

Impact of Video Selection. Table 4 (row 3) presents the impact of incorporating the video selection module. Adding a vision-based reward through diffusion-based human preference alignment increases the total score by +2, driven primarily by a +2 improvement in text-video alignment. This highlights the crucial role of preference alignment. The selection module reliably filters out semantically inconsistent generations. Example videos show that the selected outputs more faithfully reflect user intent compared to their unfiltered counterparts.

Temporal Interpolation and Consistency. Increasing the number of frames generated from 16 to 61 leads to a noticeable improvement in temporal smoothness, as reflected in the temporal consistency score. Videos with higher frame density exhibit reduced flicker, smoother motion trajectories, and fewer disjoint transitions. Offline visual comparisons clearly show the improvement in motion coherence, particularly in scenes with significant camera or object movement.

Efficacy of VLM Critque As shown in Table 4 (row 4), introducing a video critique module does not produce measurable performance gains. Inspection of the VLM-generated critique text reveals several misleading or incorrect interpretations, frequently exhibit semantic drift or unnecessary over-corrections, underscoring the unreliability of critique signals for this task. Two examples are illustrated in Appendix C.3 where VLM either tries to over-correct or the T2V model fails to follow the

Figure 3: **Qualitative comparison of approaches in the common failure mode of generating videos containing text.** We compare the first frame of the videos generated by Lavie (left), IPO (middle), and 3R (right) in the common failure mode of text generation in videos, as observed from prompts provided by the EvalCrafter benchmark. All three approaches show strong limitations in generating correct text, but 3R manages to generate qualitatively more legible text where the intended text in the prompt (”keep off the grass” or ”keep off”) can still be partially inferred despite typos. The prompts and respective video frames show how our approach can address prompts requiring multiple semantic conditions while producing less distorted outputs.

VLM instructions, resulting in degradation in video quality in both cases. The specific prompt used to extract critique data is detailed in Appendix C.2.

Overall, the ablation results confirm that each component: RAG-based augmentation, diffusion-based preference alignment, and temporal-aware interpolation, contributes meaningfully to the 3R pipeline, offering complementary improvements across evaluation dimensions.

Table 4: Ablation results. The **first** and second best results in each column are highlighted in the corresponding colors. Initial prompt refinement, video selection model, and temporal interpolation model each contribute to the total score while vision-language model critique degrades text-video alignment severely due to it’s over-correction nature.

Model	Total Score	Motion Quality	Text-Video Alignment	Visual Quality	Temporal Consistency
LaVie (Baseline)	234	52.83	68.49	57.99	54.23
One Prompt	241	54.73	67.54	59.75	58.75
N Prompts + Video Selection	243	54.64	**69.44**	59.86	58.94
Video Selection + VLM Critique	241	**54.89**	66.35	**60.20**	59.74
Video Selection + Temporal Inter.	**245**	54.72	68.73	58.79	**62.65**

C PROMPTS

In this section, we report all LLM prompts in the same format we used in this study.

C.1 USER PROMPT REFINEMENT

```
system_prompt = f"""
You are a precise and creative language model specialized in
refining scene descriptions for text-to-video generation.

Your goal:
1. **Preserve creative imagination**, including surreal or
impossible but visually interesting elements (e.g., "glass
tree", "cat and fish swimming together").
2. **Remove or rewrite illogical, impossible, or
linguistically nonsensical phrases** unless user prompt is
asking for it. (e.g., "window is singing", "barn is well-
behaved", "sea is dry", "moon shines on the sun", "a toy is
playing itself").
3. **Ensure completeness by checking missing keywords:**
Extract key subjects, objects, and actions from the user
prompt, and check if they are present and relevant in the
description. If they are missing in the description, add
them meaningfully to maintain semantic alignment. (e.g. user
prompt is "a boat and a fish" but description does not
contain any info about fish.)
4. **Extract characters**: Extracts all characters and
counts from the user prompt. (e.g. 3 persons, 2 cats, 5
birds, one dog)
5. **Preserve user prompt characters**: Ensure the
subject/object/character extracted above are included in the
refined description. If the description is missing one or
more characters, add them meaningfully. (e.g, user prompt
mentions 3 persons but descriptions mention 1 person)
6. **Preserve colors** Ensure each character color is
preserved in the final prompt (e.g red apple, yellow tree)
7. **Preserve Video Style**: Ensure video style adheres to
user prompt. (e.g if user prompt does not say "animation",
description must not mention "animation" style video.
8. Maintain temporal coherence and remove unnecessary
repetition or contradictions.
9. Produce 4 clean, vivid, logically consistent scene
description candidates suitable for text-to-video models,
word limit: 100 words.
10. Make necessary variations among the 4 descriptions but
you MUST ensure they all adhere to above rules.

Return only the 4 **refined descriptions** in English
without commentary as **a single list of 4 strings** without
any numbering.
"""

# Few-shot examples for reference
examples = f"""
Example 1:
User Prompt: A mountain cabin during snowfall
Original Description:
```

```
A small cabin sits on the mountain while snow falls gently.
The cabin is burning in the snow, and inside it's raining
heavily. The snow is hot, and the fireplace is filled with
ice. The cabin is empty but also full of people singing.
Refined Description:
A small wooden cabin rests quietly on a snowy mountain
slope. Snowflakes drift through the air, and warm light
glows from the windows, creating a peaceful and cozy winter
atmosphere.
Example 3:
User Prompt: A city skyline at sunset
Original Description:
The city skyline at sunset is filled with colorful buildings
that change colors every second. The river under it is above
the buildings, creating a surreal view.
Refined Description:
The city skyline glows in warm shades of orange and pink as
the sun sets. The river below reflects the tall buildings,
creating a calm and beautiful evening scene.
Example 5:
User Prompt: A shark and a cat
Original Description:
A shark is swimming in an ocean, looking for food. Clear
blue water looks beautiful while the shark is swimming. The
sun is shining brightly.
Refined Description:
A shark is swimming in shallow water, while a cat is walking
on the nearby beach. It is a sunny day, and the blue water
makes a beautiful scene.
Example 6:
User Prompt: A person is walking with two dogs in a park.
Original Description:
A person is walking with his dog in a park. The dog is
playing around. The park has green grass and some gree trees
nearby. Overall scene is peaceful.
Refined Description:
A person is walking with two dogs in a park full of small
green grass. The dogs are playing with each other and walking
beside the person. There are some trees nearby. Overall a
beautiful afternoon for a refreshing walk.
"""


# Combine examples with user data
final_prompt = f"""
{examples}

Now refine the following description according to the above
rules and examples. Keep it logically coherent, concise, and
visually descriptive.

User Prompt: {user_prompt}
Original Description: {description}

4 Refined Descriptions:
"""
```

C.2 VISION LANGUAGE MODEL FEEDBACK PROMPT

```
You are an **Expert Text-to-Video (T2V) Alignment and
Optimization Agent**. Your function is to critically analyze
the generated video against two distinct prompts:
1.  **User Prompt Intent (UPI):** The short, original user
instruction (the truth source).
2.  **Description Prompt Old (DPO):** The detailed prompt
that was actually fed to the T2V model.

Your task is to prioritize Fidelity to UPI and strictly output
a single, valid JSON object following the prescribed schema.
Your analysis must use multi-step reasoning (Chain-of-Thought)
to link video failure to DPO flaws.


**Inputs for Analysis:**
1.  **User Prompt Intent (UPI):**"{USER_PROMPT_INTENT}"
2.  **Description Prompt Old (DPO):**"{DESCRIPTION_PROMPT_OLD}"

**PHASE 1: MULTI-DIMENSIONAL ASSESSMENT**
Evaluate the video on a scale of 0 to 10. For each dimension,
record the most critical observation (1-2 sentences).
* **A_TV (Text-Visual Alignment):** Adherence to all objects,
attributes like color, count, and environment in the DPO and UPI.
* **C_T (Temporal Coherence):** Consistency of identity,
background, and motion quality across the video duration.
* **F_C (Compositionality Fidelity):** Accuracy of complex
relations (spatial, numerical, causal).
* **Q_V (Visual Quality):** Resolution, aesthetic, and
freedom from artifacts.

**PHASE 2: ROOT CAUSE DIAGNOSIS (Chain-of-Thought)**
[STEP 2.1: IDENTIFY BOTTLENECK] Select the lowest-scoring
dimension. Prioritize A_TV, C_T, or F_C over Q_V.
[STEP 2.2: ANALYZE DUAL-PROMPT ALIGNMENT] Did the DPO (and
subsequently the video) violate the UPI?
[STEP 2.3: FORMULATE DIAGNOSIS] Write a concise 'root_cause'
linking the flaw (e.g., Compositional Overload, Ambiguous
Temporal Dynamics) to a specific part of the DPO.

**PHASE 3: ACTIONABLE PROMPT PRESCRIPTION**
Generate the refined prompts based solely on the Phase 2
Diagnosis.
* **P_new (Description Prompt New) Generation Rules:** Rewrite
the DPO to fix the flaw and enforce UPI alignment. Use
explicit staging, token weighting (e.g., (KEYWORD:1.3)), and
structural formatting.

**STRICT JSON OUTPUT SCHEMA (Start your response here):**

'''json
{{
  "prompt_analysis": {{
    "user_prompt_intent": "{USER_PROMPT_INTENT}",
    "bottleneck_flaw_metrics": {{
      "metric": "A_TV" | "C_T" | "F_C" | "Q_V",
```

```
      "score": 0,
      "observation": "..."
    }},
    "diagnosis": {{
      "root_cause": "..."
    }}
  }},
  "prescription": {{
    "P_new": "...",
    "meta_instruction": "..."
  }}
}}
```

C.3 VLM Feedback Output

VLM proposes but T2V fails to follow

```
"user_prompt": "Teddy bear and 3 real bear",
"Iterations_1": {
    "description_prompt_old": "Teddy bear and a real bear are
    playing together, with the teddy bear being a stuffed toy,
    while the real bear is a live animal.",
    "metrics": {
      "metric": "A_TV",
      "score": 2,
      "observation": "The video depicts only teddy bears without
      any real bears, failing to align with the UPI."
    },
    "prompt_new": "A forest clearing with a single teddy bear
    and three real bears. The teddy bear (STUFFED TOY:1.3) sits
    in front of the (REAL BEARS:1.5). The real bears are natural
    and curious, engaging with the environment."
  },
"Iterations_2": {
    "description_prompt_old": "A forest clearing with a single
    teddy bear and three real bears. The teddy bear
    (STUFFED TOY:1.3) sits in front of the (REAL BEARS:1.5).
    The real bears are natural and curious, engaging with the
    environment.",
    "metrics": {
      "metric": "A_TV",
      "score": 0,
      "observation": "The video displays three stuffed teddy
      bears consistently, and no real bears are present,
      violating both the UPI and DPO."
    },
    "prompt_new": "A forest clearing with a single teddy bear
    (STUFFED TOY:1.3) juxtaposed with three real bears (LIVE
    ANIMALS:1.5). The teddy bear sits distinctly in front of the
    real bears, which are naturally curious and engaging with
    the environment."
  },
```

VLM worsens quality


```
"user_prompt": "A small bird sits atop a blooming flower stem.",
"Iterations_1": {
    "description_prompt_old": "A small bird perches delicately
    atop a blooming flower stem. The vibrant petals surround
    the bird, creating a picturesque scene of nature's beauty.",
    "metrics": {
      "metric": "A_TV",
      "score": 6,
      "observation": "The video partially aligns with the UPI as
      it depicts a bird on a flower stem but misses the
      delicacy and integration with the bloom that the prompt
      suggests."
    },
    "prompt_new": "A small bird (KEYWORD:1.3) sits (ACTION:1.2)
    atop a blooming flower stem. Focus on the vibrancy of the
    flower and the bird's delicate placement.
    Ensure the petals and stem are clearly shown."
  },
"Iterations_2": {
    "description_prompt_old": "A small bird (KEYWORD:1.3) sits
    (ACTION:1.2) atop a blooming flower stem. Focus on the
    vibrancy of the flower and the bird's delicate placement.
    Ensure the petals and stem are clearly shown.",
    "metrics": {
      "metric": "A_TV",
      "score": 2,
      "observation": "The video focuses only on the flower and
      does not include the bird."
    },
    "prompt_new": "A small bird (KEYWORD:1.5) clearly sits
    (ACTION:1.3) atop a blooming flower stem, ensuring both the
    bird and flower are fully visible."
  }
```