Rewind and Render: Towards Factually Accurate Text-to-Video Generation with Distilled Knowledge Retrieval

Daniel Lee[*1], **Arjun Chandra**[*2], **Yang Zhou**[3], **Yunyao Li**[1], **Simone Conia**[4]
[1] Adobe
[2] Boston University
[3] Adobe Research
[4] Sapienza University of Rome
dlee1@adobe.com, ac25@bu.edu, yazhou@adobe.com, yunyaol@adobe.com, simone.conia@uniroma1.it

Abstract

Text-to-Video (T2V) models, despite recent advancements, struggle with factual accuracy, especially for knowledge-dense content. We introduce FACT-V (Factual Accuracy in Content Translation to Video), a system integrating multi-source knowledge retrieval into T2V pipelines. FACT-V offers two key benefits: i) improved factual accuracy of generated videos through dynamically retrieved information, and ii) increased interpretability by providing users with the augmented prompt information. A preliminary evaluation demonstrates the potential of knowledge-augmented approaches in improving the accuracy and reliability of T2V systems, particularly for entity-specific or time-sensitive prompts.

Demo Video — https://bit.ly/factv-demo
Dataset — https://bit.ly/factv-dataset

Introduction and Related Work

Recent years have witnessed significant advancements in Text-to-Video (T2V) models, marked by the emergence of both closed-source solutions, such as Pika Labs, RunwayML's Gen-2, and OpenAI's Sora (Pika Labs 2023; Runway 2023; OpenAI 2024b), as well as open-source alternatives such as CogVideoX and Open-Sora (Yang et al. 2024; Zheng et al. 2024). These developments have led to substantial improvements in various aspects of video generation, including text alignment, video resolution, and temporal consistency (Cho et al. 2024). Despite these advancements, a critical challenge remains largely unexplored: the generation of videos containing knowledge-dense or factually liable content. This issue extends beyond text alignment, focusing on the production of content with concrete, supporting factual details that may not be explicitly captured in the prompts. This challenge is particularly pronounced in scenarios that involve specific categories/classes of entities such as people, events, locations, or evolving knowledge.

Although similar problems have been addressed in other generative domains, particularly in Natural Language Processing (NLP), through approaches like retrieval-augmented generation (Lewis et al. 2020, RAG), their application in visual domains, such as Text-to-Image (T2I) and T2V, remains limited. Recent work has begun to explore this area in T2I (Lim and Shim 2024; Wan et al. 2024), but research in T2V is further behind.

A common approach to improving factual representation is retraining the model with more comprehensive and up-to-date data. However, this strategy faces significant challenges in the T2V domain. The availability of high-quality training data is limited, and there is considerable difficulty in curating resources, especially those containing up-to-date real-world knowledge. Moreover, the need for frequent retraining to keep up with rapidly changing information poses a substantial logistical challenge (Girdhar et al. 2024). These factors render the retraining approach suboptimal, particularly for dynamic, rapidly evolving knowledge domains.

Drawing parallels from both textual and visual generation domains, we propose that information retrieval via web search and large language models may offer a more efficient and adaptable method for integrating accurate, up-to-date information in T2V systems.

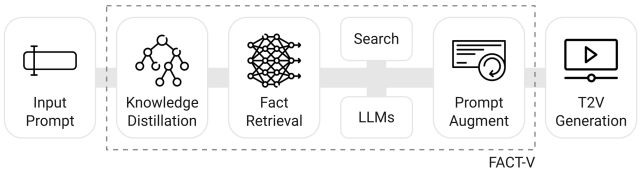

Figure 1: System flow chart for FACT-V, including *knowledge distillation*, *fact retrieval*, and *prompt augmentation*.

System Overview

We introduce FACT-V (Factual Accuracy in Content Translation to Video), a system that integrates retrieval mechanisms into a T2V pipeline (Figure 1).

FACT-V relies on dynamic knowledge retrieval to improve factual accuracy as well as provides increased interpretability in generated videos, especially required for specific categories of entities and evolving knowledge, with the aim of providing a novel framework for more transparent and adaptable T2V systems.

[*]These authors contributed equally.

Input Prompt: Scientist Discovering the First Leprosy Treatment.
Fact: Alice Ball was an [African American] [Female] Chemist.

Input Prompt: Statue of Liberty in 1890
Fact: Statue of Liberty is a [copper] [reddish-brown] figure.

Figure 2: Text-to-Video Model Output Comparisons: Examples of videos generated without (left) and with (right) FACT-V.

Knowledge Distillation. We employ a manually-curated visual ontology tailored to entity types, which is used to supplement the initial prompt. Our approach involves sequentially and dynamically chaining LLMs (i.e., GPT-4o) to extract and distill core information (Khot et al. 2023). The ontology covers various entity types including people, events, locations, and others. Based on this decomposition, we generate refined prompts for the subsequent retrieval phase. For instance, as shown in Figure 2, we identify the entity (Alice Ball), and distill relevant information based on her entity type (people), such as ethnicity, gender, and other attributes.

Fact Retrieval. This component combines two sources to gather information: (1) web search and (2) large language models (LLMs). For each query q, we route it to the optimal knowledge base. The retrieval process is formulated using LLM calls with OpenAI's GPT-4o and Perplexity API (OpenAI 2024a; Perplexity 2023).

Prompt Augmentation. We augment and rewrite the initial prompt leveraging the collected information. Only factual information with visual implications is filtered and incorporated from the raw retrieved facts. Finally, rewriting using effective prompting techniques for T2V models (Wang and Yang 2024) results in the knowledge-augmented prompt for submission to the T2V model.

Evaluation. We conducted an evaluation of FACT-V, which comprised of a preferential rating across (1) factuality and (2) alignment. Our study focused on a small scale evaluation of challenging prompts which revealed a notable improvement in factuality, whereas alignment slightly exceeded parity with a high level of annotator agreement.

User Interface Guide. The FACT-V user interface (Figure 3) – modeled after Conia et al. (2024) – is divided into three main sections:

- **Section 1 – Input Prompt:** Users enter a prompt p and select a T2V model m for video generation.
- **Section 2 – System Comparison:** Enables a side-by-side comparison between a standard and FACT-V-enhanced T2V model, allowing users to assess the impact of distilled knowledge retrieval.
- **Section 3 – Retrieved Knowledge:** Displays information fetched by the knowledge retrieval component, improving the transparency and interpretability of the generation process.

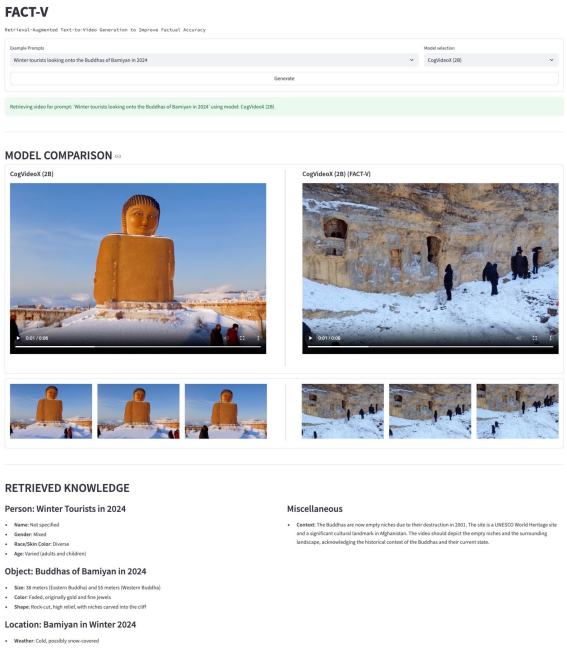

Figure 3: User interface of FACT-V, including a side-by-side comparison with and without FACT-V and the knowledge retrieved and used by our system.

Conclusion and Future Work

FACT-V showcases how the integration of retrieval-augmented knowledge into T2V generation can enhance the process in two key ways: i) producing videos with improved factual accuracy; and, ii) offering greater interpretability in the generation process by allowing users to observe the retrieved information utilized by the system. FACT-V suggests promising avenues for advancing T2V models, with future research directions including: i) exploring more sophisticated retrieval mechanisms to incorporate diverse and contextually relevant information; and, ii) investigating the impact of retrieval-augmented generation on reducing biases and improving fairness in visual outputs.

Acknowledgements

We are grateful to the anonymous reviewers for their valuable feedback. We also thank Yoonjoo Lee, Ronak Pradeep and Revanth Gangi Reddy for their insightful discussions. Simone Conia gratefully acknowledges the support of the PNRR MUR project PE0000013-FAIR, which fully funds his work since October 2023.

References

Cho, J.; Puspitasari, F. D.; Zheng, S.; Zheng, J.; Lee, L.-H.; Kim, T.-H.; Hong, C. S.; and Zhang, C. 2024. Sora as an AGI World Model? A Complete Survey on Text-to-Video Generation. arXiv:2403.05131.

Conia, S.; Lee, D.; Li, M.; Minhas, U. F.; and Li, Y. 2024. Enhancing Machine Translation Experiences with Multilingual Knowledge Graphs. In *AAAI Conference on Artificial Intelligence*.

Girdhar, R.; Singh, M.; Brown, A.; Duval, Q.; Azadi, S.; Rambhatla, S. S.; Shah, A.; Yin, X.; Parikh, D.; and Misra, I. 2024. Emu Video: Factorizing Text-to-Video Generation by Explicit Image Conditioning. arXiv:2311.10709.

Khot, T.; Trivedi, H.; Finlayson, M.; Fu, Y.; Richardson, K.; Clark, P.; and Sabharwal, A. 2023. Decomposed Prompting: A Modular Approach for Solving Complex Tasks. arXiv:2210.02406.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Lim, Y.; and Shim, H. 2024. Addressing Image Hallucination in Text-to-Image Generation through Factual Image Retrieval. arXiv:2407.10683.

OpenAI. 2024a. GPT-4o. Https://openai.com/index/hello-gpt-4o/.

OpenAI. 2024b. Sora. Https://openai.com/index/sora.

Perplexity. 2023. API. Https://docs.perplexity.ai/home.

Pika Labs. 2023. Pika. Https://pika.art/home.

Runway. 2023. Gen-2. Https://runwayml.com/research/gen-2.

Wan, Y.; Wu, D.; Wang, H.; and Chang, K.-W. 2024. The Factuality Tax of Diversity-Intervened Text-to-Image Generation: Benchmark and Fact-Augmented Intervention. arXiv:2407.00377.

Wang, W.; and Yang, Y. 2024. VidProM: A Million-scale Real Prompt-Gallery Dataset for Text-to-Video Diffusion Models. arXiv:2403.06098.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; Yin, D.; Gu, X.; Zhang, Y.; Wang, W.; Cheng, Y.; Liu, T.; Xu, B.; Dong, Y.; and Tang, J. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. arXiv:2408.06072.

Zheng, Z.; Peng, X.; Yang, T.; Shen, C.; Li, S.; Liu, H.; Zhou, Y.; Li, T.; and You, Y. 2024. Open-Sora: Democratizing Efficient Video Production for All.