Grounding Pixels in Facts: Distilled Knowledge Retrieval for Factual Text-to-Video Generation

Daniel Lee, Arjun Chandra, Yang Zhou, Yunyao Li, Simone Conia
dlee1@adobe.com, ac25@bu.edu, yazhou@adobe.com, yunyaol@adobe.com, simone.conia@uniroma1.it

Overview

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:

- Improved factual accuracy of generated videos through dynamically retrieved information.
- Increased interpretability by providing users with the augmented prompt information.

Method

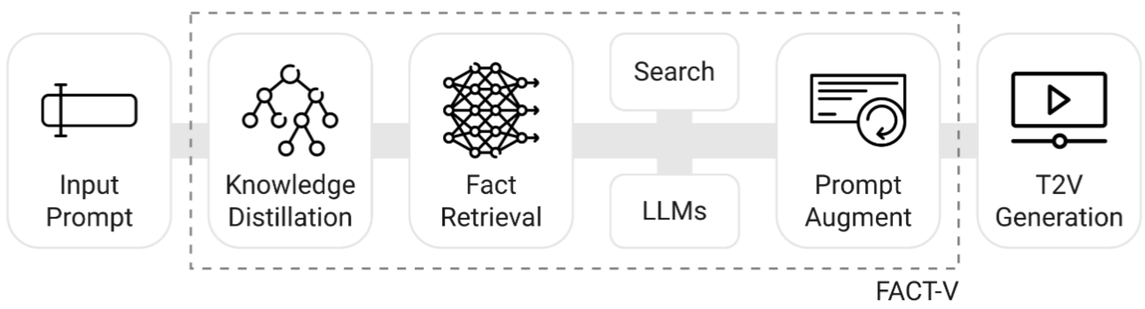

Figure 1. *System Flow Chart for FACT-V*

To improve the factual accuracy of generated content, FACT-V leverages three components:

- **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. 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.

- **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.

- **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 results in the knowledge-augmented prompt for submission to the T2V model.

Results

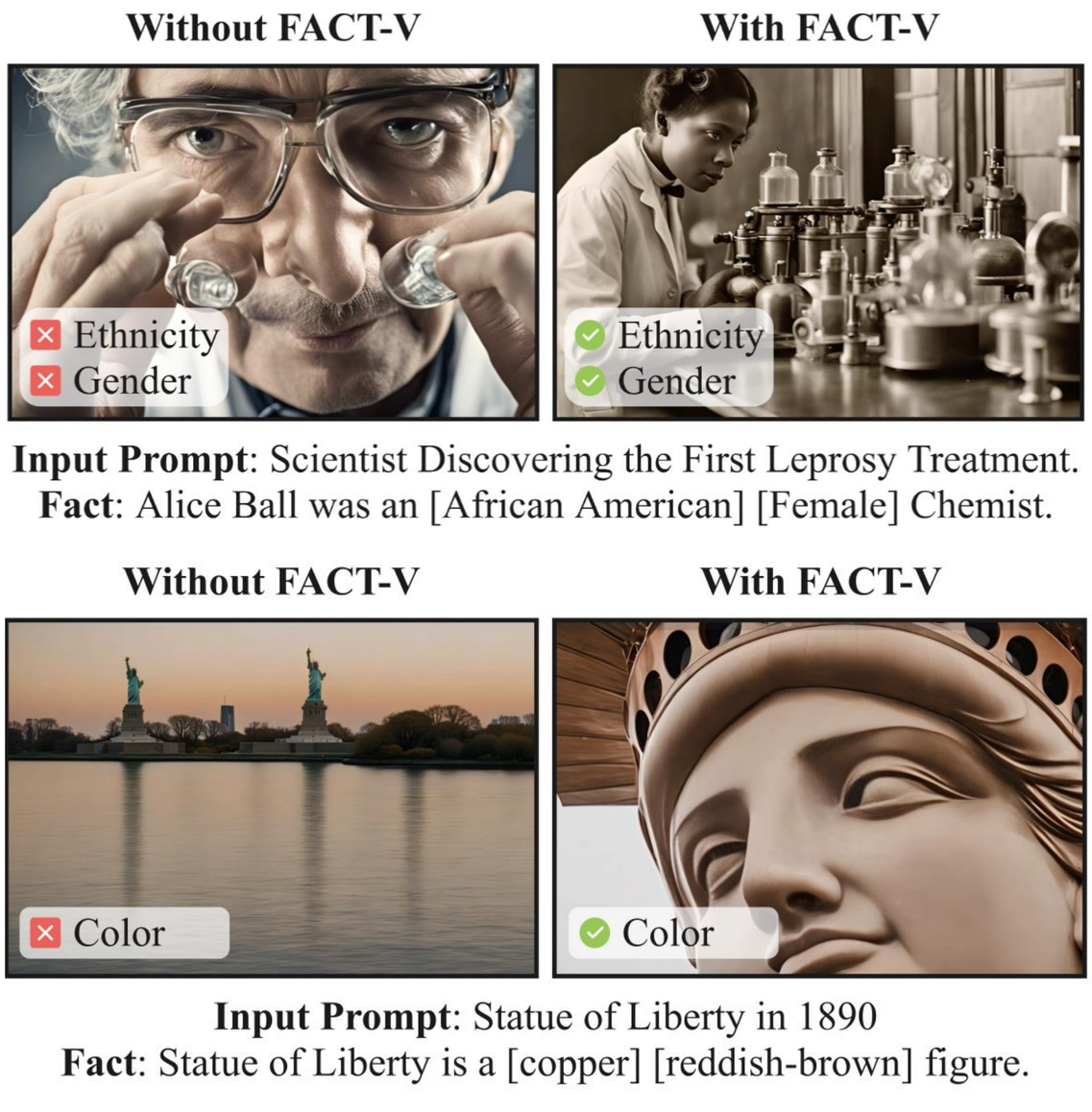

Figure 2. *FACT-V Improves Factual Accuracy in T2V Generation*

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.

Model	Factuality (%)	Alignment (%)
Runway Gen-2	0.67 / 36.67 / **62.67**	2.00 / **76.67** / 21.33
CogVideoX-2B	6.00 / 41.33 / **52.67**	12.67 / **61.33** / 26.00

Table 1. *Preferential ranking formatted as [WITHOUT FACT-V / SAME / FACT-V] for Runway's Gen-2 and CogVideoX-2B*

Discussion

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.