Issues in Information Systems
Volume 26, Issue 1, pp. 377-393, 2025

DOI: https://doi.org/10.48009/1_iis_128

TTV: Towards advancing text-to-video generation with generative AI models and a comprehensive study of model fidelity, performance, and human perception

Tasnim Akter Onisha, *Georgia Southern University, tonisha@students.kennesaw.edu*
Hayden Wimmer, *Georgia Southern University, hwimmer@georgiasouthern.edu*
Carl M. Rebman, Jr., *University of San Diego, carlr@sandiego.edu*

Abstract

Text-to-video generation has rapidly evolved as a groundbreaking application of generative AI, with the potential to revolutionize both creative and industrial sectors. Despite these advancements, the fidelity, performance, and real-world applicability of current models remain inadequately explored. This research aims to address this gap by evaluating the performance of three cutting-edge text-to-video models: Runway Gen2, CogVideoX-2B, and CogVideoX-5B. The primary objectives of this study are to (1) conduct a comprehensive evaluation of these models using rigorous mathematical assessments such as Frechet Inception Distance (FID), Frechet Video Distance (FVD), and CLIPScore to measure video quality, realism, and alignment with text input; (2) gather human perceptual data to assess perceived realism, quality, and accuracy; and (3) compare the models to identify strengths, weaknesses, and areas for improvement. To uncover how AI-generated videos measure up to human expectations, this study asked 60 participants to rate outputs from three leading text-to-video models using a 7-point Likert scale, 10 diverse prompts, and 10 real-world benchmarks. While CogVideoX-2B impressed with its precision and alignment, CogVideoX-5B stood out for its striking realism in the eyes of human viewers. These findings reveal a compelling trade-off between technical accuracy and perceptual appeal which highlights the need for evaluation methods that balance both.

Keywords: text-to-video generation, Generative AI, TTV, text-to-video generative models, transformer models, Runway Gen-2, CogVideoX, CogVideoX-2B, CogVideoX-5B

Introduction

Text-to-video (TTV) generation is emerging as a transformative technology with wide-reaching implications across creative industries, education, and entertainment. Unlike prior generative AI models focused on static images, TTV adds complexity by generating coherent, temporally consistent video content directly from textual input. This shift enables richer, more dynamic media experiences that align with human narrative understanding. Despite rapid advancements, current TTV models remain underexplored in terms of fidelity, alignment with prompts, and perceptual quality. Most evaluations rely heavily on objective metrics, such as Frechet Inception Distance (FID), Frechet Video Distance (FVD), and CLIPScore which, while rigorous, often fail to capture subjective human experiences like realism and emotional coherence. Consequently, a critical gap exists in understanding how well these models align with human perception. Subsequently, there is a lack of systematic, dual-evaluation frameworks that combine both quantitative assessments and perceptual insights to benchmark model performance. This limits our ability to holistically evaluate TTV models and guide their development for real-world applications. Addressing this gap requires a dual evaluation framework that integrates both computational and human-centered assessments

Previous research has shown that generative models are often evaluated using a combination of objective metrics such as Frechet Inception Distance (FID), Frechet Video Distance (FVD), and CLIPScore to measure quality, realism, and accuracy (Hong, Ding, Zheng, Liu, & Tang, 2022). However, these mathematical assessments do not always capture subjective qualities, such as how real or engaging a generated video feels to human viewers. As such, there is a growing need to evaluate these models through both objective and perceptual lenses to fully understand their strengths and limitations. Human perceptual analysis has also been employed in related fields, such as deepfake detection, where (Onisha, Wimmer, & Rebman Jr, 2024) utilized facial expression analysis tools and survey to discern emotional differences between real and manipulated videos, underscoring the importance of combining computational and perceptual evaluations to assess video authenticity.

To bridge this gap, this study provides a comprehensive evaluation of three state-of-the-art TTV models: Runway Gen2, CogVideoX-2B, and CogVideoX-5B using both computational metrics and a structured human perception study. We aim to (1) assess the models using FID, FVD, and CLIPScore; (2) evaluate user perception of realism, quality, and accuracy via a 60-participant Likert-scale survey; and (3) analyze trade-offs between metric performance and human judgment. This dual approach contributes to a benchmark for TTV model evaluation and offers insights into how generative models can be improved to better meet user expectations and real-world application needs. This can inform the design of future models and lead to innovations that enhance their quality, realism, and practical utility.

Literature Review

Generating high-quality, coherent videos from text has remained a more complex and less explored challenge. The growing interest in video generation reflects the potential of this technology to revolutionize various industries, including entertainment, marketing, education, and more. The following sections provide an in-depth exploration of recent advancements in text-to-video generation models, their evaluation methodologies, and the challenges faced by researchers in this rapidly developing field.

Recent Advancements in Text-to-Video Generation Models

Show-1 by (D. J. Zhang et al., 2024) integrates pixel-based and latent-based Video Diffusion Models (VDMs) to enhance video quality, alignment, and motion fidelity. Trained on WebVid-10M, Show-1 outperforms models like CogVideo and Video LDM on benchmarks such as UCF-101 and MSR-VTT, showcasing improved video realism and text-video alignment. This highlights the growing sophistication in generating videos that align with textual input, a key area of focus in this study.

Similarly, CogVideo (Hong et al., 2022), a large-scale transformer-based text-to-video generation model, demonstrates superior performance on both machine and human evaluations. With its multi-frame-rate hierarchical training strategy, it improves text-video alignment, setting a benchmark for the models under study in this research. (Singer et al., 2022) presents a method Make-A-Video for generating high-quality videos from text prompts. In the zero-shot evaluation on MSR-VTT, Make-A-Video achieved an FID of 13.17 and a CLIPSIM of 0.3049, outperforming prior models, including CogVideo, which had an FID of 23.59 and CLIPSIM of 0.2631. Human evaluations confirmed that Make-A-Video was preferred over VDM and CogVideo for video quality and text-video faithfulness, with preference percentages of 84.38% and 78.13%, respectively.

(Kim, Joo, & Kim, 2020) introduced TiVGAN, a Text-to-Image-to-Video GAN that stabilizes training and improves video quality using a step-by-step process. Tested on datasets like KTH, MUG, and Kinetics, it showed superior results in FID and Inception scores. (Y. Zhang et al., 2023)proposed ControlVideo, a training-free framework that enhances appearance consistency and temporal stability, outperforming methods like Tune-A-Video and Text2Video-Zero in video quality. However, it struggles with generating videos beyond input motion sequences. (Oh et al., 2025) developed MEVG, a method for generating videos with multiple events from text, using a last-frame-aware diffusion process. This outperforms other models in temporal consistency, semantic accuracy, and overall video quality.

Evaluation Methodologies and Benchmarking

A major aspect of this study involves evaluating the performance of text-to-video models using rigorous metrics such as FID, FVD, and CLIPScore, which are commonly used in the field to assess visual quality, temporal coherence, and alignment with textual input. Studies such as FETV by (Liu et al., 2024) proposed benchmark which evaluates text-to-video (T2V) models using three main criteria: Alignment, Temporal Coherence, and Visual Quality. Findings from the benchmark show that while models perform well in Visual Quality, they struggle with Temporal Coherence, showing difficulty in maintaining smooth transitions and consistent actions. Alignment scores also reveal some discrepancies, indicating room for improvement in generating videos that accurately reflect text descriptions.

These findings provide key insights for enhancing Text to video model performance. Furthermore, by developing a new benchmark of text prompts representing various levels of dynamics, (Liao et al., 2024) demonstrates that DEVIL (Dynamics-Evaluating Video Inference Learning) correlates highly with human evaluations, achieving over 90% Pearson correlation. The proposed metrics and benchmark provide valuable insights for advancing T2V generation models, highlighting the importance of dynamics in creating realistic and contextually accurate video content.

Challenges in Text-to-Video Generation

Despite significant advancements, generating high-quality, long-duration videos that maintain coherence throughout remain a significant hurdle in the field. Moreover, the evolution of text-to-video generation models, exemplified by advancements such as Sora, underscores the increasing complexity of integrating vision, language, and temporal dynamics, yet challenges remain in areas like dataset quality, evaluation metrics, and human-AI interaction, highlighting the field's infancy and the need for further research to realize a true world model and advance toward AGI (Cho et al., 2024). A survey by (Lei, Wang, Ma, Huang, & Liu, 2024) showed that the integration of generative models such as GANs and diffusion models has also shown promise in improving video quality and realism. These models aim to address key challenges such as motion consistency, occlusion, and appearance instability in generated videos.

Human Perception and Model Evaluation

Studies like (Leiker, Gyllen, Eldesouky, & Cukurova, 2023) examines the effectiveness of AI-generated videos in online education, showing that participants often cannot differentiate between AI-generated and traditional instructor-led videos. This highlights the growing importance of human perceptual evaluations in assessing video realism and quality. By incorporating human evaluations in this research, we aim to better understand the subjective qualities of generated videos, an essential step in advancing TTV models that align with user expectations.

Recent works, such as ModelScopeT2V (Wang et al., 2023) introduced ModelScopeT2V, a text-to-video synthesis model based on Stable Diffusion. The model incorporates spatio-temporal blocks to ensure smooth frame transitions and temporal coherence. It achieved state-of-the-art results, outperforming existing methods like Make-A-Video and Imagen Video in terms of FID-vid, FVD, and CLIPSIM scores, with the lowest FID-vid (11.09) and FVD (550) scores. These results demonstrate superior visual fidelity, temporal consistency, and semantic alignment with text prompts. ModelScopeT2V shows strong potential for text-to-video synthesis, offering high-quality video generation. (Yang, Zhou, Liu, & Loy, 2023) present a video-to-video translation framework that ensures high temporal consistency by propagating key frames across video sequences, showcasing the importance of maintaining coherence across frames in generated videos. This methodology informs the ongoing exploration of temporal coherence in TTV models and serves as a reference for evaluating the performance of the models under consideration in this study.

Evaluation Framework Rationale

Although recent literature highlights advance in text-to-video generation, most evaluations isolate either computational metrics or user preferences, rarely combining both in a structured manner. Models such as CogVideo, Make-A-Video, and ModelScopeT2V are often assessed using single-mode evaluations. This study selects Runway Gen2, CogVideoX-2B, and CogVideoX-5B to represent architectural diversity and benchmark prominence. FID, FVD, and CLIPScore are chosen for their reproducibility and widespread use, yet these metrics alone may not fully capture perceived quality. A structured human evaluation is therefore incorporated to complement quantitative analysis. This dual framework directly responds to gaps in prior work and enables a more holistic assessment of both fidelity and user-aligned realism.

Methodology

This study assesses the performance of text-to-video generation models through a structured methodology involving model selection, video generation, and evaluation via computational metrics and human perception. Figure 1 illustrates the overall methodology.

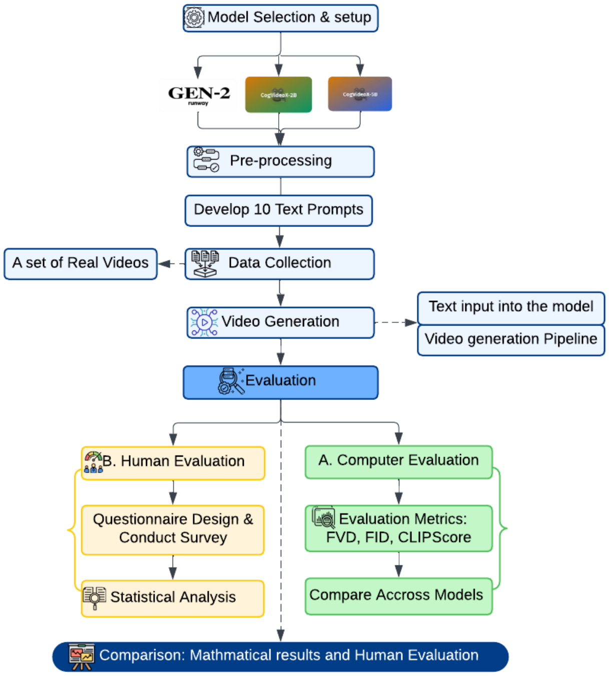


Figure 1. Proposed Methodology

The research focuses on three prominent models: Runway Gen-2, CogVideoX-2B, and CogVideoX-5B, pretrained on Hugging Face and tested for their ability to generate realistic and high-quality videos. As benchmarks, ten real videos from diverse sources were collected. Utilizing identical 10 text prompts, ten sets of videos were generated using each of the models. Two techniques were employed to evaluate the generated videos: computer evaluation and human evaluation via a survey. For computer evaluation, metrics such as FID score, FVD, and CLIPSCORE were computed for each set of generated VIDEOS in comparison to the real ones. The results were analyzed and further validated through a survey involving 60 participants. Statistical analysis was performed and compared the results across the models, providing comprehensive insights into their performance. The three models selected for text-to-video generation in this study are:

Model 1: Runway Gen2

This model is an advanced text-to-video generation model designed for high-quality video synthesis from text descriptions. It builds upon generative AI techniques and incorporates temporal consistency across video frames. The process begins with text encoding, where a pre-trained language model converts the textual prompt into a latent vector, capturing its semantic and contextual nuances. In the video generation pipeline, the text embeddings are mapped to a latent space that defines the video's structure and style. Frames are generated iteratively using an image generator while ensuring temporal coherence through RNNs or attention mechanisms. Finally, post-processing techniques, including noise reduction, frame interpolation, and color correction, enhance the video's quality and smoothness. Figure 2 describes how this model works, and Figure 3 shows our generated video frames (e.g. only 5 frames) through Runaway Gen-2 Model.

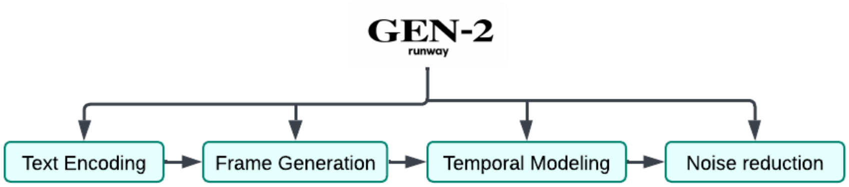
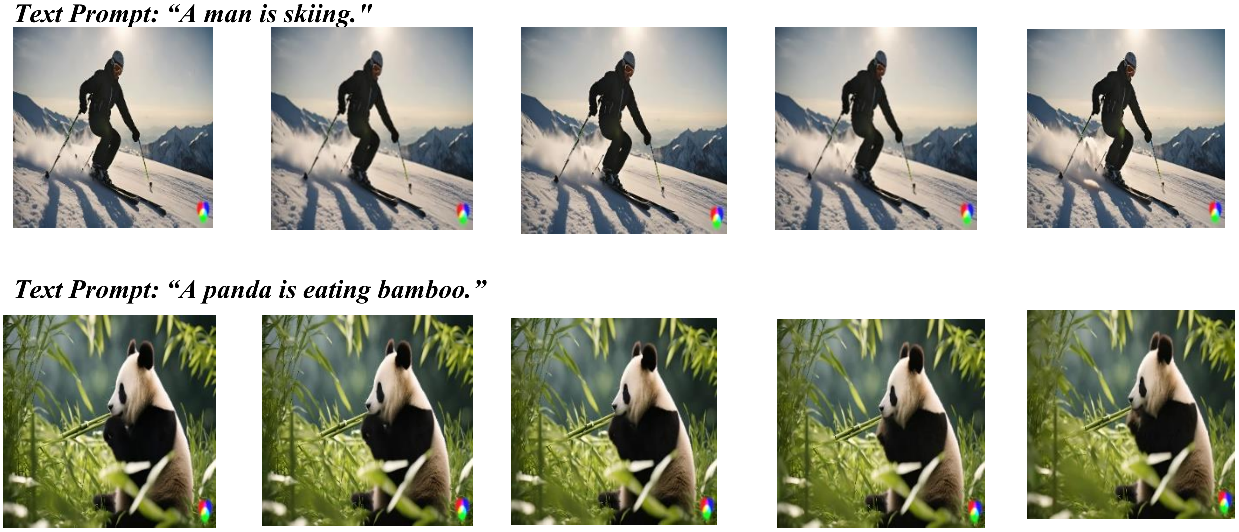


Figure 2. Text to Video generation of Runway Gen-2



Text Prompt: "A man is skiing."

Text Prompt: "A panda is eating bamboo."

Figure 3. Examples of Generated Video frames

Model 2: CogVideoX-2B

CogVideoX-2B, a latent diffusion model optimized for text-to-video generation, was utilized to evaluate its performance and efficiency. With 2 billion parameters, it strikes a balance between computational efficiency and video quality, making it suitable for generating videos from concise text prompts. CogVideoX-2B employs a diffusion-based pipeline, starting with random noise and refining it iteratively over multiple inference steps to produce coherent video frames. The process involves encoding text prompts into latent embeddings, which guide the generation of video frames sequentially, ensuring temporal consistency. The model leverages a Variational Autoencoder (VAE) for latent space operations and utilizes guidance mechanisms, such as the guidance scale, to enhance the alignment of generated videos with the text prompt. This architecture enables CogVideoX-2B to produce realistic and semantically relevant video, while maintaining computational efficiency. Figure 4 illustrates how this generation model works, and Figure 5 shows our generated video frames (e.g. only 5 frames) through CogVideoX-2B Model.

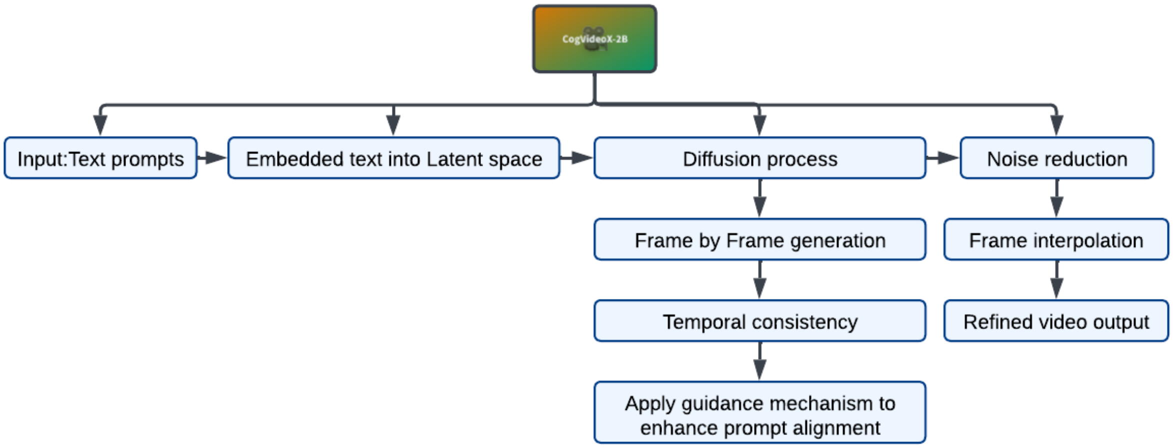
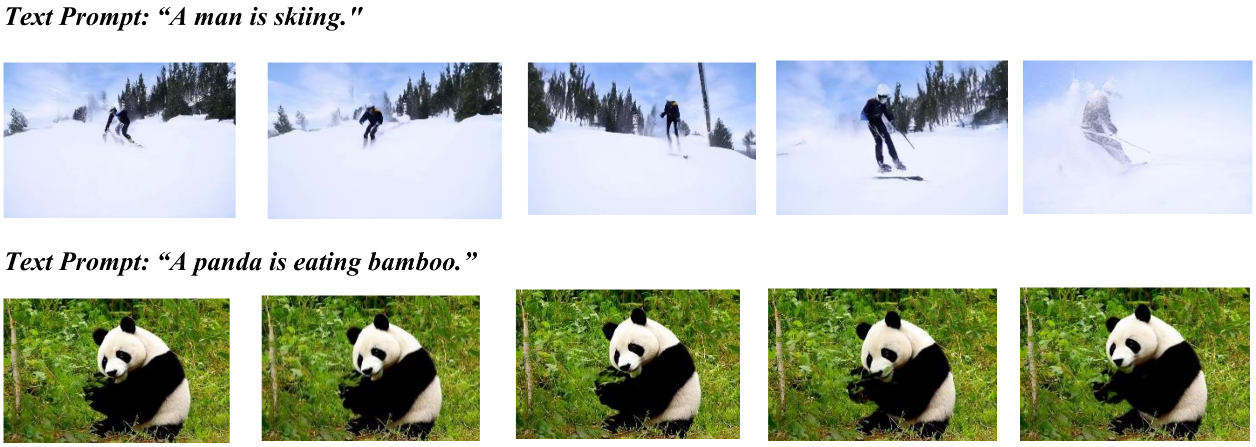


Figure 4. Text-to-Video Generation of CogVideoX-2B


Text Prompt: "A man is skiing."

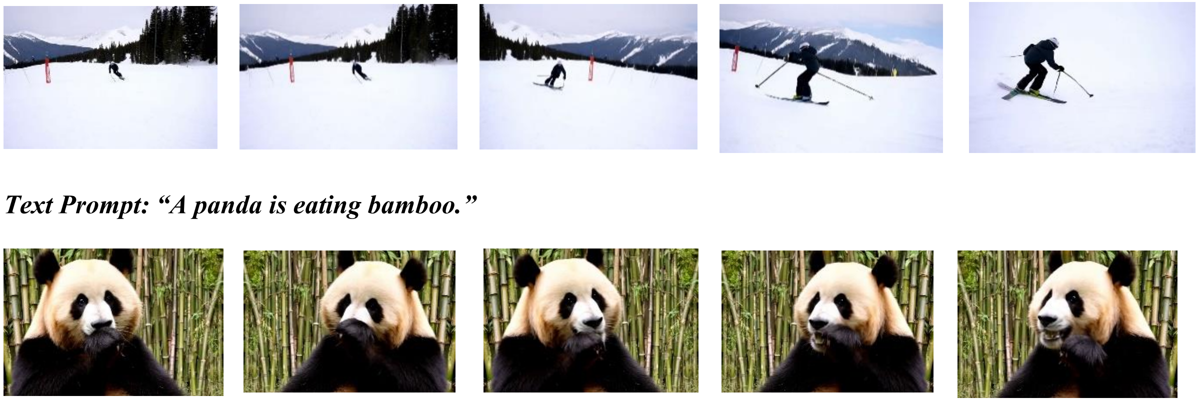
Text Prompt: "A panda is eating bamboo."

Figure 5. Example of generated video frames

Model 3: CogVideoX-5B

In this research, CogVideoX-5B, a more advanced version of CogVideoX, enhances the video generation quality with its 5 billion parameters. Similar to CogVideoX-2B, this model employs a latent diffusion pipeline, but with a larger parameter set, it produces videos with finer detail, better temporal consistency, and a more accurate representation of complex text prompts. The model follows the same general process of encoding text into latent embeddings, using a Variational Autoencoder (VAE) for refining the latent space and applying guidance mechanisms to align the generated videos with the provided text prompt. The larger scale of CogVideoX-5B allows for higher-quality outputs, making it suitable for applications requiring more nuanced and realistic video generation. Figure 6 shows our generated video frames (e.g. only 5 frames) through CogVideoX-5B Model.

Text Prompt: "A man is skiing."



Text Prompt: "A panda is eating bamboo."

Figure 6. Example of generated frames

Data Collection

Real video datasets were obtained from various sources, each representing different scenarios and contents. These videos were divided into individual frames that serve as a reference for comparison with generated videos. The dataset includes 10 real video folders, each containing frames extracted from a single video. Figure 7 shows some random frames as examples of real video.

Figure 7. Example of real video frames with Text Prompt "A person is surfing huge

Generated Videos

The generated video dataset includes videos produced by the three models. Each generated video was split into frames similar to the real videos. The synthetic video set also contains 10 folders corresponding to different generated videos, each with multiple frames.

Text Prompts

The study uses 10 distinct text prompts to generate the videos. These prompts are:

"A cat is playing with a yarn in a sofa."
"A panda is eating bamboo."
"A dog is playing with a ball in the field."
"Bald Eagle soaring in the sky."
"A man is skiing."
"Leaves are rustling in the wind."
"A person is surfing huge waves."
"3-Year-Old Riding Bike Without Training Wheels."
"A campfire burning at night."
"A fish swimming in an aquarium with coral reefs."

Preprocessing and Fine-Tuning

In this implementation, a list of ten diverse prompts was created, such as 'A cat is playing with a yarn on a sofa' and 'A panda is eating bamboo,' to test the model's capabilities in generating various scenes. The first step involves extracting frames from both real and generated videos using an automated frame extraction script. These frames are then resized and normalized to ensure uniformity in resolution and visual consistency across both datasets. For fine-tuning, video-text pairs were used to ensure semantic alignment between the text and generated video content, optimizing the model's ability to produce coherent video sequences. The inference process generates videos by iteratively refining a random noise representation into video frames through a fixed number of inference steps (50 in our setup), with each video comprising 49 frames. A guidance scale of 6 ensures that the model closely follows the provided prompts, enhancing text-to-video alignment. Once the video is generated, it is exported to a video file format (MP4), individual frames are extracted for further analysis, and both the video and frames are downloaded in a convenient format.

Two distinct evaluation methods were employed to assess the realism, quality, accuracy of the generated videos from text prompts: **Method A (Mathematical Evaluation)** and **Method B (Human Study).** Below is a detailed explanation of both methods, including their mathematical formulations and how the evaluation Type equation here.was conducted.

Method A: Mathematical Evaluation

In Method A, this study used several quantitative metrics to assess the realism, quality and accuracy with text alignment of the generated videos. These metrics were computed using Python frameworks and libraries to compare the generated videos with real videos.

Fréchet Inception Distance (FID)

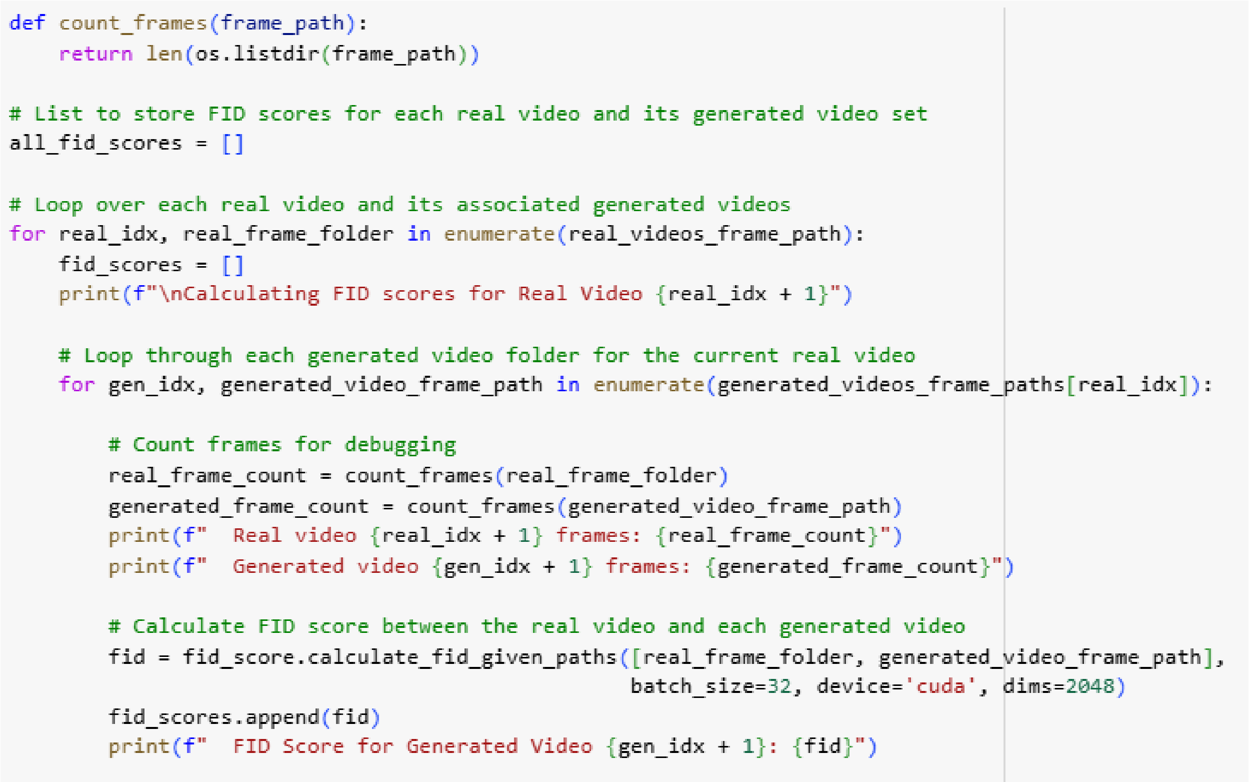
The FID score is used to evaluate the quality of generated images and videos by comparing the distribution of generated data to that of real data. A lower FID indicates higher similarity between the two distributions, indicating better realism (Brownlee, 2019). The FID score was calculated using the activations from an Inception-v3 model, which was used to extract feature representations from both real and generated videos as shown in Figure 8. The distance between the distributions of these activations was computed to obtain the FID score. The equation 1 formula for FID is used:

Mathematical Equation:

$$FID = \| \mu_r - \mu_g \|^2 + Tr\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{\left\{\frac{1}{2}\right\}}\right) \text{ (Equation 1)}$$

- μ_r *and* μ_g are the mean feature vectors of the real and generated video frames, respectively
- $\Sigma_r \Sigma_g$ are the covariance matrices of the real and generated video frames features, respectively
- Tr denotes the trace of the matrix.

a. Libraries Used: torchvision for using the Inception v3 model to extract features from video frames., torch (PyTorch) for working with neural networks, scipy for calculating the Frechet distance.



```
def count_frames(frame_path):
    return len(os.listdir(frame_path))

# List to store FID scores for each real video and its generated video set
all_fid_scores = []

# Loop over each real video and its associated generated videos
for real_idx, real_frame_folder in enumerate(real_videos_frame_path):
    fid_scores = []
    print(f"\nCalculating FID scores for Real Video {real_idx + 1}")

    # Loop through each generated video folder for the current real video
    for gen_idx, generated_video_frame_path in enumerate(generated_videos_frame_paths[real_idx]):

        # Count frames for debugging
        real_frame_count = count_frames(real_frame_folder)
        generated_frame_count = count_frames(generated_video_frame_path)
        print(f"  Real video {real_idx + 1} frames: {real_frame_count}")
        print(f"  Generated video {gen_idx + 1} frames: {generated_frame_count}")

        # Calculate FID score between the real video and each generated video
        fid = fid_score.calculate_fid_given_paths([real_frame_folder, generated_video_frame_path],
                                                   batch_size=32, device='cuda', dims=2048)
        fid_scores.append(fid)
        print(f"  FID Score for Generated Video {gen_idx + 1}: {fid}")
```

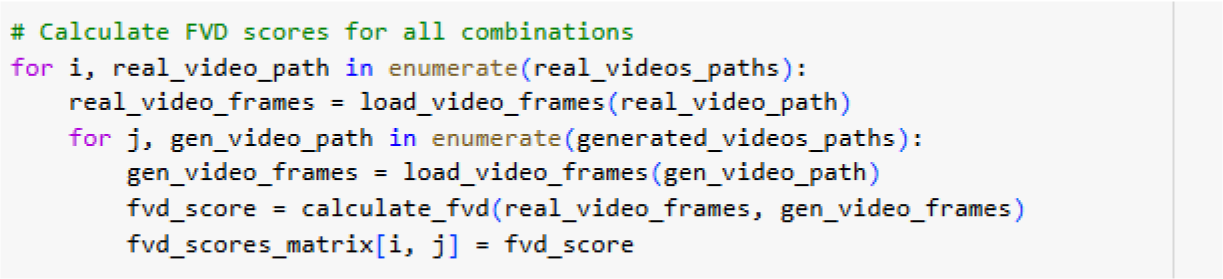
Figure 8. Code snippet of FID calculation

Fréchet Video Distance (FVD)

This metric is an extension of FID, designed to evaluate the quality of generated videos by comparing feature distributions of real and generated videos. It operates similarly to FID but considers temporal dynamics and video features. The main difference is that FVD works with sequences of frames (video). To calculate FVD as shown in Figure 9, frames are extracted from real and generated videos while preserving temporal order. These frames are resized and processed through a pre-trained video feature extractor like I3D to capture spatial and temporal features. The mean and covariance of these features are computed, and FVD is calculated by comparing the feature distributions of real and generated videos, providing a metric that accounts for both spatial quality and temporal coherence. The equation 2 is the formula of FVD:

$$FVD = \| \mu_r - \mu_g \|^2 + Tr\left(C_r + C_g - 2\left(C_r C_g\right)^{\left\{\frac{1}{2}\right\}}\right) \text{(Equation 2)}$$

- $\mu_r - \mu_g$ are the means of the feature vectors for real and generated videos
- $C_r C_g$ are the covariance matrices of the feature vectors for real and generated videos
- Tr denotes the trace of the matrix



```python
# Calculate FVD scores for all combinations
for i, real_video_path in enumerate(real_videos_paths):
    real_video_frames = load_video_frames(real_video_path)
    for j, gen_video_path in enumerate(generated_videos_paths):
        gen_video_frames = load_video_frames(gen_video_path)
        fvd_score = calculate_fvd(real_video_frames, gen_video_frames)
        fvd_scores_matrix[i, j] = fvd_score
```

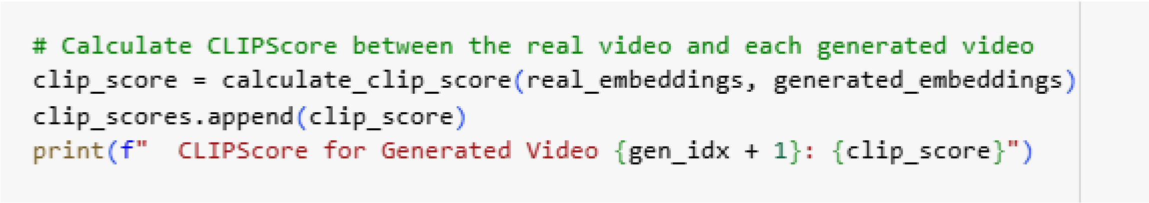
Figure 9. Code snippet of FVD Calculation

CLIPScore

To measure similarity between real and generated videos using CLIP model embeddings. Frames from both videos are processed to extract embeddings, which are high-dimensional representations of the content. Cosine similarity is calculated between the embeddings of corresponding frames, with normalization ensuring unit-length vectors. The average cosine similarity across all frames determines the CLIPScore, reflecting how closely the generated video matches the real one. The provided code extracts embeddings using get_clip_embeddings and computes scores with calculate_clip_score. Figure 10 shows the code snippet for the calculation of clipscore. Equation 3 is the formula of the CLIP score calculation. equation averages the cosine similarities of the embeddings across all frames:

$$CLIPScore = \frac{1}{N}\sum_{i=1}^{N}\frac{E_{\text{real},i}\cdot E_{\text{gen},i}}{|E_{\text{real},i}||E_{\text{gen},i}|} \text{ (Equation 3)}$$

- N is the number of frames
- $E_{\text{real},i}$ is the embedding of the i-th frame of the real video
- $E_{\text{gen},i}$ is the embedding of the i-th frame of the generated video



```python
# Calculate CLIPScore between the real video and each generated video
clip_score = calculate_clip_score(real_embeddings, generated_embeddings)
clip_scores.append(clip_score)
print(f"  CLIPScore for Generated Video {gen_idx + 1}: {clip_score}")
```

Figure 10. Code snippet of Clipscore

Method B Human Perception Study

The study aims to evaluate the human perception of video quality, realism, and accuracy generated by three different text-to-video models: Runway Gen 2 (Model A), CogVideoX-2B (Model B), and CogVideoX-5B (Model C). The primary objective is to compare these models based on three dimensions of perception: Perceived Realism, Perceived Quality, and Perceived Accuracy. A survey was conducted to gather human responses to videos generated by these models based on a predefined set of prompts.

Participant Recruitment

A total of 60 participants completed an online survey administered via Qualtrics. Participants were recruited voluntarily through academic and professional networks. The survey was anonymous, with no collection of demographic or personally identifiable information. As the study involved minimal risk and no personal data, it qualified as exempt research and did not require formal IRB review or informed consent. A brief introductory statement informed participants of the voluntary nature of their participation. The survey questions were designed based on established best practices for Likert-scale usability and perception evaluations. The participants were asked to evaluate the videos based on their subjective perception.

Survey Design

The survey was created using Qualtrics and consisted of 3 identical prompts across the 3 models. The videos for each model were generated based on specific text descriptions, and each video was rated on three parameters:

- Q1: Perceived Realism - Participants rated how realistic the video appeared on a scale from 1 (Less Realistic) to 7 (Highly Realistic).
- Q2: Perceived Quality - Participants rated the overall quality of the video on a scale from 1 (Low Quality) to 7 (High Quality).
- Q3: Perceived Accuracy - Participants rated how accurately the video represented the text description provided on a scale from 1 (Not at all) to 7 (Completely).

Each model was tested across three videos with different prompts. The videos were presented in a random order to mitigate any bias caused by the sequence of presentation.

Videos and Prompts

Each participant was asked to rate the perceived realism, quality, and accuracy of the three videos generated by each model. Three distinct video prompts were used:

- Video 1: "A cat is playing with a yarn on a sofa."
- Video 2: "A panda is eating bamboo."
- Video 3: "A dog is playing with a ball in the field."

These videos were generated by three different models:

- Model A (Runway Gen 2): A generative model for video creation.
- Model B (CogVideoX-2B): Another generative video model with a different architecture.
- Model C (CogVideoX-5B): A further refined version of the CogVideoX model, expected to have improved capabilities.

Statistical Analysis

The collected data was analyzed to compare the ratings between models for each video across the three survey questions. The survey responses were organized and cleaned for analysis. Missing or invalid responses were handled accordingly.

Paired T-Test Analysis

Paired T-tests were conducted to compare responses between the models for each survey questions.
Statistical significance was determined using a two-tailed test with $\alpha = 0.05$. P-values were reported to determine whether model differences in perception were statistically significant.
The following model comparisons were made:

- M1 vs M2: Comparison between Model A (Runway Gen 2) and Model B (CogVideoX-2B).
- M2 vs M3: Comparison between Model B (CogVideoX-2B) and Model C (CogVideoX-5B).
- M3 vs M1: Comparison between Model C (CogVideoX-5B) and Model A (Runway Gen 2).

Hypothesis Testing

Each hypothesis was tested independently for Q1, Q2, and Q3. For each of the three surveys, the following hypotheses were tested:

- Q1 Hypothesis:
  - H_0: There is no significant difference in perceived realism between the models.
  - H_1: There is a significant difference in perceived realism between the models.

- Q2 Hypothesis:
  - H_0: There is no significant difference in perceived quality between the models.
  - H_1: There is a significant difference in perceived quality between the models.
- Q3 Hypothesis:
  - H_0: There is no significant difference in perceived accuracy between the models.
  - H_1: There is a significant difference in perceived accuracy between the models.

The p-value for each test:

- If $p \leq 0.05$, the null hypothesis (H_0) was rejected, indicating a significant difference between the models.
- If $p > 0.05$, the null hypothesis (H_0) was not rejected, indicating no significant difference between the models for that question.

Results

This section reports both objective evaluation metrics (FID, FVD, CLIPScore) and human perception results based on realism, quality, and accuracy. Descriptive outcomes are first presented, followed by statistical comparisons to assess significance.

FID Results

Descriptive findings: Table 1 shows the **FID** scores of three models (Runway Gen-2 as Model A, CogVideoX-2B as Model B, CogVideoX-5B as Model C) across ten generated video sets (V1 to V10), highlighting variations in video quality. Overall, Model A consistently shows higher FID scores, indicating lower quality compared to Models B and C. In notable cases, Model A performs better in V1 (224.62) than Models B (329.46) and C (347.47), suggesting higher video quality for this set. For V2, all models perform similarly with scores between 60 and 66. However, Model A performs poorly in V6 (388.04) compared to Models B (321.02) and C (341.15). In V9, Model B excels with the lowest score (68.50), significantly outperforming Model A (212.10) and Model C (150.16).

Table 1. FID Scores across models

FID Scores			
Videos	**Model A**	**Model B**	**Model C**
V1	224.6161541	329.4603214	347.47433
V2	60.42996281	65.61818585	64.02015358
V3	308.769304	284.6546953	305.2183099
V4	122.8095693	151.2050764	167.9691011
V5	179.6019706	196.2462777	164.0402013
V6	388.042079	321.016455	341.1479454
V7	220.3954802	150.3805624	161.23807
V8	173.2379683	132.3945732	153.918537
V9	212.1018731	68.49737404	150.1646496
V10	129.4089092	211.5387445	227.3118233

Statistical Test results

To determine whether the observed differences in FID scores between the models are statistically significant or simply due to random variation, the T.test is performed between the models. The T-test analysis of FID scores, which measure the quality of generated videos (lower is better), was conducted to test the following hypothesis:

> H_0: There is no significant difference in FID quality between the models.
> H_1: There is a significant difference in FID quality between the models.

Table 2. T-Test of FID

Statistical Analysis (FID)		
Model A vs. Model B	**Model A vs. Model C**	**Model B vs. Model C**
P = 0.6562984681	P= 0.7611937585	**P= 0.08429753595**

The results in table 2, indicate no statistically significant differences for Model A vs. Model B (p = 0.656) and Model A vs. Model C (p = 0.761), supporting the null hypothesis (H_0) in these cases. For the comparison between Model B and Model C, the p-value (0.084) suggests a significant borderline difference, where Model B may slightly outperform Model C.

However, this difference is not strong enough to reject the null hypothesis. Overall, the analysis concludes that there is no significant difference in FID quality between the models, particularly between Model B and Model C, indicating comparable performance in terms of generated video quality.

FVD Results

Descriptive findings: The **FVD** score analysis in Table 3 highlights key performance trends among the models. Model A consistently achieved the lowest FVD scores for V1 (5045.16), V3 (5966.67), and V10 (7976.94), demonstrating better temporal quality in these cases. Model B outperformed in V4 (18055.56) and V5 (9884.65), while Model C frequently had the highest FVD scores, particularly for V1 (19522.04), V3 (20166.44), and V6 (26762.32), indicating weaker performance in maintaining video quality and temporal consistency.

Table 3. FVD Scores across models

FVD Scores			
Videos	**Model A**	**Model B**	**Model C**
V1	5045.16	13165.28	19522.04
V2	19742.81	17299.05	19433.13
V3	5966.67	13638.4	20166.44
V4	39457.36	18055.56	13655.46
V5	20449.64	9884.65	10111.38
V6	9235.5	18490.59	26762.32
V7	18286.62	16206.19	20620.05
V8	11184.1	9095.27	12576.02
V9	25287.53	53800.5	69186.31
V10	7976.94	14411.73	21568.47

Statistical Test results

Statistical T-tests shown in table 4, revealed no significant differences between Model A vs. Model B (p = 0.624) or Model A vs. Model C (p = 0.258). However, the comparison between Model B and Model C (p

= 0.016) showed a statistically significant difference, with Model B performing notably better. These results emphasize Model B's advantage over Model C in terms of FVD-based video quality, providing strong evidence for its superior temporal and visual consistency.

Table 3. T- Test of FVD

Statistical Analysis (FVD)		
Model A vs. Model B	**Model A vs. Model C**	**Model B vs. Model C**
P= 0.6240904552	P= 0.2578538068	**P= 0.01552723295**

CLIP score Results

Descriptive findings: Table 5 reveals that Model A achieved relatively high scores, particularly for V1 (0.8077), V4 (0.8430), and V10 (0.7945), indicating good semantic alignment with the reference text in these cases. Model B demonstrated the highest CLIP scores in V2 (0.8165), V4 (0.8725), and V9 (0.8488), suggesting strong semantic similarity and the best overall performance in aligning with the prompt. Model C also performed well in V4 (0.8728) and V6 (0.8815) but had lower scores in V1 (0.6887) and V10 (0.7104), reflecting weaker alignment with the reference text in those instances.

Table 4. CLIP scores across models

CLIP Scores			
Videos	**Model A**	**Model B**	**Model C**
V1	0.80768472	0.71078885	0.68871158
V2	0.74609685	0.81652713	0.83545774
V3	0.57861584	0.63737339	0.60806930
V4	0.84298593	0.87253356	0.87278634
V5	0.70627141	0.80038387	0.76232082
V6	0.78993267	0.85826993	0.88150454
V7	0.79002547	0.77788723	0.77859783
V8	0.62289506	0.69248575	0.70532203
V9	0.75978786	0.84880829	0.86017573
V10	0.79449558	0.76734495	0.71037650

Statistical Test results

In table 6, the statistical T-test results showed in Table 6 that there were no significant differences between any pair of models, with all p-values (Model A vs. Model B: 0.111, Model A vs. Model C: 0.302, Model B vs. Model C: 0.380) above the 0.05 threshold. Despite these findings, Model B is identified as the most accurate in terms of alignment with the prompt, with Model A showing similar accuracy, while Model C performed the least accurately.

Table 5. T- Test of FVD across models

Statistical Analysis (CLIPscore)		
Model A vs. Model B	**Model A vs. Model C**	**Model B vs. Model C**
P= 0.1110924346	P= 0.3018590847	**P= 0.380125658**

Mathematical and Statistical Evaluation Summary:

- **FID (Quality):** No significant difference in quality between CogVideoX-2B and CogVideoX-5B, indicating similar performance.
- **FVD (Quality):** CogVideoX-2B significantly outperforms CogVideoX-5B in video quality, with lower FVD scores providing strong statistical evidence.
- **Clipscore (Accuracy):** CogVideoX-2B is the most accurate in aligning with the prompt, with Runway Gen-2 performing similarly and CogVideoX-5B the least accurate. Statistical analysis shows no significant difference between the models in terms of CLIP accuracy.

Result of Human Perception Study

Table 7 shows the statistical analysis of human perception:

Table 6. T-Test Statistical analysis of Human Perception Study

Models	Q1: Perceived Realism	Q2: Perceived Quality	Q3: Perceived Accuracy
Model A vs Model B	0.766862	0.112376	**0.000048 (Significant)**
Model B vs Model C	0.112376	**0.000186 (Significant)**	**0.000768 (Significant)**
Model C vs Model A	**0.000048 (Significant)**	0.084168	0.259123

In the human study, the following hypotheses were tested for perceived realism, quality, and accuracy:

Q1: Perceived Realism

For perceived realism, Model C significantly outperformed Model A (p = 0.000048), supporting the alternative hypothesis that there is a significant difference. However, no significant differences were found between Model A and Model B (p = 0.766862), or between Model B and Model C (p = 0.112376), leading to the acceptance of the null hypothesis in these comparisons.

Q2: Perceived Quality

For perceived quality, Model B significantly outperformed Model C (p = 0.000186), indicating a significant difference. No significant difference was observed between Model A and Model B (p = 0.112376), or between Model C and Model A (p = 0.084168), resulting in the acceptance of the null hypothesis for these comparisons.

Q3: Perceived Accuracy

For perceived accuracy, Model A significantly outperformed Model B (p = 0.000048), and Model B significantly outperformed Model C (p = 0.000768). However, no significant difference was found between Model C and Model A (p = 0.259123), leading to the acceptance of the null hypothesis for this comparison.

Human perception study for this video generation models summary:

CogVideoX-5B was rated the best for perceived realism, while CogVideoX-2B outperformed CogVideoX-5B in perceived quality and accuracy. However, Runway Gen-2 performed better than CogVideoX-2B in terms of accuracy.

Discussion

The findings from both human perceptual evaluations and mathematical metrics provide valuable insights into the performance of the models under different evaluation criteria. This research is limited by the data size used for evaluation. A small dataset of video prompts may not fully represent the diverse scenarios required to comprehensively assess model performance. Despite its limitations, this research has a significant impact in advancing the field of text-to-video generation.

In terms of human perception, CogVideoX-5B was regarded as the best model for realism, while CogVideoX-2B emerged as the leader in quality and accuracy. Notably, Runway Gen-2 outperformed CogVideoX-2B in terms of accuracy, which suggests that while CogVideoX-2B excels in quality, Runway Gen-2 offers better alignment with the task at hand for some specific use cases. When considering mathematical evaluations, a different picture emerges.

For FID (quality), paired t-tests show that there is no statistically significant difference between CogVideoX-2B and CogVideoX-5B, suggesting that both models perform similarly in terms of image fidelity. However, FVD (video quality) clearly favors CogVideoX-2B, as it produces a significantly lower score, indicating superior video temporal quality compared to CogVideoX-5B. Additionally, Clipscore (accuracy) corroborates human evaluations, with CogVideoX-2B again leading in accuracy. This aligns with the subjective assessments that ranked CogVideoX-2B the highest for this criterion, followed by Runway Gen-2, while CogVideoX-5B performed the least accurately.

These observations reveal a critical divergence: perceived realism (favoring CogVideoX-5B) does not fully align with objective metrics, which favor CogVideoX-2B. This gap may be attributed to several factors. Human perception is influenced by semantic coherence, emotional resonance, and natural motion, which are not fully captured by metrics like FID or FVD. Conversely, automated metrics may emphasize pixel-level or feature-space accuracy without accounting for perceptual or contextual salience. Such discrepancies emphasize the need for hybrid evaluation frameworks that capture both measurable fidelity and experiential quality.

In sum, while this study highlights consistent strengths in CogVideoX-2B across both technical and perceptual dimensions, the divergence in realism ratings reflects the evolving nature of evaluation in generative video research. Future work should incorporate larger datasets, diverse prompts, validated perceptual tools, and demographic profiling to ensure broader and deeper insights.

Conclusion

A dual evaluation approach proved essential for a fuller understanding of model strengths and weaknesses. This study evaluated three text-to-video models: CogVideoX-2B, CogVideoX-5B, and Runway Gen-2 using both human perception and computational metrics. CogVideoX-5B was preferred for realism based on human ratings, while CogVideoX-2B performed better in quality and accuracy, as supported by FVD and clip score results. These findings suggest that CogVideoX-2B is more balanced in objective performance, whereas CogVideoX-5B stands out in perceptual realism. This divergence underscores the gap between computational metrics and human perception.

A combined evaluation approach enables a more holistic assessment, ensuring that models are both technically sound and perceptually effective. Future research may focus on expanding dataset diversity and enhancing methodological rigor to improve generalizability.

References

Brownlee, J. (2019). How to implement the frechet inception distance (fid) for evaluating gans. *Retrieved Dec, 5*, 2019.

Hong, W., Ding, M., Zheng, W., Liu, X., & Tang, J. (2022). Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*.

Kim, D., Joo, D., & Kim, J. (2020). Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access, 8*, 153113-153122.

Lei, W., Wang, J., Ma, F., Huang, G., & Liu, L. (2024). A Comprehensive Survey on Human Video Generation: Challenges, Methods, and Insights. *arXiv preprint arXiv:2407.08428*.

Leiker, D., Gyllen, A. R., Eldesouky, I., & Cukurova, M. (2023). Generative AI for learning: Investigating the potential of synthetic learning videos. *arXiv preprint arXiv:2304.03784*.

Liao, M., Lu, H., Zhang, X., Wan, F., Wang, T., Zhao, Y., . . . Wang, J. (2024). Evaluation of text-to-video generation models: A dynamics perspective. *arXiv preprint arXiv:2407.01094*.

Liu, Y., Li, L., Ren, S., Gao, R., Li, S., Chen, S., . . . Hou, L. (2024). Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems, 36*.

Oh, G., Jeong, J., Kim, S., Byeon, W., Kim, J., Kim, S., & Kim, S. (2025). *Mevg: Multi-event video generation with text-to-video models.* Paper presented at the European Conference on Computer Vision.

Onisha, T. A., Wimmer, H., & Rebman Jr, C. M. (2024). Facial expressions analysis for deep fake and genuine video recognition. *Issues in Information Systems, 25*(1), 159-174.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., . . . Gafni, O. (2022). Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., & Zhang, S. (2023). Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Yang, S., Zhou, Y., Liu, Z., & Loy, C. C. (2023). *Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation*. Paper presented at the SIGGRAPH Asia 2023 Conference Papers, Sydney, NSW, Australia. https://doi.org/10.1145/3610548.3618160

Zhang, D. J., Wu, J. Z., Liu, J.-W., Zhao, R., Ran, L., Gu, Y., . . . Shou, M. Z. (2024). Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, 1-15.

Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., & Tian, Q. (2023). Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*.