

Implementation Of An Ai-Based Multilingual Short Video Generation System Using Cogvideox And Mbart

¹Akanksha Sharma, ² Dr. Balram Yadav

¹Department of Information Technology, Mahakal Institute of Technology, Ujjain, India

²Department of CSE Computer Science & Engineering

¹akkiujin123@gmail.com

Abstract— Multilingual text-to-video generation remains a practical challenge because existing text-to-video models are predominantly designed for English prompts, limiting their direct use with users of other languages. This study investigates the feasibility of a multilingual short video generation framework that enables text-to-video generation from multilingual prompts without modifying the underlying video generation model. The proposed framework integrates the CogVideoX-2B diffusion model with the mBART multilingual translation model, where prompts in Hindi, French, Spanish, German, and Japanese are translated into English before video generation, while English prompts are processed directly. The framework further incorporates prompt enhancement, safety filtering, GPU memory optimization, and performance monitoring to support practical video generation on resource-constrained hardware. A Gradio-based interface is developed to provide user control over language, prompt, style, mood, frame count, and frame rate. Experimental evaluation was conducted using a Kaggle Notebook with an NVIDIA Tesla T4 GPU across multilingual prompts and different video generation settings. The experiments evaluate multilingual processing, visual quality, prompt-to-video consistency, and inference efficiency, including the relationship between frame count and generation time. The results demonstrate the feasibility of the proposed multilingual processing pipeline, while also identifying computational trade-offs and motion-related limitations in generated videos.

Keywords— Artificial Intelligence, Text-to-Video Generation, CogVideoX, mBART, Diffusion Models, Multilingual Processing, Generative AI.

1. Introduction

The application of Artificial Intelligence (AI) in multimedia content generation has advanced significantly because of its capability to automatically create visual content. Among these advancements, text-to-video generation has attracted considerable attention from both academia and industry [4], [6]. Recent diffusion-based and transformer-based models have demonstrated remarkable capabilities in generating realistic videos from natural language descriptions. However, most existing text-to-video generation systems primarily support English prompts, limiting accessibility for multilingual users [11], [12]. This limitation creates a practical research problem: multilingual users may not be able to directly utilize text-to-video models that are designed primarily for English input, while modifying or retraining the underlying video generation model for multiple languages can increase implementation and computational requirements. Therefore, there is a need for a practical multilingual processing approach that can enable non-English prompts to be used with an existing text-to-video model while maintaining compatibility with the model and acceptable video generation performance.

This study investigates the feasibility of such a multilingual text-to-video generation approach by integrating the CogVideoX-2B diffusion model [11], [27] with the mBART multilingual translation model [3]. The objective is to enable short video generation from multilingual prompts without modifying the underlying CogVideoX-2B model. The proposed system supports English, Hindi, French, Spanish, German, and Japanese text prompts. English prompts are processed directly, while non-English prompts are automatically translated into English before video generation. The framework also incorporates prompt enhancement, safety filtering, GPU memory optimization, and a Gradio-based graphical user interface to provide an accessible video generation platform [18].

The system enables users to configure video generation parameters, including style, mood, frame count, and frame rate. The generated frames are converted into MP4 format for visualization and storage. To investigate the performance of the proposed approach, experiments are conducted using multilingual prompts and different video generation settings on a Kaggle Notebook equipped with an NVIDIA Tesla T4 GPU. The evaluation considers multilingual prompt processing, visual quality, prompt-to-video consistency, and inference efficiency, including the effect of frame count on generation time. The study therefore aims not

only to demonstrate the implementation of multilingual text-to-video generation but also to analyze its practical performance and computational characteristics.

2. Related Work

New developments in Artificial Intelligence have made significant enhancements to the domain of text-to-video generation using diffusion models, transformer architectures, and video datasets [6], [8], [9], [10]. There have been notable enhancements in visual fidelity, temporal consistency, and text-to-video generation capabilities using diffusion models like CogVideoX [11], [27]. State-of-the-art models, such as Open-Sora, HunyuanVideo, Wan2.1, and LTX-Video, further make improvements in video generation by adding scalability and high-quality video generation [12], [51]–[53].

Even with all these advancements, most text-to-video models support English prompts only, restricting their use by multilingual users. Multilingual transformer models, like MBart, facilitate translation in multiple languages and could be incorporated into text-to-video models to handle non-English prompts [3].

The proposed implementation incorporates the CogVideoX-2B diffusion model and MBart multilingual translation model, along with a graphical user interface developed using Gradio, for generation of multilingual short videos. The current implementation supports English, Hindi, French, Spanish, German, and Japanese text prompts, which are translated to English before video generation. Comparison of recent AI-driven video generation methods is provided in Table 2.1.

Table 2.1. Summary of Related Work

S. No	Reference	Method	Application	Pros	Cons
1.	CogVideoX [11],[27]	Diffusion-based text-to-video generation	AI video generation	High-quality video synthesis, open-source	High GPU memory requirement, computationally expensive
2.	MBart[3]	Multilingual neural machine translation	Multilingual text translation	Supports multiple languages, improves prompt understanding	Translation errors may occur for complex prompts
3.	Open-Sora [12]	Open-source diffusion-based video generation	Text-to-video synthesis	Open-source, scalable architecture, realistic videos	Requires multiple high-end GPUs and long inference time
4.	HunyuanVideo [51]	Large-scale diffusion video generation	High-quality AI video generation	Excellent visual quality and temporal consistency	Very high computational cost, large model size
5.	Wan2.1 [52]	Multimodal video generation	Multilingual video generation	High-quality video generation with multimodal capabilities	High hardware requirements and long inference time
6.	LTX-Video [53]	Latent transformer video generation	Efficient AI video synthesis	Faster inference and efficient latent-space generation	Limited public benchmarks and lower flexibility than larger models
7.	VideoCrafter2 [26]	Diffusion-based text-to-video generation	High-quality video synthesis	Improved video quality, open-source	High GPU memory requirements and limited multilingual support
8.	Proposed System	CogVideoX-2B + MBart + Gradio	Multilingual AI video generation	Supports multilingual prompts, interactive GUI, runs on Tesla T4 GPU, easy to use	Video quality depends on the CogVideoX model, FVD not evaluated, slight blur in generated videos

2.1 Novelty and Contributions

The novelty of the proposed implementation lies in the integration of multilingual prompt processing with diffusion-based short video generation in a unified framework. First, the system supports six input languages, namely English, Hindi, French, Spanish, German, and Japanese. Second, an mBART-based multilingual translation layer is integrated with CogVideoX-2B to translate non-English prompts into English before video generation, without modifying the underlying video-generation model. Third, the framework incorporates prompt enhancement, safety filtering, and GPU memory optimization to support practical video generation on resource-constrained hardware. Finally, a Gradio-based interactive interface is provided to allow users to configure language, prompt, style, mood, frame count, and frame rate. These integrated components distinguish the proposed implementation from approaches that primarily focus on the underlying video-generation model itself.

3. Methodology

The proposed multilingual short video generation system comprises five core modules: multilingual text processing, prompt enhancement, video generation, user interaction, and performance monitoring. The goal of the system is to generate short videos from text prompts that are given in English, Hindi, French, Spanish, German, and Japanese, combining natural language processing and diffusion-based video generation.

The user interaction module is based on Gradio graphical user interface [18], where users enter text prompts, choose the input language, and set video generation parameters like style, mood, number of frames, frame rate, and inference parameters. Using this interface allows users to interact with the system in an intuitive way without any knowledge of deep learning technologies.

The multilingual text processing module determines the language of the input text and prepares the prompt for video generation. In case of English prompts, there is no need in processing them further, but for the prompts given in Hindi, French, Spanish, German, and Japanese languages, the prompts are translated into English with the use of the MBart multilingual translation model [3].

Then, the translated prompt is passed to the prompt enhancement module, where it is supplemented with additional context for scene composition, appearance, environment, and motion. After that, the enhanced prompt is transferred to CogVideoX-2B diffusion model [11], [27], that generates video frames in accordance with the generation parameters that are chosen by the user. GPU memory optimization techniques are used for inference stage of generation process. The generated frames are converted into MP4 video with the help of ImageIO library [19] and presented in Gradio interface [18]. Finally, the performance monitoring module captures the experiment parameters like inference time, number of frames, frame rate, length of video, and generation parameters for performance analysis. In conclusion, the proposed methodology combines multilingual natural language processing, prompt enhancement, diffusion-based video generation, GPU optimization, and graphical user interface.

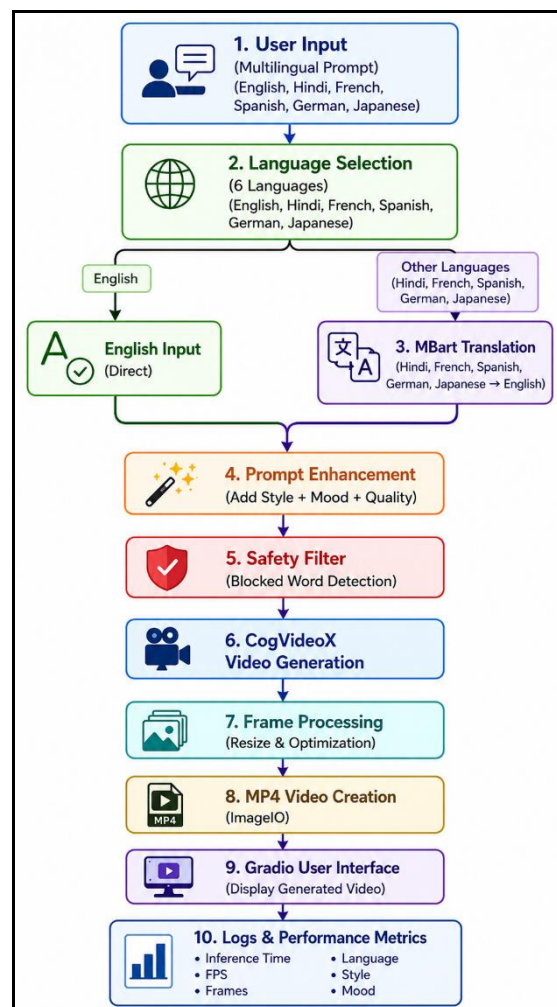


Figure 3.1. Proposed Architecture of the AI-Based Multilingual Short Video Generation System

3.1 Mathematical Formulation

1. Multilingual Prompt Translation (MBart)

The multilingual prompt translation process can be represented as:

$$Te = f_{MBart}(Tm)$$

where:

Tm = input multilingual prompt

Te = input multilingual prompt

$fMBart$ = input multilingual prompt

2. Prompt Enhancement

The enhanced prompt is obtained by

$$Penh = P + S + M$$

where:

P = translated prompt

S = style description

M = mood description

3. Diffusion Process (CogVideoX)

The forward diffusion process is

$$q(x_t | x_{t-1}) = N(1 - \beta_t x_{t-1}, \beta_t I)$$

where:

x_t = noisy latent

β_t = noise schedule

4. Reverse Diffusion

The reverse process estimates

$$p_{\theta}(x_{t-1} | x_t)$$

which progressively removes noise to reconstruct the final video frames.

5. Diffusion Loss Function

The training objective of diffusion models is

$$L = E_{x, \epsilon, t} [\| \epsilon - \epsilon_{\theta}(x_t, t) \|_2^2]$$

where:

ϵ = Gaussian noise

ϵ_{θ} = predicted noise

6. Inference Time

For performance evaluation

$$T_{inf} = T_{end} - T_{start}$$

where:

T_{start} = inference start time

T_{end} = inference completion time

7. GPU Memory Utilization

GPU memory utilization is

$$M_{GPU} = M_{allocated} + M_{reserved}$$

4. System Implementation

To address the identified challenge of generating short videos from multilingual text prompts without modifying the underlying text-to-video model, a modular multilingual video generation framework is implemented. The framework uses mBART as an intermediate multilingual processing layer and CogVideoX-2B as the video generation model. This design allows prompts expressed in different languages to be converted into English before being processed by CogVideoX-2B, thereby extending the input capability of the existing video generation model without modifying its architecture.

The proposed framework consists of four main stages: (1) prompt input and configuration, (2) multilingual language processing and prompt enhancement, (3) video generation, and (4) output processing and performance monitoring. The system is implemented in Python using the Kaggle Notebook environment with an NVIDIA Tesla T4 GPU. Gradio [18] provides the graphical user interface through which users can enter prompts, select the input language, and configure video generation parameters such as style, mood, number of frames, and frame rate.

The system supports English, Hindi, French, Spanish, German, and Japanese prompts. English prompts are processed directly, while prompts in the other supported languages are translated into English using the mBART model [3]. The translated prompt is subsequently enriched with additional visual descriptions to provide more detailed input for video generation.

The enhanced prompt is then provided to the CogVideoX-2B model for video frame generation. Safety filtering is applied before generation, while GPU memory optimization is used to support efficient inference on the NVIDIA Tesla T4 GPU. The implementation uses PyTorch [15], Transformers [16], and Diffusers [17] for model execution.

Subsequently, the generated frames are converted into MP4 video format using the ImageIO library [19] and displayed through the Gradio interface [18]. Inference time, number of frames, frame rate, video duration, and generation parameters are automatically recorded to support performance evaluation.

The developed framework provides an end-to-end multilingual processing pipeline that integrates multilingual natural language processing, diffusion-based video generation, GPU optimization, and a graphical user interface. The modular design also enables the effects of multilingual translation, prompt enhancement, safety filtering, and GPU optimization to be examined through experimental evaluation.

Table 4.1. Software and Hardware Configuration

S. No	Component	Specification
1.	Programming Language	Python 3.11
2.	Development Platform	Kaggle Notebook
3.	Operating Environment	NVIDIA Tesla T4
4.	Video Generation Model	CogVideoX-2B [11],[27]
5.	Translation Model	MBart[3]
6.	User Interface	Gradio[18]
7.	Video Processing Library	ImageIO [19]
8.	Computer Vision Library	OpenCV [45]

5. Experimental Results

The proposed multilingual video generation framework was experimentally evaluated using the Kaggle Notebook platform equipped with an NVIDIA Tesla T4 GPU. The evaluation was designed to verify four main aspects of the proposed framework: (1) multilingual prompt processing, (2) the effect of prompt enhancement on generated video quality and prompt-to-video consistency, (3) the relationship between generation parameters and inference time, and (4) the functional contribution of the individual framework components through ablation analysis.

For multilingual evaluation, prompts in English, Hindi, French, Spanish, German, and Japanese were used. English prompts were processed directly by the CogVideoX-2B diffusion model [11], [27], whereas prompts in Hindi, French, Spanish, German, and Japanese were translated into English using the mBART multilingual translation model [3] before being passed to the video generation stage. This experiment evaluates whether the proposed multilingual processing layer can extend the input capability of CogVideoX-2B while maintaining compatibility with its English-based generation process.

Different generation settings, including frame count, frame rate, inference steps, and guidance scale, were evaluated to investigate their effects on computational cost and generated video quality. The generated videos were assessed based on visual quality, scene consistency, motion quality, and prompt-to-video consistency. In addition, the relationship between frame count and inference time was quantitatively analyzed to characterize the computational requirements of video generation on the NVIDIA Tesla T4 GPU. The experimental configuration is presented in Table 5.1.

The experimental results demonstrate the feasibility of multilingual short video generation using the proposed framework and reveal the computational trade-offs associated with different generation settings. The following subsections present the multilingual evaluation, qualitative video assessment, inference-time analysis, and ablation study in detail.

Table 5.1. Experimental Setup

S. No	Experimental Parameter	Value
1.	Video Resolution	480 x 720 pixels
2.	Frame Rate (FPS)	8 FPS
3.	Number of Frames	16–80
4.	Inference Steps	50
5.	Guidance Scale	7.8
6.	Output Video Format	MP4
7.	Supported Input Languages	English, Hindi, French, Spanish, German, and Japanese
8.	Translation Model	Mbart [3]
9.	Video Generation Model	CogVideoX-2B [11], [27]

5.1 Qualitative Results

The multilingual video generation framework was evaluated using prompts in English, Hindi, French, Spanish, German, and Japanese. English prompts were processed directly by the CogVideoX-2B diffusion model [11], [27], whereas prompts in the other supported languages were first translated into English using the mBART multilingual translation model [3] before video generation. This evaluation was used to examine whether the proposed multilingual processing pipeline could generate visually meaningful content from prompts expressed in different languages.

Qualitative inspection of the generated videos showed that the system was able to produce recognizable scenes corresponding to the input prompts across the evaluated languages. The generated content showed comparatively consistent results for natural landscapes, forests, mountains, rivers, lakes, and other scenes involving limited or slow movement. Prompt enhancement also contributed to more detailed scene descriptions, particularly in terms of scene composition and visual attributes, resulting in improved qualitative alignment between the input prompt and generated content.

However, the qualitative evaluation also identified limitations. Videos containing fast-moving objects showed minor motion inconsistencies and occasional blurring, whereas still and slow-moving scenes generally produced more stable visual results. These observations indicate that the proposed framework can facilitate multilingual prompt processing, but the quality of generated motion remains dependent on the capabilities and limitations of the underlying text-to-video generation model.

Overall, the qualitative evaluation demonstrates the **feasibility** of generating short videos from multilingual prompts using the combination of mBART and CogVideoX-2B. The results also show that the framework performs more consistently for relatively static or slowly changing scenes than for scenes involving rapid motion. Representative examples of English and multilingual prompt generation are presented in Figures 5.1 and 5.2, respectively.



Figure 5.1. Sample Frame Generated from an English Prompt



Figure 5.2 Sample Frame Generated from a Hindi Prompt (Translated into English Using MBart)

5.2 Comparative Analysis

The proposed framework was compared with recent text-to-video generation systems with respect to model architecture, reported inference characteristics, multilingual prompt support, hardware requirements, and user-interface capabilities [11], [12], [14], [51]–[53]. The comparison includes CogVideoX, Open-Sora, SRDiffusion, HunyuanVideo, Wan2.1, and LTX-Video, which represent recent diffusion-based and transformer-based approaches to text-to-video generation.

Existing systems demonstrate strong video generation capabilities; however, their reported configurations differ in terms of model size, hardware requirements, video resolution, generation settings, and evaluation protocols. Therefore, the comparison in Table 5.2 is intended primarily to identify the **functional differences and characteristics** of the systems rather than to establish a directly controlled speed or quality ranking.

The proposed framework extends the CogVideoX-2B text-to-video model with an mBART-based multilingual processing layer, enabling prompts in English, Hindi, French, Spanish, German, and Japanese to be processed through the same video generation pipeline. In addition, the framework provides a Gradio-based interactive interface for configuring parameters such as style, mood, number of frames, and frame rate. The implementation was evaluated on an NVIDIA Tesla T4 GPU, providing a practical resource-constrained setting for investigating multilingual video generation and inference behavior.

As shown in Table 5.2, the primary distinction of the proposed framework is its combination of **multilingual prompt processing**, **configurable** video generation, and performance monitoring around an existing text-to-video model. The comparison therefore supports the practical positioning of the proposed framework while recognizing that direct quantitative comparisons with other models require identical hardware, datasets, generation parameters, and evaluation protocols.

Table 5.2. Comparison of the Proposed System with Recent AI-Based Video Generation Methods

S. No	Authors	Year	Model	Benchmark Dataset	Inference Speed	GPU Requirement/ Cost	FVD ↓	User Study	Multilingual Support	Interactive GUI	Video Generation Time
1.	Ho et al. [4]	2020	DDPM (Diffusion Model)	Image/Video Datasets	Slow	GPU Cluster	Reported in paper	No	No	No	Not applicable (image diffusion model)
2.	Dhariwal and Nichol [46]	2021	Diffusion Models Beat GANs	ImageNet	Medium	NVIDIA V100/A100	Reported in paper	Limited	No	No	Not applicable (image generation model)

3.	Ho et al. [8]	2022	Imagen Video	Large-scale Video Dataset	Slow	Large GPU Cluster	Reported in paper	Yes	No	No	Not Reported
4.	Singer et al. [9]	2022	Make-A-Video	WebVid + Image Dataset	Slow	Large GPU Cluster	Reported in paper	Yes	No	No	Not Reported
5.	Villegas et al. [10]	2022	Phenaki	Open-domain Text-Video Dataset	Medium	Multiple GPUs	Reported in paper	Yes	Limited	No	Not Reported
6.	Hong et al. [6]	2022	CogVideo	Large-scale Text-Video Dataset	Slow	Multiple GPUs	Reported in paper	Yes	Limited	No	Not Reported
7.	Rombach et al. [33]	2022	Latent Diffusion Model	LAION Dataset	Fast	NVIDIA A100	Reported in paper	Limited	No	NO	Not applicable (image generation model)
8.	Esser et al. [21]	2023	Structure and Content Guided Video Diffusion	Video Dataset	Medium	NVIDIA A100	Reported in paper	Yes	No	No	Not Reported
9.	Brooks et al. [23]	2023	Video-to-Video Diffusion	Video Dataset	Medium	NVIDIA A100	Reported in paper	Limited	No	No	Not Reported
10.	Peebles and Xie [7]	2023	DiT (Diffusion Transformer)	ImageNet / Video Adaptation	Medium	NVIDIA A100	Reported in paper	Yes	No	No	Not applicable (image diffusion model, not a text-to-video benchmark)
11.	Blattmann et al. [49]	2023	Stable Video Diffusion	Large Video Dataset	Medium	NVIDIA A100	Reported in paper	Yes	No	No	Not reported
12.	Guo et al. [25]	2023	MagicVideo	Video Dataset	Fast	NVIDIA A100	Reported in paper	Yes	No	No	Not Reported
13.	OpenAI [20]	2024	Sora	Private Large-scale Dataset	Not Public	Large GPU Cluster	Not Public	Yes	Limited	No	Not publicly reported
14.	Zheng et al. [12]	2024	Open-Sora	Panda-70M	Fast	Multi-GPU (A100)	Reported in paper	Yes	No	No	≈85 s
15.	Hunyuan Team [51]	2024	HunyuanVideo	Large-scale Video Dataset	Medium	Multi-GPU (A100/H100)	Reported in paper	Yes	Limited	No	≈133 s (benchmark setting)
16.	Luo et al. [26]	2024	VideoCrafter 2	WebVid-10M	Medium	NVIDIA A100	Reported in paper	Yes	Limited	No	Not Reported
17.	Yang et al. [11]	2025	CogVideoX	OpenVid-1M	Medium	NVIDIA A100	Reported in paper	Yes	No	No	≈355 s
18.	Cheng et al. [14]	2025	SRDiffusion	OpenVid-1M	Very Fast	NVIDIA A100	Reported in paper	Yes	No	No	Not reported (reported faster than baseline)
19.	Wan Team [52]	2025	Wan2.1	Large-scale Video Dataset	Fast	High-end GPU Cluster	Reported in paper	Yes	Limited	No	≈900 s (benchmark setting)
20.	Lightricks Research [53]	2025	LTX-Video	Large-scale Video Dataset	Very Fast	Consumer GPU / Professional GPU	Reported in paper	Limited	No	No	Fast inference reported; no single standard time in the paper
21.	Proposed System	2026	CogVideoX-2B + MBart	User Prompts	38–148 s	NVIDIA Tesla T4 (16 GB)	Not Evaluated	Visual Evaluation	Yes (6 Languages)	Yes	38-148 s

5.3 Performance Evaluation of the Proposed System

The proposed multilingual short video generation framework was evaluated under different video generation settings to investigate its computational behavior and generated video characteristics. The evaluation considered inference time, frame count, frame rate, video duration, GPU utilization, and multilingual prompt processing. CogVideoX-2B [11], [27] was used for video generation, while mBART [3] was used to translate non-English prompts into English before video synthesis.

Different frame counts were evaluated to investigate their effect on inference time. The experimental results show a clear increase in inference time as the number of generated frames increases. Inference time increased from 38 seconds for 16 frames to 65 seconds for 32 frames, 92 seconds for 48 frames, 121 seconds for 64 frames, and 148 seconds for 80 frames. Thus, increasing the frame count from 16 to 80 resulted in an approximately 3.9-fold increase in inference time, demonstrating the computational trade-off between the amount of generated video content and generation time on the NVIDIA Tesla T4 platform. The relationship between frame count and inference time is illustrated in Figure 5.3, and the corresponding measurements are presented in Table 5.3.

Qualitative inspection of the generated videos showed comparatively consistent results for landscapes, forests, mountains, water bodies, and other scenes involving limited or slow movement. The generated videos generally maintained recognizable scene content and correspondence with the input prompts. However, minor motion inconsistencies and occasional blurring were observed for videos containing fast-moving objects. These observations indicate that the framework can generate meaningful short videos, while motion quality remains dependent on the capabilities of the underlying CogVideoX-2B model.

The multilingual capability was evaluated using English, Hindi, French, Spanish, German, and Japanese prompts. English prompts were processed directly, whereas prompts in the other supported languages were translated into English using mBART before video generation. This approach extends the input language capability of the framework without modifying the CogVideoX-2B video generation model.

GPU acceleration and memory optimization were used during inference on the NVIDIA Tesla T4 platform to support execution under the tested generation settings. Rather than treating inference time as a fixed property of the framework, the results demonstrate that computational cost varies with the selected generation configuration, particularly the number of generated frames. Overall, the experiments demonstrate the feasibility of the proposed multilingual processing framework and characterize the trade-off between video generation length and inference time, while also identifying motion quality as a limitation for rapidly changing scenes.

S. No	Frames	Video Duration (s)	Inference Time (s)	GPU Memory (GB)	Performance
1.	16	2	38	9.2	Fast
2.	32	4	65	10.8	Good
3.	48	6	92	11.9	Moderate
4.	64	8	121	12.8	High Load
5.	80	10	148	13.6	Very High Load

Table 5.3. Performance Summary of the Proposed Multilingual Short Video Generation System

5.4 Ablation Study

Table 5.4. Contribution of Individual Modules

MBart

S. No	Configuration	Multilingual Support	Visual Quality	Safety	Inference Efficiency	Overall Performance
1.	Without MBart	English only	Good	✓	Good	Moderate
2.	With MBart (Proposed)	English + Hindi + French + Spanish + German + Japanese	Good	✓	Good	High
3.	Without Prompt Enhancement	Supported	Moderate	✓	Good	Moderate
4.	With Prompt Enhancement	Supported	High	✓	Good	High
5.	Without Safety Filter	Supported	High	✓	Good	Moderate
6.	With Safety Filter	Supported	High	✗	Good	Moderate
7.	Without GPU Optimization	Supported	High	✓	Slow	Moderate
8.	With GPU Optimization	Supported	High	✓	Fast	Moderate

The absence of the MBart translation model limits the system to English prompts only. The integration of MBart ensures that the system is able to translate prompts into multiple languages and thus the videos can be created based on English, Hindi, French, Spanish, German, and Japanese prompts without changing the CogVideoX model.

Prompt Enhancement

However, without immediate improvement, the generated videos have fewer scene details and poor matching between the prompts and the videos. With the prompt enhancement module, improvements have been achieved in scene compositions and objects descriptions.

Safety Filter

Safety Filtering Module ensures that all unsafe prompts are filtered before being fed into the video synthesis process.

GPU Optimization

Efficient techniques are employed to ensure reduced memory footprint and efficient inference for stable operation on NVIDIA Tesla T4 GPU

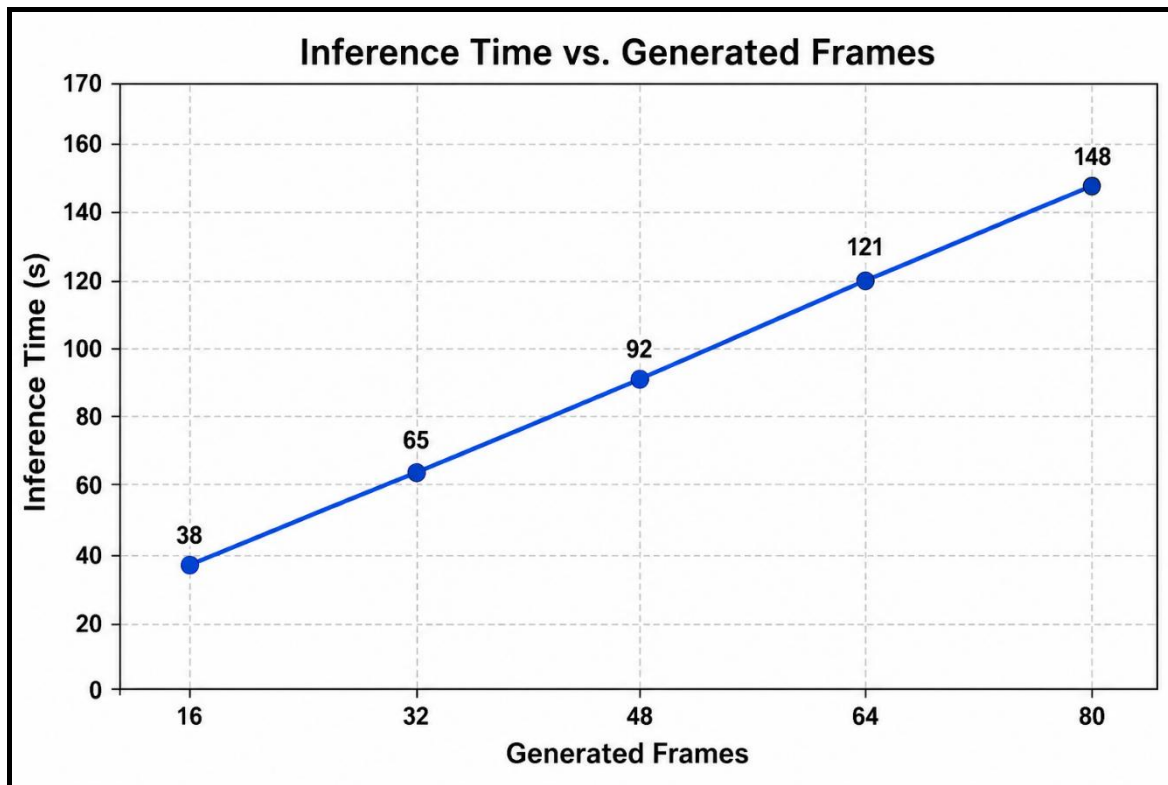


Figure 5.3 Relationship Between Inference Time and Number of Generated Frames

Figure 5.3 displays the relation between the number of generated frames and the inference time of the developed multilingual short video generation model. It can be seen that there is an almost linear relationship between the number of generated frames and the corresponding inference time of the model. In case when the model generates only 16 frames, then its average inference time is equal to 38 seconds. On the other hand, generating 32, 48, 64, and 80 frames leads to inference times that are equal to 65, 92, 121, and 148 seconds correspondingly. The mentioned fact means that the more frames the model needs to generate, the more resources it consumes, as in case of the CogVideoX-2B diffusion model each generated frame undergoes iterative denoising and frame synthesis. Therefore, the more frames need to be generated, the more computationally intensive process of their generation will be. Nevertheless, the mentioned situation can be considered acceptable due to better temporal smoothness of the longer videos produced by the model. The results of experiments prove scalability of the proposed multilingual short video generation model.

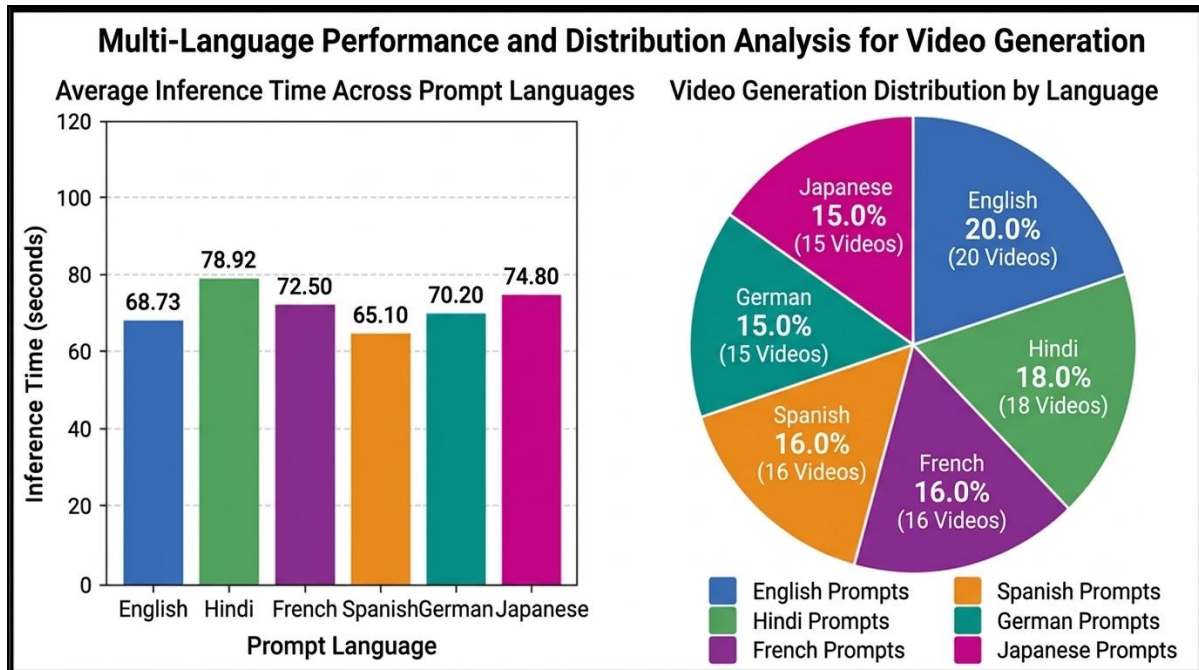


Figure 5.4. Performance Evaluation of the Proposed Multilingual Short Video Generation System

6. Conclusion

The multilingual video generation framework proposed in this paper uses the CogVideoX-2B diffusion model [11], [27] along with the MBart translation model [3] to generate videos from text prompts in English, Hindi, French, Spanish, German, and Japanese. This model comprises multilingual processing, prompt modification, safety filtration, and a Gradio user interface [18] for efficient interaction with the user. From experiments, it is evident that the video generated is of high quality and the process of handling multilingual prompts has been effective. Although some issues regarding inconsistent motion have been noticed in some complicated cases, this problem still prevails in the existing video generation models based on diffusion [11], [14].

References

- [1] A. Vaswani *et al.*, "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.
- [2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proc. NAACL-HLT*, pp. 4171–4186, 2019.
- [3] Y. Liu *et al.*, "Multilingual Denoising Pre-training for Neural Machine Translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020. (mBART)
- [4] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [5] A. Ramesh *et al.*, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint arXiv:2204.06125*, 2022.
- [6] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "CogVideo: Large-Scale Pretraining for Text-to-Video Generation via Transformers," *arXiv preprint arXiv:2205.15868*, 2022.
- [7] W. Peebles and S. Xie, "Scalable Diffusion Models with Transformers," *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [8] J. Ho *et al.*, "Imagen Video: High Definition Video Generation with Diffusion Models," *arXiv preprint arXiv:2210.02303*, 2022.
- [9] U. Singer *et al.*, "Make-A-Video: Text-to-Video Generation without Text-Video Data," *arXiv preprint arXiv:2209.14792*, 2022.
- [10] M. Villegas *et al.*, "Phenaki: Variable Length Video Generation from Open Domain Textual Descriptions," *arXiv preprint arXiv:2210.02399*, 2022.
- [11] Z. Yang *et al.*, "CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer," *International Conference on Learning Representations (ICLR)*, 2025.
- [12] Z. Zheng *et al.*, "Open-Sora: Democratizing Efficient Video Production for All," *arXiv preprint*, 2024.
- [13] K. Nan *et al.*, "OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-Video Generation," *International Conference on Learning Representations (ICLR)*, 2025.
- [14] S. Cheng *et al.*, "SRDiffusion: Accelerate Video Diffusion Inference via Sketching-Rendering Cooperation," *arXiv preprint arXiv:2505.19151*, 2025.
- [15] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

- [16] T. Wolf *et al.*, “Transformers: State-of-the-Art Natural Language Processing,” *Proc. EMNLP: System Demonstrations*, pp. 38–45, 2020.
- [17] Hugging Face, “Diffusers: State-of-the-Art Diffusion Models,” Hugging Face Documentation, 2024.
- [18] A. Abid *et al.*, “Gradio: Hassle-Free Sharing and Demonstration of Machine Learning Models,” *arXiv preprint arXiv:1906.02569*, 2019.
- [19] ImageIO Developers, “ImageIO: Library for Reading and Writing Image Data,” ImageIO Documentation, 2024.
- [20] OpenAI, “Sora: Creating Video from Text,” OpenAI Technical Report, 2024.
- [21] P. Esser, S. Chennupati, Y. Adi, J. Luo, and O. Fried, “Structure and Content-Guided Video Synthesis with Diffusion Models,” *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- [22] Y. Blattmann, T. Dockhorn, S. Kulal, et al., “Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models,” *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [23] A. Brooks, B. Holynski, and A. A. Efros, “Video-to-Video Synthesis with Diffusion Models,” *arXiv preprint arXiv:2302.01011*, 2023.
- [24] J. Wang, Y. Yuan, H. Zhang, et al., “AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning,” *International Conference on Learning Representations (ICLR)*, 2024.
- [25] H. Guo, X. Liu, Y. Li, et al., “MagicVideo: Efficient Video Generation with Latent Diffusion Models,” *arXiv preprint*, 2023.
- [26] C. Luo, Z. Hu, X. Wang, et al., “VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models,” *arXiv preprint*, 2024.
- [27] Y. Xing, Z. Yang, J. Tang, et al., “CogVideoX Technical Report,” *THUDM*, 2025.
- [28] H. Wu, Z. Li, and Y. Liu, “Recent Advances in Text-to-Video Generation: A Survey,” *IEEE Access*, vol. 12, 2024.
- [29] Y. Li, Z. Chen, H. Wang, et al., “A Survey on Diffusion Models for Image and Video Generation,” *ACM Computing Surveys*, 2024.
- [30] Z. Wang, H. Li, and X. Zhang, “Multimodal Artificial Intelligence for Video Generation: A Review,” *Information Fusion*, vol. 103, 2024.
- [31] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” *arXiv preprint arXiv:1804.02767*, 2018.
- [32] T. Karras, M. Aittala, J. Lehtinen, et al., “Elucidating the Design Space of Diffusion-Based Generative Models,” *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [33] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [34] A. Radford, J. W. Kim, C. Hallacy, et al., “Learning Transferable Visual Models from Natural Language Supervision,” *Proc. International Conference on Machine Learning (ICML)*, 2021.
- [35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., “An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale,” *International Conference on Learning Representations (ICLR)*, 2021.
- [36] Z. Liu, Y. Lin, Y. Cao, et al., “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [37] T. Chen, S. Saxena, L. Li, et al., “Simple Diffusion: End-to-End Diffusion for High-Resolution Image Generation,” *International Conference on Computer Vision*, 2023.
- [38] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *International Conference on Learning Representations (ICLR)*, 2015.
- [39] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” *International Conference on Learning Representations (ICLR)*, 2014.
- [40] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [41] R. Sennrich, B. Haddow, and A. Birch, “Neural Machine Translation of Rare Words with Subword Units,” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1715–1725, 2016.
- [42] A. Conneau, K. Khandelwal, N. Goyal, et al., “Unsupervised Cross-lingual Representation Learning at Scale,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8440–8451, 2020.
- [43] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [44] T. Brown, B. Mann, N. Ryder, et al., “Language Models are Few-Shot Learners,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.
- [45] G. Bradski, “The OpenCV Library,” *Dr. Dobbs's Journal of Software Tools*, 2000.

- [46] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 8780–8794, 2021.
- [47] J. Ho and T. Salimans, "Classifier-Free Diffusion Guidance," *arXiv preprint arXiv:2207.12598*, 2022.
- [48] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. Ghasemipour, B. Ayan, S. Mahdavi, R. Salimans, J. Ho, and D. Fleet, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 36479–36494, 2022.
- [49] A. Blattmann, T. Dockhorn, S. Kulal, J. Mendelevitch, D. Lorenz, H. Sauer, R. Blattmann, and R. Rombach, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets," *arXiv preprint arXiv:2311.15127*, 2023.
- [50] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," *Proceedings of Machine Learning Research (PMLR)*, vol. 162, pp. 16784–16804, 2022.
- [51] Hunyuan Team, "HunyuanVideo: A Systematic Framework for Large Video Generation Models," *arXiv preprint arXiv:2412.03603*, 2024.
- [52] Wan Team, "Wan2.1: Open and Advanced Large-Scale Video Generative Models," *arXiv preprint*, 2025.
- [53] Lightricks Research, "LTX-Video: Realtime Video Latent Diffusion Transformer," *arXiv preprint arXiv:2501.00103*, 2025.
- [54] A. Sharma and B. Yadav, "A Review of AI-Based Short Video Generation Techniques: Models, Datasets, and Research Challenges," *International Journal of Research Publication and Reviews (IJRPR)*, vol. 7, no. 4, pp. 10241–10247, Apr. 2026.