

AI-Driven Text-to-Multimedia Content Generation: Enhancing Modern Content Creation

Dhiraj Jadhav

Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, INDIA

dhiraj.jadhav@vit.edu

Sakshee Agrawal

Department of Computer Engineering
Vishwakarma Institute of Technology,
Pune, INDIA

sakshee.agrawal21@vit.edu

Sakshi Jagdale

Department of Computer Engineering
Vishwakarma Institute of Technology,
Pune, INDIA

ashok.sakshi21@vit.edu

Pranav Salunkhe

Department of Computer Engineering
Vishwakarma Institute of Technology,
Pune, INDIA

pranav.salunkhe21@vit.edu

Raj Salunkhe

Department of Computer Engineering
Vishwakarma Institute of Technology,
Pune, INDIA

raj.salunkhe21@vit.edu

Abstract — This research investigates the use of artificial intelligence for the creation of multimedia content from textual sources, such as press articles, stories, and scripts. The system utilizes Natural Language Processing and image generation techniques to create relevant visuals and voiceovers, enhancing content creation efficiency without sacrificing quality. The system processes various input formats, using Optical Character Recognition (OCR) for non-text formats and advanced large language models (LLMs) like LLaMA2 and Gemini for text summarization. It automatically extracts keywords and uses them to generate synchronized visual and audio content, utilizing tools like Google Text-to-Speech (GTTS) and AI-based image generation models such as DALL-E and Stable Diffusion. The multimedia content is assembled and rendered using MoviePy, ensuring seamless integration of text, audio, visuals, and subtitles. Applications in journalism, marketing, and education are examined to showcase how this technology improves accessibility and audience engagement. Overall, the study demonstrates the potential of AI in automating media content creation, streamlining the process while maintaining narrative coherence and enhancing the user experience.

Keywords— *Media Content Creation, Audience Engagement, Text-to-Video Generation, Natural Language Processing (NLP), Computer Vision, Optical Character Recognition (OCR), Large Language Models (LLMs).*

I. INTRODUCTION

In today's fast-paced digital world, the demand for engaging multimedia content is higher than ever. With the rise of social media, online news platforms, and video-sharing websites, consumers are becoming more interested in visual content that grabs their attention and provides information rapidly. Because of this, writers, journalists, and marketers are looking for creative ways to turn written materials—like press articles, narratives, and scripts—into multimedia content that not only captures attention but also delivers information efficiently. This requirement has resulted in tremendous technological developments, especially in artificial intelligence (AI).

Automation of text-to-multimedia generation is one of the most promising advances in this field. This technology makes it easier for creators to reach their audiences by converting written material into captivating multimedia content. This system analyzes text, extracts key information, and generates

corresponding visuals, animations, and voiceovers using Natural Language Processing (NLP) and computer vision. It uses Google Text-to-Speech (GTTS) for natural-sounding narration and powerful Large Language Models (LLMs) such as LLaMA2 and Gemini for efficient summarization. Additionally, web scraping and AI-driven image generation ensure that each piece of multimedia content visually corresponds to the text. This streamlined technique not only improves the end product's quality, but it also expands the reach of information, catering to the growing need for multimedia content.

The ability to quickly create multimedia content from text has multiple applications across various disciplines. For example, automated multimedia content generation in journalism can help news organizations provide viewers with more engaging and timely updates. News organizations can reach a wider audience by turning their articles into visual content, particularly for individuals who prefer acquiring knowledge through visual media. Similar to this, companies can use customer testimonials or product descriptions to produce promotional movies for marketing purposes, which increases the attraction and popularity of their product. Another field where text-to-multimedia creation can have a big impact is education. Educators can transform complex concepts into visually appealing lessons that cater to diverse learning styles and increase student comprehension. By using visual multimedia content that conveys information efficiently, educators can present material in a more engaging and effective manner.

There are drawbacks to this technology in addition to its many benefits. It is critical to ensure that the created multimedia content accurately reflects the original text while maintaining the desired message. Customization options are also necessary in order to adapt material to certain circumstances and audiences. Meeting these obstacles will be essential to ensure that the technology reaches its full potential as it develops further.

This paper explores the automation of text-to-multimedia content generation, examining its applications in journalism, marketing, and education. It highlights the advantages of this technology, such as increasing efficiency, enhanced audience engagement, and easier access to high-quality content. This research intends to analyze the major impact of automated multimedia generation on modern content creation and

communication practices by examining the integrated use of AI, NLP, and Computer Vision.

II. LITERATURE REVIEW

The field of text-to-video generation has advanced rapidly due to the incorporation of cutting-edge artificial intelligence tools. AutoGAN, a technique that combines Generative Adversarial Networks (GANs) with Neural Architecture Search (NAS) to optimize generator architectures, is one of the first attempts in this field [1]. AutoGAN's innovative approach has set new benchmarks, achieving Fréchet Inception Distance (FID) scores on datasets such as CIFAR-10 and STL-10. This represents a major advancement in generative modeling since it shows how NAS could significantly enhance the performance of GANs.

Expanding upon this basis, FineGAN presents an unsupervised GAN structure intended to produce images of finely detailed object classifications in a hierarchical manner [2]. FineGAN outperforms previous clustering approaches in disentangling the appearance, shape, and background of an item. This development highlights the model's capacity to autonomously recognize and produce fine-grained object categories, an essential skill for producing high-caliber text-to-video output. A related breakthrough that has been demonstrated to be effective on several datasets is the Masked Convolutional Generative Flow model [3]. Using masked convolutions, this model pushes the limits of AI-driven content production by enhancing the synthesis process and providing a strong substitute for conventional generative models.

There has also been substantial advancement in the field of multimodal capabilities. A noteworthy approach combines pre-trained image encoder and decoder models with frozen text-only large language models, enabling various multimodal applications, including image retrieval and unique image production [4]. This integration shows a viable path towards improving text-to-video systems, as it performs better when taking on complex language tasks. The ControlNet architecture was presented as a solution to the requirement for more precise control in text-to-image generation [5]. This architecture integrates spatially localized input conditions into a pre-trained diffusion model, thereby improving the precision and quality of generated images. The integration of diffusion probabilistic models with denoising score matching has proven helpful in improving the fidelity of created content. Other models achieve outstanding image fidelity and alignment by combining transformer language models with high-fidelity diffusion models [6]. This approach raises the standard for text-to-image generation, employing advanced language understanding to produce photorealistic graphics, an essential feature for realistic text-to-video applications.

The evolution of transformer architectures for natural language processing has been carefully investigated, with an emphasis on the transformer architecture's significance in diverse NLP applications [7]. Hugging Face's open-source library is emphasized as an essential resource for using and deploying pre-trained models, enabling more progress in AI-powered video production. Within the field of video editing, there is a new method that permits text input to be used to modify and alter videos in various ways, including adding, removing, and altering words [8]. This approach is a major step toward fully text-based editing of audiovisual content,

with possible implications for dynamic video production. It achieves realistic changes by utilizing automatic annotation and seamless stitching techniques.

A technique that creates cohesive and narrative videos from abstract stimuli [9] serves as an example of further developments in narrative-driven video production. This design, which uses optimized large language models and a diffusion model, shows how LLM-guided zero-shot video creation can produce visually engaging stories with minimal input. Using text-to-image diffusion models that have been trained beforehand, a novel method for one-shot video tuning is presented [10]. This method creates films from a single text-video pair. This approach emphasizes the adaptability of AI in managing a variety of content types by efficiently fine-tuning models to process new video input and producing remarkable text-driven video production outcomes.

A transformer-based method for latent video diffusion modeling has also been devised, making video generation more efficient and effective [11]. This model demonstrates significant capabilities in class-conditional video creation and frame prediction, highlighting the importance of transformers in text-to-video applications. A novel approach that accomplishes both local and global consistency without additional training is used to guarantee temporal consistency in text-guided video-to-video translation [12]. Using previously learned image diffusion models, this zero-shot video translation approach offers a reliable method for generating temporally coherent video sequences. Another advancement in the text-to-video domain focuses on tailoring motion in generated videos using a parallel spatial-temporal architecture [13]. This method improves the creation of realistic motion in video by distancing motion from appearance and adding reference motions to the underlying model.

A novel approach employs masked token modeling and multi-task learning to efficiently complete various video production tasks [14]. This method tokenizes videos into a low-dimensional spatial-temporal manifold by using a 3D quantization model, producing high-quality films appropriate for various applications. Finally, a technique that applies developments in text-to-image generation to word-to-video production is suggested [15]. This approach uses unsupervised video data to teach the link between words and visuals, stressing the benefits of unsupervised learning in expediting training and generating realistic motion in produced videos, and highlighting AI's potential to transform video content creation.

III. DESIGN AND METHODOLOGY

The primary objective of this research is to develop a system that generates multimedia content from text input provided by the user. The input, which can be in the various forms, undergoes several processing steps to produce the final output. The following methodology outlines these steps in detail, including the various modules and tools used.

A. Input Handling

The system begins by accepting text input from the user, which can be in various formats, including plain text, PDF, or other document types. If the input is already in text format, it is directly passed to the subsequent processing modules. However, if the input is in PDF or another non-text format, the system employs Optical Character Recognition (OCR) to

extract the textual content. The OCR process uses a pre-trained model that is specifically designed to handle diverse document layouts, fonts, and formats, ensuring accurate conversion of the content into machine-readable text. This approach standardizes all input types into a usable text format, allowing for seamless integration with the following processing steps. The robustness of the OCR model enables it to capture complex document structures and varied text presentations, making it a critical component for handling non-standardized inputs.

B. Text Processing and Summarization

The extracted text is then processed and summarized using advanced LLMs such as LLaMA2 and Gemini, which are chosen for their superior performance in natural language understanding and summarization tasks. The summarization process is achieved by prompting these models to condense the input text into concise summaries, typically ranging from 5 to 10 lines, depending on the length and complexity of the original content. These prompts are carefully designed to extract the main points and essential messages, ensuring the summary captures the critical aspects of the text. Over time, the LLMs learn the desired pattern and style of the summaries based on iterative feedback, refining their output to align closely with the specific requirements of the research.

This summarization step is essential for distilling the input text into a form that is more manageable and aligned with the subsequent stages of multimedia content creation. It involves deep linguistic analysis, ensuring the retained summary accurately reflects the core message and important details of the original article. Additionally, keywords are automatically extracted from the summarized text using natural language processing techniques. These keywords are vital as they guide the generation of visual and audio content by identifying the core elements that need to be represented. The keyword extraction process employs named entity recognition and context analysis, ensuring the selection of the most relevant and impactful terms to maintain the integrity and focus of the content.

C. Subtitle, Audio Generation and Synchronisation

The process begins by segmenting the summarized text to facilitate synchronization of visual, audio, and subtitle components. Logical breaking points are identified within the text, ensuring that phrases with a natural flow remain together, while pauses and transitions are assigned to separate segments. This segmentation provides a foundation for aligning narration and subtitles with the corresponding visuals.

For narration, the Google Text-to-Speech (GTTS) library is employed, converting text into audio with context-aware intonation to produce a more natural and engaging speech. Corresponding subtitles are generated for each text segment, ensuring they are temporally aligned with the audio.

Synchronization of multimedia content is achieved by linking each audio clip with its corresponding subtitle and visual segment. The segmented text aids this process by providing natural points for alignment. Python scripts are used to automate the synchronization, ensuring that each subtitle is displayed for the precise duration of the audio playback. The process involves the following steps:

1. *Text Segmentation*: The summarized text is divided into coherent segments based on linguistic structure, ensuring that natural pauses and continuous phrases are appropriately separated.
2. *Audio Generation*: GTTS is used to convert each text segment into an audio clip, with intonation that reflects the context of the text.
3. *Subtitle Generation*: For each audio segment, a corresponding subtitle is generated to match both content and timing.
4. *Synchronization*: Python scripts are employed to link each audio clip to its corresponding subtitle. The scripts ensure that subtitles remain visible on screen for the exact duration of the audio clip, and are synchronized with any associated visual elements.

By mapping each subtitle and audio clip to its corresponding visual element, the synchronization process ensures that all components—audio, subtitles, and visuals—are aligned seamlessly. This method guarantees a coherent and fluid multimedia experience, where each component enhances the overall narrative.

D. Visual Content Generation

Suitable images are found for each of the extracted keywords. This process involves three methods:

1. *Database Search*: A pre-existing image database is searched to find relevant images that match the extracted keywords. The database contains a diverse collection of high-quality images categorized by themes, topics, and subjects, ensuring broad coverage. This method allows for quick retrieval of images that are relevant to the content being generated.
2. *Web Scraping*: When the database does not contain suitable images, automated web scraping techniques are employed to gather images from the internet. Using Python libraries such as *BeautifulSoup* and *Selenium*, scripts are written to search for and download images from various online sources. Web scraping ensures that a wide variety of visuals are available, especially when dealing with niche or uncommon topics.
3. *Image Generation*: For abstract or highly specific keywords where suitable images cannot be found through the above methods, image generation is performed using advanced deep learning models. Diffusion models such as RunwayML, DALL-E, and Stable Diffusion are used to create high-quality images based on the provided keywords. These models employ deep learning techniques to generate images that capture the thematic elements of the keywords, ensuring relevance and consistency with the overall narrative.

The images retrieved and generated are then processed to align with the corresponding text segments. Image segmentation is applied to analyze the visual content and identify key elements within each image. Annotations are added to provide metadata that links each image to its corresponding text segment, ensuring that the visual elements are contextually accurate and enhance the narrative flow. This approach ensures that multimedia content is enriched with high-quality, thematically consistent visuals, aligned with the audio and subtitles to provide a cohesive and engaging experience.

E. Multimedia Assembly

The assembly of multimedia content is achieved using the *MoviePy* library, which integrates the generated images, audio, and subtitles into a cohesive multimedia presentation. In this process, subtitles are overlaid on images and synchronized with the audio narration. Each segment—comprising text, audio, and visuals—is carefully aligned to ensure smooth transitions and consistency. *MoviePy*'s ability to handle high-resolution content ensures that the final multimedia output is of professional quality. Transitions between segments are added programmatically to enhance the flow and coherence of the final product.

F. Final Output

Once assembled, the multimedia content is rendered in high quality using *MoviePy*, which ensures that all components—audio, visuals, and subtitles—are properly synchronized. The rendering process optimizes the file for playback on various

dataset of press articles in both PDF and text formats. The system's performance was assessed based on several criteria: accuracy of text extraction, quality of summarization, relevance of keyword extraction, synchronization of subtitles and audio, and overall quality of the generated multimedia.

The accuracy of text extraction was ensured by the OCR processing module, which effectively converted PDF documents into machine-readable text with high precision. This reliability is crucial, as errors at this stage could propagate through the system, negatively impacting the quality of the final multimedia output.

The summarization module, utilizing advanced Language Learning Models (LLMs) such as LLaMA2 and Gemini, produced concise and coherent summaries of the press articles. Human evaluators rated these summaries highly for their coherence, conciseness, and relevance, indicating that the LLMs effectively distilled the essential points of the

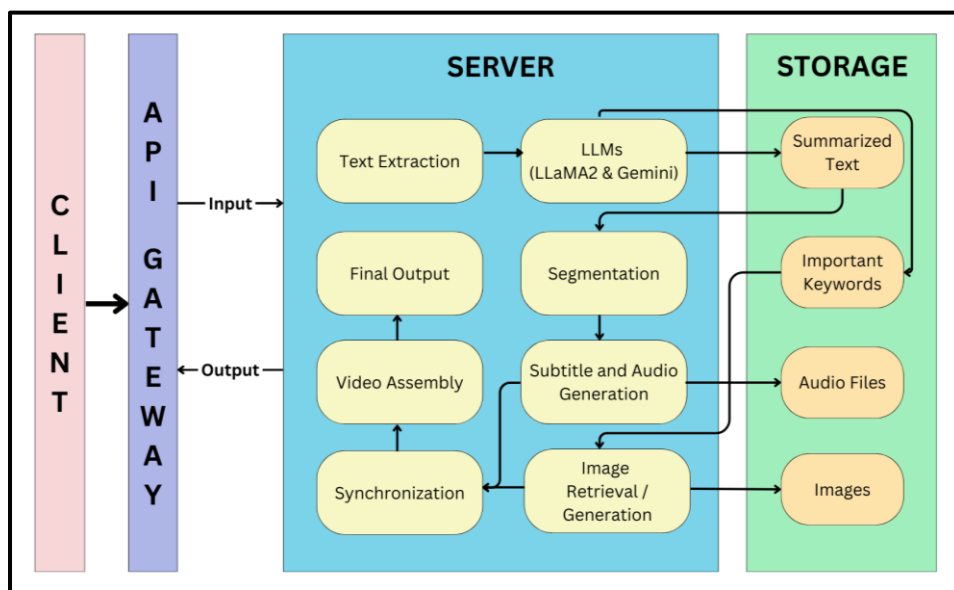


Fig. 3.1 Architecture Diagram

devices, ensuring compatibility and smooth performance across platforms. The final multimedia content is delivered through direct playback, providing an immediate and accessible viewing experience. Additionally, users are provided with the option to download the content for sharing or offline viewing.

An architecture diagram, illustrated in Fig. 3.1, depicts the system structure and the interaction between different modules, from initial input processing to final multimedia output. By integrating these components, the system effectively transforms textual input into engaging multimedia content. Each step ensures contextual accuracy and visual coherence, demonstrating the technical feasibility of automated text-to-multimedia generation. This approach has potential applications across various domains such as journalism, marketing, and education, and the system's modular design allows for scalability and adaptability, making it a versatile solution for multimedia content generation.

IV. RESULTS AND DISCUSSION

The proposed system for generating multimedia content from press articles was implemented and evaluated using a diverse

articles. Fig. 4.1 shows script generated, effectiveness of the process.

Character	Dialogue
Narrator	A sly fox, his belly rumbling with hunger, prowled through the forest searching for a meal. He searched high and low, but found nothing to satisfy his appetite.
Fox	(Grumbling) Where can a hungry fox find a decent meal in this place? I haven't eaten all
day! Narrator	Just as the fox was about to give up hope, he stumbled upon a farmer's wall. Peeking over the top, he saw the most tempting sight - a vine laden with big, juicy grapes, their skin a deep, inviting purple.
Fox	(Eyes widening) Ah, what luck! Those grapes look absolutely
delicious.	The fox licked his lips, imagining the sweet, juicy taste. He backed up, took a running leap, and jumped... but he couldn't reach the grapes. He fell back to the ground, frustrated. He tried again and again, jumping higher each time, but the grapes remained just out of reach.
Fox	(Panting) Almost... almost... Just a little
higher! Narrator	Finally, exhausted and defeated, the fox gave up. He turned away from the wall, brushing the dirt from his
coat. Fox	(Scoffs) Well, who needs those grapes anyway? They probably were sour. Yes, definitely sour. I wouldn't have liked them anyway.
Narrator	And with that, the fox trotted away, pretending he hadn't really wanted the grapes in the first place.

Fig 4.1 Accuracy of text summarization.

Keyword extraction was highly effective, accurately identifying the most important terms from the summarized text. These keywords guide the visual content generation, ensuring that the selected images align with the text. The high relevance of extracted keywords enhances the narrative by making the visual elements contextually appropriate and engaging. Visual content generation, particularly through

diffusion models, produced high-quality images relevant to the keywords. However, it was noted that generating these images takes approximately 15 minutes in local machine, which is a consideration for processing time. Fig. 4.2 and 4.3 illustrate examples of images generated using diffusion models, highlighting the capability and quality of the visual content produced.



Fig 4.2 Diffusion model-generated image for keyword "grapes"



Fig 4.3 Diffusion model-generated image for keyword "fox"

Synchronization of subtitles and narrative audio was another strong point of the system, achieving high accuracy in aligning the audio narration with the subtitles. This synchronization is essential for maintaining a seamless and coherent narrative flow, which is critical for viewer engagement. Accurate synchronization ensures that the subtitles complement the audio, making the multimedia easier to follow and more engaging.

The overall quality of the generated videos was evaluated based on visual appeal, coherence, and engagement. Human evaluators provided high ratings, reflecting the system's ability to produce high-quality, engaging videos. The integration of subtitles, narrative audio, and relevant images resulted in a cohesive and immersive viewing experience, effectively conveying the core message of the press articles in an engaging and visually appealing manner.

While this approach for AI-driven text-to-multimedia generation proves to be effective, there exist few limitations. The process of generation of summary and images demands extensive computational resources and time, particularly for producing high-quality and detailed outputs, which can be a challenge in resource-constrained environments like local machines. The quality of the generated results is highly dependent on the models used. Misinterpretations by generative models remain a concern, as they can generate outputs that do not align with context of user input. In practical use, generating the desired multimedia may require multiple iterations, as repeating the generation process tends to yield more accurate results over time. This iterative approach, again increases the need for computational power and time, creating a trade-off between efficiency and quality.

The proposed system demonstrates significant potential in transforming textual press articles into dynamic videos. While the process of generating images requires considerable time, the high accuracy in text extraction, summarization, keyword relevance, and synchronization of multimedia elements contributes to the production of high-quality videos. Future improvements could focus on optimizing the image generation process to further enhance the system's efficiency and overall user experience.

V. CONCLUSION

This research successfully demonstrates the feasibility and effectiveness of generating multimedia content from various textual inputs using a combination of OCR, advanced LLMs, keyword extraction, and image generation techniques. The system exhibited high performance in text extraction, summarization, keyword relevance, subtitle-audio synchronization, and overall content quality. High accuracy in OCR ensured a strong foundation, while LLMs produced coherent summaries, maintaining the core message of the articles. Accurate keyword extraction guided relevant visual and audio content, and precise subtitle-audio synchronization enhanced viewer engagement.

However, there are areas for improvement. Enhancing the OCR module and incorporating advanced pre-processing steps can mitigate issues from poor-quality scans, increasing text extraction accuracy. The current 15-minute media generation process can be optimized to reduce time and improve efficiency. Integrating more AI models and expanding training datasets could enhance media relevance and diversity. Additionally, advanced editing and animation techniques can make multimedia more dynamic and lifelike. Future research should focus on scaling the system for larger datasets and automating the multimedia content generation pipeline for real-time applications in newsrooms. Real-time generation capabilities and interactive features could further enhance user engagement and practical utility.

In conclusion, this research establishes a robust framework for automated multimedia content generation from press articles, demonstrating significant potential to transform journalism. Continued advancements in related technologies will refine the system, paving the way for more efficient, and versatile content creation.

ACKNOWLEDGMENT

We would like to express our deepest gratitude to Prof. Anant Kaulage and Prof. Ganesh Bhutkar from VIT Pune, India, for their invaluable guidance and support throughout the course

of this research project. Their expert insights, constructive feedback, and encouragement were instrumental in shaping the direction of our work. We are sincerely thankful for their mentorship, which played a crucial role in the successful completion of this project.

REFERENCES

- [1] Gong, X., Chang, S., Jiang, Y. and Wang, Z., 2019. Autogan: Neural architecture search for generative adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3224-3234).
- [2] Singh, K.K., Ojha, U. and Lee, Y.J., 2019. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6490-6499).
- [3] Ma, X., Kong, X., Zhang, S. and Hovy, E., 2019. Macow: Masked convolutional generative flow. *Advances in Neural Information Processing Systems*, 32.
- [4] Koh, J.Y., Fried, D. and Salakhutdinov, R.R., 2024. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36.
- [5] Zhang, L., Rao, A. and Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3836-3847).
- [6] Ho, J., Jain, A. and Abbeel, P., 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, pp.6840-6851.
- [7] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G. and Salimans, T., 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv 2022. arXiv preprint arXiv:2205.11487*.
- [8] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Davison, J., 2020, October. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).
- [9] Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D.B., Genova, K., Jin, Z., Theobalt, C. and Agrawala, M., 2019. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4), pp.1-14.
- [10] Hong, S., Seo, J., Shin, H., Hong, S. and Kim, S., 2023. DirecT2V: Large Language Models are Frame-Level Directors for Zero-Shot Text-to-Video Generation. *arXiv preprint arXiv:2305.14330*.
- [11] Zhangjie Wu, J., Ge, Y., Wang, X., Lei, W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X. and Shou, M.Z., 2022. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv e-prints*, pp.arXiv-2212..
- [12] Gupta, A., Yu, L., Sohn, K., Gu, X., Hahn, M., Fei-Fei, L., Essa, I., Jiang, L. and Lezama, J., 2023. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*.
- [13] Yang, S., Zhou, Y., Liu, Z. and Loy, C.C., 2023, December. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers* (pp. 1-11).
- [14] Zhang, Y., Tang, F., Huang, N., Huang, H., Ma, C., Dong, W. and Xu, C., 2023. Motioncrafter: One-shot motion customization of diffusion models. *arXiv preprint arXiv:2312.05288*.
- [15] Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I. and Jiang, L., 2023. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10459-10469).