Personalised Video Generation: Temporal Diffusion Synthesis With Generative Large Language Model

Akide Liu [*†] Hanwen Wang [*] Mong Yuan Sim [*]

Group 18 - COMP SCI 4816 - Applied Machine Learning Honours - 2023

The University of Adelaide, Australia

Generated Result Preview: https://vmv.re/PVG-result

Demo Preview: https://vmv.re/PVG-Demo

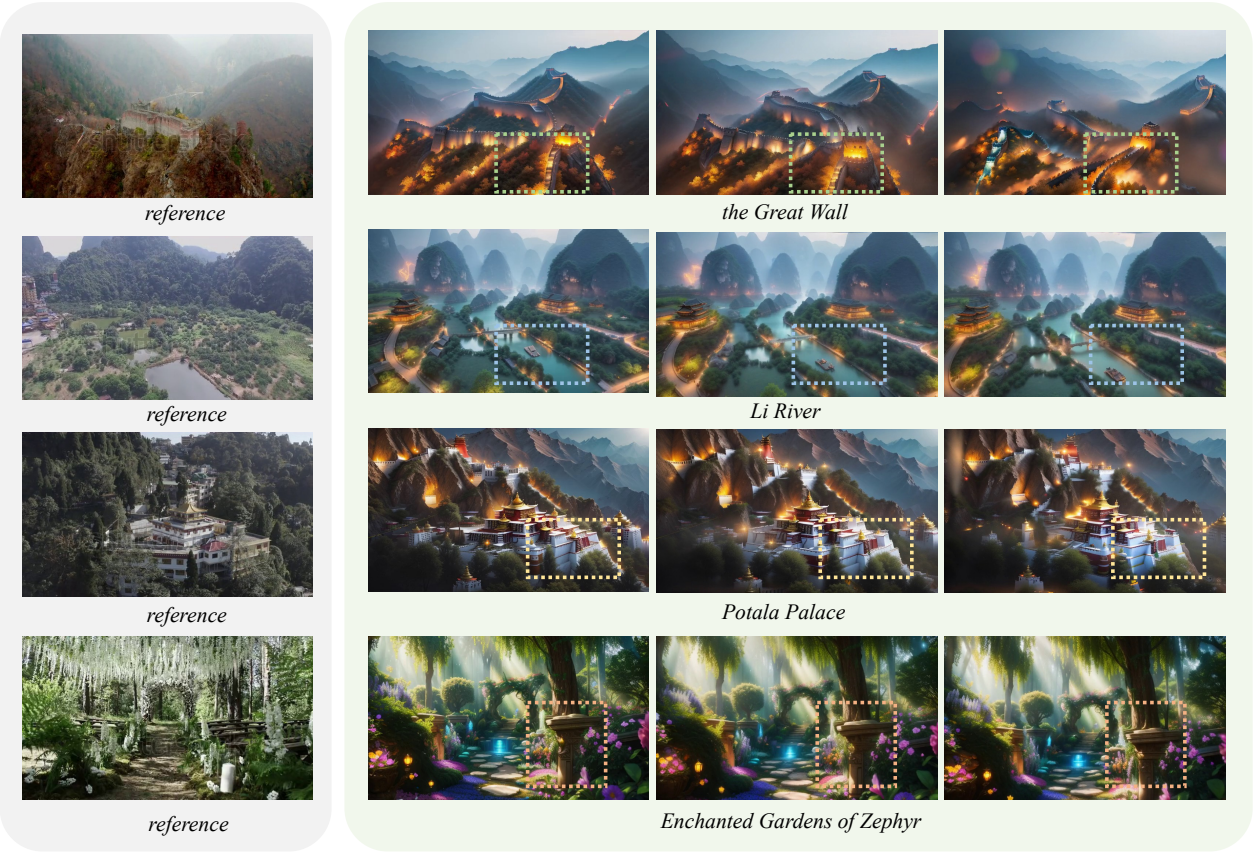

Figure 1. The image presents a trio of scenes that demonstrate the exceptional transformation capabilities of our framework. Initially, we address the challenge of temporal consistency in video generation by incorporating reference videos as input, which guides the temporal transformation process to produce coherent results. On the left, the raw input images from the original video are shown, slightly compromised by watermarks. Despite this, the framework skillfully interprets detailed textual prompts into high-resolution, watermark-free videos (768 × 448), as seen in the generated output. These prompts, which are meticulously detailed in Table 5, not only direct the removal of visual imperfections but also enhance the original content, weaving a dynamic and engaging visual narrative that underscores our framework's advanced capabilities in personalized video generation.

[*]Equal contribution

[†]Project lead, corresponding email: akide.liu@adelaide.edu.au

Abstract

Personalized video generation is a key driver of innovation in multimedia content creation. Diffusion processes have become increasingly important in this field, as they show great potential in generation settings. Our research leverages the emerging potential of diffusion processes in video synthesis to present a novel framework that fuses a large language generative model (LLM) with a text-to-video (T2V) diffusion synthesis system. The aim of this system is to produce bespoke video content from user-given text. To achieve this, we tap into the capabilities of advanced LLMs to expand brief story prompts into detailed scripts that are then harmonized with the T2V diffusion models. Our synthesis approach addresses and mitigates challenges such as frame inconsistency, high computational overheads, and the scarcity of high-quality datasets. The key components of our framework are a complex temporal attention mechanism and a robust multi-tiered pipeline, which includes key frame generation, interpolation, LDM decoding, and upsampling. These components enable us to produce seamless, richly contextual, and computationally efficient videos. To enhance the temporal consistency in video generation, we augment the text-to-video (T2V) process by integrating reference videos as guidelines. This approach facilitates the transformation of still images into fluid sequences with movement awareness, yielding more coherent and temporally consistent results. We evaluate our system using metrics such as Fréchet inception distance (FID) and Fréchet video distance (FVD), and demonstrate its capability to outperform current leading methods in crafting personalized videos. We also show how we can turn the LLM-generated prompts into an efficient and expressive text-to-video model with resolution up to 768 × 448 using LDM stable diffusion. The preliminary results indicate not only improved performance but also a novel pathway for users to create videos that vividly embody their personal stories and artistic visions.

Keywords: Large Language Generative Model (LLM), Text-to-Video (T2V), Latent Diffusion Model (LDM), Temporal Attention Mechanism(TAM)

1. Introduction

The field of computer vision is undergoing rapid advancements, with text-to-image generation [43, 44] emerging as a focal area of research. Existing methodologies [44] predominantly concentrate on generating static images based on textual prompts. While these approaches have demonstrated utility, they fall short of addressing the growing demands of video content creation. To bridge this gap, the present study embarks on an innovative endeavour. Our primary objective is to fine-tune a video diffusion model capable of generating a sequence of coherent frames. Furthermore, we aim to augment this model by integrating a large language model, which will transform textual inputs into both captions and simplified prompts. These elements will be synergistically combined within the video diffusion framework. Ultimately, the study aims to deliver a comprehensive solution for video generation, featuring seamless integration of text and visual frames.

A particularly salient field within this landscape is text-to-image generation. While a multitude of existing methodologies have made commendable strides in transforming textual descriptions into static images, they conspicuously fall short when it comes to the generation of dynamic video content [28]. The prevailing research paradigm has been largely skewed towards the creation of static images, thereby leaving a discernible void in the arena of producing seamless video narratives based on user prompts.

At the heart of this intricate challenge lies a dual conundrum. First, there's the task of generating cogent and compelling textual narratives. For this, we commence by utilizing an initial user-provided story title as a foundational prompt. Leveraging the prowess of state-of-the-art language models such as GPT-3.5 [48], GPT-4 [47], and Vicuna [83] and LLAMA [59, 61], we aspire to achieve near-human language generation capabilities. The empirical data underscores the adeptness of these models in weaving intricate narratives, and we harness this potential to flesh out the initial prompt into a comprehensive textual storyline.

The subsequent hurdle, and arguably the more intricate of the two, is translating these meticulously crafted textual narratives into dynamic visual content. The chasm between human cognition and the underlying algorithms of text-to-video diffusion models is pronounced. This disparity necessitates a nuanced approach, leading us to devise a tailored transformation technique. This method seamlessly morphs the textual narrative into a structured format that's primed for the diffusion models, encapsulating textual cues, chromatic patterns, geometric shapes, and other pivotal video generation parameters.

In essence, the overarching challenge is the seamless fusion of cutting-edge linguistic models with avant-garde video synthesis techniques, a confluence that promises to revolutionize the domain of dynamic content creation.

To actualize the envisioned seamless video synthesis, we architect a robust multi-stage pipeline. At its heart lies the Latent Diffusion Model (LDM) [51], a beacon of consistency and quality in frame rendering. Augmented with temporal attention mechanisms, the LDM ensures that the video frames progress with a rhythmic cadence, preserving both continuity and fluidity. The culmination of this pipeline is a harmonious blend of linguistic prowess and visual synthesis, promising personalized video content tailored to individual narratives. In the sections that follow, we unfurl the detailed

tapestry of our methodology, buttressed with empirical validations and benchmark comparisons. We are optimistic that our approach, a nexus of linguistic and visual AI capabilities, ushers in a significant stride forward in the realm of personalized video generation.

Our framework, informed by a symphony of pioneering research and novel methodologies, establishes itself as an innovative paradigm in the domain of video generation. Drawing inspiration from seminal works [6]. we have integrated the Latent Diffusion Model (LDM) at the heart of our architecture. This integration is pivotal in upholding the fluidity and continuity that are paramount in frame sequences. When synergized with temporal layers, the LDM stands as a beacon of consistency, ensuring that the resulting videos exhibit smooth and seamless transitions between frames. To enhance the temporal consistency in video generation, we augment the text-to-video (T2V) process by integrating reference videos as guidelines. This approach facilitates the transformation of still images into fluid sequences with movement awareness, yielding more coherent and temporally consistent results.

Central to our framework is the harmonious interplay between the LLM and T2V components. The LLM, acting as the genesis, meticulously crafts textual narratives that form the bedrock for subsequent visual interpretations. In contrast, the T2V system undertakes the task of morphing these narratives from mere textual descriptions into captivating visual tales. While each component has its own distinct realm of operation, they converge in a collaborative dance, ultimately producing a cohesive video output. This video not only resonates with visual allure but also faithfully mirrors the user's initial textual prompt, exemplifying the framework's adeptness in bridging the textual and visual realms.

Our contributions in this research are multifaceted. Our research advances the field through a novel framework that synergistically integrates advanced language models with state-of-the-art video synthesis techniques, producing personalized video content with high precision. Addressing the temporal consistency issues prevalent in diffusion architecture, we introduce a suite of technologies that refine video continuity. We have identified a gap in current research regarding long video generation and its deficiency in training data that captures the essence of extended durations. To bridge this gap, we propose utilizing reference videos from extensive video-text paired datasets to inform and enhance subframe generation. Moreover, we harness advanced large language models for central reasoning, establishing a comprehensive dataflow from tile to storyline, then to video caption, and finally to video generation prompt—fortifying the content's coherence. The forthcoming sections will elaborate on our methodology, validate it empirically, and showcase its efficacy across standard benchmark datasets.
Our principal contributions are articulated as follows:

- We unveil a cutting-edge framework that orchestrates the capabilities of large-scale generative language models with text-to-video diffusion synthesis systems, enabling the creation of tailor-made video content from textual prompts provided by users.
- Our framework's effectiveness and adaptability are proven across a wide range of domains and scenarios, showcasing its broad applicability.
- Through rigorous experimentation and evaluation, we substantiate the superior quality and variety of the video content produced by our system. Additionally, we provide evidence of heightened user satisfaction and preference for our approach.

2. Related Work

We start by reviewing the achievements of LLMs, which are pre-trained on massive text corpora and improved by reinforcement learning techniques. We show how LLMs can perform various natural language tasks, such as comprehension, generation, interaction, and reasoning, as well as in-context and instruction-based learning, and chain-of-thought prompting. We then move on to the domain of DMs, which are generative models that learn to synthesize realistic data by adding and removing noise iteratively. We highlight the recent developments in latent DMs, which compress the data into a lower-dimensional latent space and enable high-fidelity and efficient video synthesis from user descriptions. We also discuss the challenges and solutions in text-to-image generation, emphasizing the transition from GANs to models that use latent spaces for more expressive and controllable outputs. Finally, we examine the dynamic field of video diffusion, where we present the latest techniques that advance the text-to-video synthesis.

Large Language Model Large Language Models (LLMs) have gained significant attention in both academic and industrial sectors due to their exceptional performance across a wide array of Natural Language Processing (NLP) tasks [8, 11, 49, 60, 78, 80]. The architecture of these LLMs is fundamentally built upon extensive pre-training on massive text corpora, which is further enhanced by the integration of Reinforcement Learning from Human Feedback (RLHF) [49]. This synergistic approach equips LLMs with unparalleled capabilities in language comprehension, generation, interaction, and reasoning. The advent of LLMs has spurred the emergence of new research avenues that aim to exploit the inherent potential of these models. These burgeoning domains include, but are not limited to, in-context learning [8, 41, 73], instruction-based learning [12, 34, 68, 69], and chain-of-thought prompting [20, 37, 66, 71]. LLMs have been observed to adhere to scaling laws [31, 35], which contribute to their robust reasoning capabilities. These properties enable LLMs to understand natural language intricacies and

solve complex tasks with remarkable efficacy. Owing to their proven success across diverse applications [82], LLMs have experienced exponential growth in recent years. Their applications have also expanded beyond mere text generation to include interactive NLP tasks, such as embodied AI and story generation.

Diffusion Models Diffusion models (DMs) are a class of generative models that learn to synthesize realistic data by adding and removing noise iteratively. They have been widely used for video synthesis tasks, such as text-to-video generation, video inpainting, and video super-resolution. DMs are trained using denoising score matching, which minimizes the difference between the output of a denoising model and either the random noise or a specific target vector. The noise level is controlled by a schedule that depends on the diffusion time and the signal-to-noise ratio. The diffusion process can be modeled by stochastic differential equations, which are usually discretized for practical implementation.

DMs suffer from some drawbacks, such as high computational cost due to the large number of diffusion steps and low resolution due to the pixel-space representation. Moreover, DMs often require additional conditioning information, such as text prompts, to generate diverse and coherent videos. Latent diffusion models (LDMs) [51] are a novel extension of DMs that address these issues by compressing the input data into a lower-dimensional latent space using a regularized autoencoder. The latent space allows for smaller and less memory-intensive DMs that can generate high-fidelity reconstructions using fewer diffusion steps. Furthermore, LDMs can incorporate an adversarial objective to enhance the photorealism of the generated videos. In this paper, we propose a new LDM framework for personalized video synthesis based on user descriptions. We leverage state-of-the-art large language models to expand brief story prompts into detailed scripts that are then harmonized with the LDMs. We also introduce a temporal attention mechanism and a multi-stage pipeline to ensure smooth and contextually-rich video content. We evaluate our framework using various metrics and demonstrate its superiority over existing methods in terms of quality, diversity, and efficiency.

Image Diffusion Text-to-image (T2I) generation is a challenging task that aims to synthesize realistic images from natural language descriptions. In recent years, various methods have been proposed to tackle this problem, ranging from generative adversarial networks (GANs) to diffusion models. In this section, we review some of the most relevant works in this field and highlight their strengths and limitations.

GANs are a popular choice for T2I generation, as they can learn to generate sharp and diverse images by optimizing an adversarial objective. However, GANs also suffer from some issues, such as mode collapse, instability, and difficulty in capturing long-term dependencies. Early works, such as Reed et al., directly adapted GANs for T2I tasks by conditioning the generator and the discriminator on text embeddings. Later works, such as StackGAN++ [79] and AttnGAN [74], improved upon this by introducing progressive generation techniques and enhanced text-image alignment mechanisms.

A major breakthrough in T2I was achieved with DALL-E [50], which treated the generation process as a sequence-to-sequence problem, utilizing a discrete variational autoencoder (VQVAE) paired with a Transformer. This approach enabled the generation of high-quality and diverse images from complex and compositional text prompts, such as "an armchair in the shape of an avocado". Following this, various models, such as Make-A-Scene and Parti, introduced controllability into T2I generation via semantic maps and image tokenizers, respectively.

Another promising direction for T2I is to use denoising diffusion probabilistic models (DDPMs), which learn to synthesize realistic data by adding and removing noise iteratively. DDPMs have some advantages over GANs, such as stability, scalability, and flexibility. For example, GLIDE [44] utilized a T2I and an upsampling diffusion model for cascaded generation, introducing classifier-free guidance to enhance image quality and adherence to the text prompt. DALL-E combined the capabilities of CLIP's latent space with a prior model to inform the generation process. Lastly, recent advancements, such as VQ-diffusion and stable diffusion, have shifted T2I generation from pixel space to the more efficient latent space.

Video Diffusion Although text-to-image (T2I) technologies have made considerable progress, their text-to-video (T2V) counterparts still grapple with data scarcity, temporal consistency, and computational resource intensity challenges. Video Diffusion Model [28] was among the first to generate low-resolution videos through Diffusion Models (DMs) and a space-time factorized U-Net architecture. To generate high-definition videos, ImagenVideo [26] advanced the field with cascaded diffusion models and a unique v-prediction parameterization technique. Subsequent research aimed to lessen training costs by transferring knowledge from pre-trained T2I models to T2V generation. For example, Make-A-Video [56] , MagicVideo [84] , and LVDM [23] employed fine-tuning techniques, while Khachatryan et al. [36] attempted a training-free transfer. However, this method produced low-quality videos with inconsistent dynamics. Other efforts, like Gen-1 [16] and FollowYourPose [40] , focus on controlling video structure and motion dynamics using depth and pose cues, respectively. Several other studies also share our motivation to extend image Latent Diffusion Models (LDMs) to video generators. Most notably, Video-LDM [6] introduces temporal layers but retains the original weights. Our work extends these efforts by adapting a pre-trained T2I model for more efficient text-structure-guided video generation, focusing on improving video prediction mechanisms

for longer video synthesis.

In the broader area of video generation, several architectures have been employed, including Recurrent Neural Networks [1, 9, 14, 19, 38] , Autoregressive Transformers [32, 72, 75], Normalizing Flows [5, 15], and Generative Adversarial Networks (GANs) [39, 57, 63, 65, 77]. LongVideoGAN [7] generates high-resolution videos over extended durations by employing dual-resolution models. The concept of incorporating temporal layers into pre-trained models has previously been explored by MoCoGAN-HD [57] and StyleVideoGAN [18], although these were limited to object-centric videos. Unlike CogVideo [32], which relies on a strictly autoregressive architecture, our approach uses continuous Diffusion Models and offers better performance in text-to-video synthesis. More recently, Diffusion Models (DMs) have emerged as a promising technique for video synthesis. Ho et al. [29], Yang et al. [76], and Voleti et al. [64] variously used DMs for low-resolution video generation, prediction, and interpolation.

Text-to-video (T2V) generation is a challenging task that aims to synthesize realistic videos from natural language descriptions. Compared to text-to-image (T2I) generation, T2V generation has received less attention, mainly due to the lack of large-scale, high-quality text-video datasets and the high complexity of video data. In this section, we review some of the most relevant works in this field and highlight their strengths and limitations.

Early works on T2V generation focused on simple domains, such as animations of digits or constrained human actions. For example, Sync-DRAW [42] employed a variational autoencoder (VAE) with recurrent attention mechanisms to generate videos conditioned on text. Later works adapted generative adversarial networks (GANs) from image to video generation, such as TGAN and TAC-GAN, which used temporal convolutional networks and attention mechanisms, respectively.

Recent works on T2V generation have achieved significant improvements by using more advanced models and techniques. For instance, GODIVA introduced the use of a 2D VQVAE and sparse attention, enabling the generation of more realistic scenes from complex text prompts. NÜWA created a unified framework that can perform various generative tasks, including T2V, within a multitask learning setup. CogVideo augmented a pre-existing T2I model (CogView) with temporal attention modules to capture the dynamic nature of video. Video Diffusion Models (VDM) employed a space-time factorized U-Net that was trained on both image and video data.

However, most of the existing works on T2V generation rely on large private datasets with millions of text-video pairs, which limit their reproducibility and accessibility. In contrast, our work utilizes exclusively open-source datasets, such as YouCook2 and MSR-VTT, which are widely used and publicly available. Moreover, our work leverages state-of-the-art large language models and latent diffusion models to generate personalized videos based on user descriptions, which is a novel and challenging task that has not been explored before.

3. Method

Our research framework introduces a sophisticated system architecture that seamlessly integrates a Large Language Generative Model (LLM) with a Text-to-Video (T2V) Diffusion Synthesis System. The overarching aim of this integrated system is to generate personalized video content predicated on concise user-provided descriptions. Within this architecture, the LLM is specifically tasked with generating the textual components that serve as the narrative backbone of the video. Concurrently, the T2V system is responsible for rendering the video components, thereby translating the textual narrative into a visual medium.

Textual Generation. Our research introduces a sophisticated pipeline designed for the generation of personalized videos shown in Fig. 3 , a process that commences with an initial story title prompt, hereafter denoted as S^I. The function of S^I is twofold: it serves as a query parameter for scenario generation and establishes a foundational framework that guides the ensuing stages of video creation. For the critical tasks of natural language understanding and generation, we leverage state-of-the-art large language models, symbolized as G. Our implementation options include GPT-3.5 [48], GPT-4 [47], and open-source alternatives such as Vicuna [83] and LLAMA [59, 61]. As evidenced by the comparative analysis presented in Table 4, contemporary large language models exhibit capabilities in language generation [10] , knowledge utilization [3, 22] , and advanced reasoning across multiple domains [13, 24, 70] . These attributes underscore their potential utility in generating storylines that approximate human-level performance. To synthesize a nuanced and contextually rich storyline, we invoke G to produce detailed story descriptions predicated on S^I. This operation can be mathematically formalized as:

$$S^D = G(S^I). \tag{1}$$

In this equation, S^D signifies the detailed story descriptions that are algorithmically generated by G based on the initial title prompt S^I. It is imperative to recognize that the cognitive mechanisms humans employ to understand context can diverge substantially from the interpretative algorithms of text-to-video (T2V) diffusion models. To reconcile this discrepancy, we introduce a specialized processing phase. This phase is engineered to transform the human-readable story descriptions, S^D, into a format that is more conducive for T2V diffusion models. The transformation encompasses

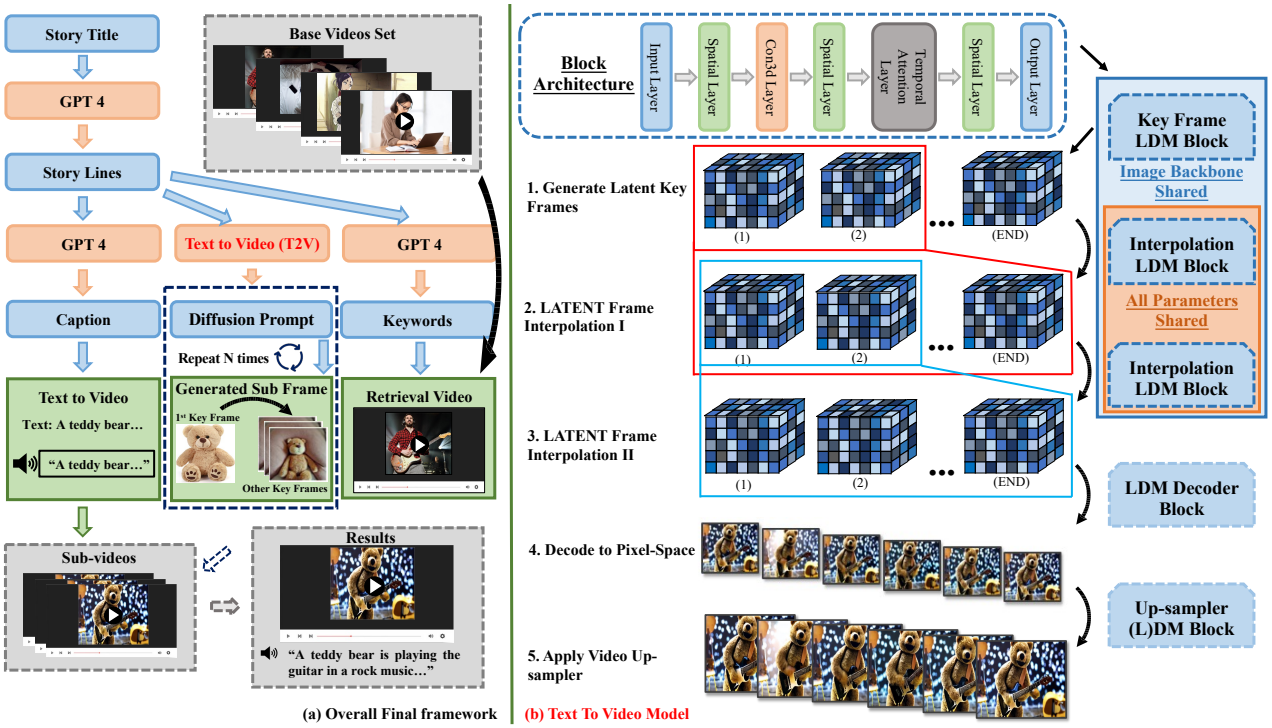

Figure 2. Personalized Video Generation Pipeline is partitioned into two primary subsystems. (a) delineates an intricate architecture that cohesively amalgamates a Large Language Generative Model (LLM) with a Text-to-Video (T2V) Diffusion Synthesis System. The overarching objective of this synergistic framework is to fabricate personalized video content, which is contingent upon succinct, user-supplied textual descriptions. Additionally, we retrieval the reference videos by the keywords, which serve as guides the temporal transformation for T2V system. Within this integrated architecture, the LLM is explicitly commissioned to synthesize the textual elements that constitute the narrative scaffolding of the resultant video. In parallel, the T2V system undertakes the task of rendering the visual constituents, thereby transmuting the textual narrative into a corresponding visual representation. (b) elucidates the architecture of our video diffusion system, which is an assemblage of five modular components: a keyframe Local Dynamic Model (LDM) block, an interpolation block, a VQ-VAE decoder block, and an upsampler block, all of which are augmented with temporally-attentive mechanisms.

the conversion of textual narratives into a structured format that may incorporate a range of attributes, including textual elements, color schemes, shapes, and categorical descriptors. We designate this refined, model-compatible description as S^P. The formal representation of this transformation is:

$$S^P = G(S^I, S^D). \tag{2}$$

Here, S^P represents the processed story description, optimized for compatibility with T2V diffusion models. It is generated by G using both the initial title prompt S^I and the detailed story descriptions S^D. By adopting this bifurcated approach, our methodology ensures the generation of videos that are not only contextually rich but also highly compatible with the computational paradigms of T2V diffusion models, thereby elevating the overall quality of the generated video content.

Tab 5 demonstrates Large Language Models' prowess in translating detailed narrative prompts into visually and emotionally captivating video scenes. These prompts, rich in metaphor and imagery, are meticulously rendered into videos where the spatial layers construct lifelike images and temporal layers enable smooth transitions, mirroring the intended narrative. Captivating depictions include the Great Wall as a symbol of resilience, the Li River's tranquil beauty, and the Potala Palace's sacredness, each reflecting the scene's unique essence and atmosphere as envisioned in the prompts.

Reference Guide Video Retrieval. To enhance the cohesion and alignment between the generated content and the retrieval system's requirements, we refine the process of keyword generation. Our approach leverages the prompts used in the diffusion model as a basis to synthesize a set of keywords S^K. These keywords are engineered to closely correspond with the expectations of the text encoder within the text-video retrieval system, ensuring that the resulting keywords serve as effective reference points for subsequent video processing steps. The generation model G operates on the set of diffusion prompts S^P, applying a transformation that yields a set of keywords optimally suited for our

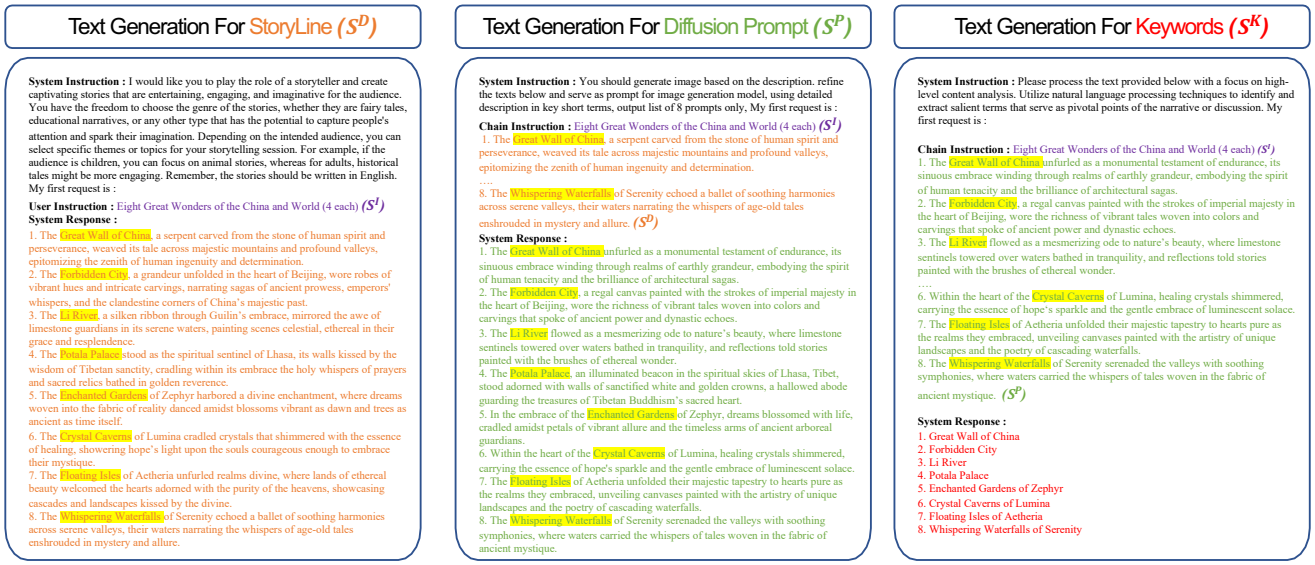

Figure 3. This figure delineates the data flow within the Large Language Model (LLM) for generative tasks. Initially, the storyline generation, denoted as S^D, is conducted by harnessing the user input S^I. This process systematically produces a list of responses corresponding to eight cities. Subsequently, the diffusion prompt S^P generation pipeline utilizes both the user input S^I and the storyline output S^D to refine the ensuing narrative further. Ultimately, this refined narrative is used to generate a targeted set of keywords S^K, which are instrumental for the ensuing text-video retrieval phase. The mathematical representation of this generative data flow is depicted as follows:

retrieval system. The formulation of this transformation is encapsulated in the following equation:

$$S^K = G(S^P), \tag{3}$$

where S^K represents the set of generated keywords and S^P denotes the set of diffusion prompts. The function G encapsulates the processes of keyword extraction, refinement, and alignment, ensuring that the output is in the desired format for the text-video retrieval system, thereby facilitating a more streamlined and coherent video generation pipeline.

The retrieval component of our system operates through a sophisticated interplay of visual and textual encodings to accurately retrieve reference videos [2]. It starts by processing visual inputs — image or video clips — by dividing them into a series of non-overlapping spatio-temporal patches. These patches are then transformed into embeddings, ready to be interpreted by a transformer that also incorporates learned temporal and spatial positional embeddings. This allows the system to understand the relative position of each patch in time and space. The embeddings pass through a series of modified space-time self-attention blocks that focus on temporal and spatial attributes sequentially, refining the video clip's representation. The [CLS] token is used as a marker for extracting the final video embedding. Parallel to this, the text encoder, a bidirectional transformer, analyzes the textual input, producing a complementary text encoding from its [CLS] token. This dual-encoding system ensures that the retrieval process is acutely aware of both the content and context of the reference material needed to guide the synthesis of new video content.

Video Generation. The endeavor of video generation is fraught with complexities, primarily due to three overarching challenges. First, the inherent frame inconsistency in generative sequences complicates the task of achieving smooth temporal progression, particularly when attempting to replicate intricate real-world scenarios. [52, 53] Second, the computational costs associated with training on video data are often prohibitive, necessitating the development of cost-effective and parameter-efficient fine-tuning strategies. [55] Third, the field currently suffers from a dearth of large-scale, high-quality, and publicly accessible datasets. [2, 67] To address these multifaceted challenges, we follow [6, 56] and introduce a comprehensive multi-stage pipeline that incorporates structural improvements, efficient fine-tuning strategies, and leverages a newly introduced text-video alignment dataset.

Addressing Temporal Consistency. To ameliorate issues related to frame-level inconsistency and to produce smoother temporal progression, we adopt a temporal attention mechanism. Our architectural design is influenced by the seminal work "Align Your Latent" [6] . Initially, we employ a pre-trained Latent Diffusion Model (LDM) parameterized by θ, which has been trained on a large-scale 2D image dataset. The original layers, which operate in the pixel space, are denoted as l_θ^i, where i represents the layer index. To introduce temporal awareness, we augment the architecture with temporal layers l_ϕ^i, which are interleaved with the existing

spatial layers l_θ^i. These temporal layers are specifically designed to align individual frames in a temporally consistent manner. The resulting architecture, denoted as $f_{\theta,\phi}$, achieves both spatial and temporal consistency. The training objective for these temporal layers, l_ϕ^i, is mathematically formulated as:

$$\arg\min_{\phi} \mathbb{E}_{\mathbf{x}\sim p_{\text{data}},\tau\sim p_\tau,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\|\mathbf{y}-\mathbf{f}_{\theta,\phi}\left(\mathbf{z}_\tau;\mathbf{c},\tau\right)\|_2^2\right] \tag{4}$$

Here, τ represents the temporal index, and $\mathbf{c}$ denotes any additional conditioning variables. The expectation is taken over the data distribution p_{data} , the temporal distribution p_τ, and a standard normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

Algorithm 1 Personalized Video Generation Framework

Input: Story title prompt S^I
Output: Personalized Video V
▷ Initialization
Initialize Latent Diffusion Model (LDM) with parameters θ
Initialize Text-to-Audio Model (TAM) with parameters ϕ
Initialize empty lists VideoList, AudioList
▷ Combined Textual and Video Generation
for $i = 1$ to N **do**
  $S^D[i] = G(S^I)$ ▷ Generate detailed story description
  $S^P[i] = G(S^I, S^D[i])$ ▷ Process for T2V model compatibility
  Generate key frames using Key Frame Generation LDM with $S^P[i]$
  ▷ Diffusion Reverse Inference
  $\mathbf{z}_s = \tilde{\boldsymbol{\mu}}_{s|t}(\mathbf{z}_t, \hat{\mathbf{x}}_\theta(\mathbf{z}_t)) + \sqrt{(\tilde{\sigma}_{s|t}^2)^{1-\gamma}(\sigma_{t|s}^2)^\gamma}\boldsymbol{\epsilon}$
  Perform frame interpolation using Interpolation LDM
  Decode latent to pixel space using LDM Decoder:

$$\mathbf{X}_{\text{decoded}}[i] = D_{\text{VQ-VAE}}(\mathbf{Z}; \theta_{\text{dec}})$$

  $A^D[i] = \text{TAM}(S^D[i]; \phi)$ ▷ Convert text to audio using TAM
  Append $\mathbf{X}_{\text{decoded}}[i]$ to VideoList
  Append $A^D[i]$ to AudioList
end for
Upscale resolution of each video in VideoList using Upsampler
Return generated video $V = \text{cat}(\text{VideoList}, S^D, A^D)$

Algorithm The algorithm 1 outlines a concise description of the Personalized Video Generation Framework.

The algorithm represents a novel procedure for converting a simple story title prompt into a detailed, dynamic, and personalized video with corresponding audio. The method involves initializing a Latent Diffusion Model (LDM) and a Text-to-Audio Model (TAM) with pre-set parameters. It proceeds by generating detailed descriptions of the story and transforming these into a format compatible with a Text-to-Video (T2V) model. This model is then used to create key frames, which are further refined through reverse inference in the diffusion process and frame interpolation. The key frames are decoded into pixels to form the video frames, while the TAM generates a synchronized audio track. These elements are compiled into a finalized video that reflects the user's personalized narrative drawn from the initial textual prompt.

Architectural Design and Multi-Stage System. At the architectural level, we adopt a modular approach that divides the overarching task of video diffusion into multiple stages. Specifically, the pipeline comprises four key components: Key Frame Generation Latent Diffusion Model (LDM), Interpolation LDM, LDM Decoder, and an Upsampler.

Key Frame Generation LDM. The Key Frame Generation LDM serves as the foundational model for generating 3D latent matrices imbued with rudimentary temporal awareness. Given the potential for large semantic changes between frames, stabilization is crucial. To this end, we employ a classifier-free diffusion guidance mechanism [27] to guide the model during the sampling process. Mathematically, this guidance is formulated as:

$$\mathbf{f}'_{\theta,\phi}\left(\mathbf{z}_\tau;\mathbf{c}_S\right) = \mathbf{f}_{\theta,\phi}\left(\mathbf{z}_\tau\right) + s\cdot\left(\mathbf{f}_{\theta,\phi}\left(\mathbf{z}_\tau;\mathbf{c}_S\right) - \mathbf{f}_{\theta,\phi}\left(\mathbf{z}_\tau\right)\right) \tag{5}$$

Interpolation LDM. The Interpolation LDM aims to enhance frame consistency between key frames. This component employs a marking-conditioning method and applies masking techniques to the target frames. Given a key frame denoted as T, the interpolation system is designed to operate in two regions: $T \rightarrow 4T$ and $4T \rightarrow 16T$, thereby achieving a higher frame rate.

LDM Decoder. The Latent Diffusion Model (LDM) Decoder serves as a pivotal component in our multi-stage video generation pipeline. This decoding process is non-trivial, as it necessitates the preservation of intricate spatial and temporal features encapsulated within the latent matrices. Mathematically, the decoding operation can be represented as follows:

$$\mathbf{X}_{\text{decoded}} = D_{\text{VQ-VAE}}(\mathbf{Z}; \theta_{\text{dec}}) \tag{6}$$

where $\mathbf{X}_{\text{decoded}}$ denotes the decoded video frames in pixel space, $\mathbf{Z}$ represents the latent matrices, $D_{\text{VQ-VAE}}$ [46, 51] is the decoding function parameterized by θ_{dec}, which are the trainable parameters of the decoder.

Upsampler. The Upsampler serves as a super-resolution system, designed to upscale the video resolution from 512×1024 to 1280×2048 through a cascaded Diffusion Model (DM). The training objective for this super-resolution

system is defined as:

$$\mathbb{E}_{\mathbf{x}\sim p_{\text{data}},(\tau,\tau_\gamma)\sim p_\tau,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\left\|\mathbf{y}-\mathbf{g}_{\theta,\phi}\left(\mathbf{x}_\tau;\mathbf{c}_{\tau_\gamma},\tau_\gamma,\tau\right)\right\|_2^2\right] \tag{7}$$

By employing this multi-stage architecture, our system is designed to address the complexities inherent in video generation tasks, thereby achieving both high-quality and computationally efficient results.

The novelty of our approach resides in the introduction of FVD, a metric specifically designed for evaluating video generation models. FVD builds upon the foundational principles of FID and represents a significant advancement in the evaluation of generative models for video synthesis. Our empirical studies substantiate that FVD provides accurate assessments of videos augmented with static 3D objects.

While human evaluation remains a vital component for assessing generative models, it is inherently subjective and can vary across individuals. Hence, it is imperative to complement human evaluations with objective metrics like FID and FVD to obtain a holistic understanding of a generative model's performance.

4. Experiments

In this section, we present the experimental results of our personalized video synthesis framework based on video latent diffusion models (Video LDMs). We compare our framework with the state-of-the-art methods on various datasets and metrics, and conduct human evaluation and ablation studies to demonstrate the effectiveness and robustness of our approach. We conduct all experiments on a cluster with 4-8 GPU nodes each equipped 4 NVIDIA A100-40G GPUs.

4.1. Dataset

InternVid The InternVid [81] dataset is a large-scale dataset for multimodal video understanding and generation, which was collected from the web using large language models (LLM). It contains 234 million video-description pairs that cover a diverse range of topics, such as education, entertainment, health, and science. The total duration of the videos is over 760,000 hours, making it one of the most comprehensive video datasets available.

WebVid-10M Dataset The WebVid-10M [2] dataset is a large-scale dataset for video generation and understanding, which was collected from the web using natural language queries. It contains 10.7 million video-caption pairs that cover a wide range of topics, such as animals, sports, music, and art. The total duration of the videos is over 52,000 hours, making it one of the largest video datasets available.

The WebVid-10M dataset is used to adapt the "Stable Diffusion" Image LDM (Latent Diffusion Model), a model initially designed for images, into one that can handle video data (Video LDM). The LDM is a generative model that learns to synthesize realistic data by adding and removing noise iteratively. The Video LDM extends the Image LDM by incorporating temporal attention and multi-stage pipelines to generate smooth and contextually-rich videos.

The videos from the WebVid-10M dataset have been resized to a resolution of 320×512 for the purpose of the study. The captions have been tokenized and encoded using a pre-trained tokenizer. The dataset is split into training, validation, and test sets, with 80%, 10%, and 10% of the data, respectively.

4.2. Evaluation metrics

Frechet Inception Distance (FID). The Fréchet Inception Distance (FID) serves as a robust metric for assessing the quality of images synthesized by Generative Adversarial Networks (GANs) and DDPM diffusion models [25]. This metric quantifies the dissimilarity between the feature vectors of real and generated images. Specifically, it computes the statistical properties of these feature vectors, which are extracted using the Inception v3 model originally designed for image classification tasks.

Fréchet Video Distance (FVD). An extension of FID, the Fréchet Video Distance (FVD), is employed to evaluate the quality of videos generated by generative models [62]. Similar to FID, FVD measures the distance between feature vectors of real and generated videos. These feature vectors encapsulate the statistics of computer vision features extracted using the Inception v3 model. A lower FVD score indicates that the real and generated videos are statistically similar, with a perfect score of 0.0 implying identical distributions. FVD has been empirically validated to align well with human qualitative judgments.

The mathematical expressions for both FID and FVD are given by:

$$\begin{aligned}\text{FID} \cong \text{FVD} = &\|\mu_{\text{real}} - \mu_{\text{generated}}\|_2^2 \\ &+ \text{Tr}\left(\Sigma_{\text{real}} + \Sigma_{\text{generated}} - 2\sqrt{\Sigma_{\text{real}}\Sigma_{\text{generated}}}\right),\end{aligned} \tag{8}$$

where μ_{real} and $\mu_{\text{generated}}$ are the means of the feature vectors for the real and generated data, respectively, and Σ_{real} and $\Sigma_{\text{generated}}$ are their corresponding covariance matrices.

Comparative Analysis. FID has been rigorously compared against other prevalent metrics such as the Inception Score (IS) [54] and Kernel Inception Distance (KID) [4]. Similarly, FVD has been benchmarked against the Structural Similarity Index (SSIM) [45] and Peak Signal-to-Noise Ratio (PSNR) [17]. Empirical evidence suggests that both FID and FVD outperform these alternative metrics in evaluating the quality of generative models.

Quantitative results We train our Video LDM framework on the WebVid-10M Dataset, which contains 10.7 million video-caption pairs from various domains. We condition on tags and the level of crowding, and randomly drop these labels during training to achieve classifier-free guidance and

Table 1. Comparison with LVG on RDS.

Method	FVD	FID
LVG [7]	478	53.5
Align Latents [6]	389	31.6
Align Latents [6] (cond.)	356	51.9
Ours	298	**27.9**
Ours (cond.)	**274**	48.3

Table 2. Ablations for context guided.

Method	FVD	FID
Pixel-baseline	639,56	59.70
End-to-end LDM	1155.10	71.26
Attention-only	704.41	50.01
Align Latents [6]	534.17	48.26
Align Latents [6] (context-guided)	508.82	54.16
Ours	435.23	**41.33**
Ours (context-guided)	**391.72**	44.81

unconditional synthesis. We do not condition on bounding boxes in this setting. We adopt the first training the image backbone LDM on video frames independently (spatial layers) and then training temporal layers on the video.

Table 1 shows our main results for the Video LDM without upsampler. We report the performance of our model with and without conditioning on tags and crowding. We use the Fréchet Video Distance (FVD) and the Frechet Inception Distance (FID) [25] as the quantitative metrics to measure the quality and diversity of the generated videos. Lower FVD and FID indicate better performance. As can be seen, our Video LDM generally outperforms LVG on both metrics, and adding conditioning further reduces FVD, indicating that our model can generate more realistic and diverse videos that match the given conditions.

We also train a Long Video Generation (LVG) adversarial network [7] on the same dataset as our main baseline, which is the previous state-of-the-art method for long-term high-resolution video synthesis. We use the same resolution and frame rate as LVG (128×256 30 fps) for fair comparison. Next, we compare our video fine-tuned pixel-space upsampler with independent frame-wise image upsampling (Table 2), using 128×256 30 fps ground truth videos for conditioning. We find that temporal alignment of the upsampler is crucial for high performance. FVD degrades significantly, if the video frames are upsampled independently, indicating loss of temporal consistency. As expected, FID is essentially unaffected, because the individual frames are still of high quality when upsampled independently.

InternVid is a popular benchmark used to evaluate video generation and has recently been employed in Text-to-Video (T2V) models. Our framework underwent finetuning of its pretrained model for class-conditional video generation. In contrast, VDM (Ho et al., 2022) performed unconditional video generation and was trained from scratch on InternVid. We contend that both approaches are suboptimal and do not directly assess the T2V generation capabilities. Furthermore, the FVD evaluation model requires the videos to be 0.5 seconds (16 frames) long, which is impractical for real-world video generation applications. Despite this, for comparison with prior work, we evaluated our framework on InternVid in both zero-shot and finetuning scenarios. As indicated in Table 3, our framework's zero-shot performance is already more competitive than other models trained on InternVid and significantly outperforms CogVideo, demonstrating superior generalization to such a specific domain. Our finetuning setting achieves state-of-the-art results with a substantial reduction in FVD, implying that our framework can produce more coherent videos than previous methods.

Qualitative results

The key idea of the provided content is that the temproal diffusion synthesis framework offers advanced capabilities in video generation and editing tasks when compared to existing technologies like CogVideo and VDM for video generation, and FILM for video interpolation. Our framework can produce videos with better motion consistency and relevance to the input text. It can also perform image animation, creating personalized videos from a single image, and can generate semantically meaningful transitions between two images, outperforming FILM in understanding the semantic content of movements within video frames. Additionally, It can create videos that are semantically similar to a source video by using averaged CLIP embeddings of the original video frames. The document highlights that due to space limitations, only three examples of each capability are shown. In Figure 1, we show conditional samples from the combined Video LDM and video upsampler model. We observe high-quality videos that match the given conditions. Moreover, using our prediction approach, we find that we can generate very long, temporally coherent high-resolution videos of multiple minutes.

5. Limitations

While we have proposed a highly flexible system, there are certain limitations to be noted. One such limitation pertains to the upgrading mechanism, which is currently reliant on human feedback. This dependence may affect the user experience. To address this, one potential solution is the integration of a sentiment analyzer that gauges the sentiment of the responses generated by the large language models (LLMs). There are instances when LLMs may fail to generate an appropriate response due to a mismatch between the local document and the user query. In such cases, LLMs might return a response like "The text provided is not related to the query". By employing sentiment analysis, the system can discern this as a negative sentiment from the

Table 3. Video generation evaluation on InternVid for both zero-shot and fine-tuning settings.

Method	Pretrain	Class	Resolution	IS (↑)	FVD (↓)
Zero-Shot Setting					
CogVideo (Chinese)	No	Yes	480×480	23.55	751.34
CogVideo (English)	No	Yes	480×480	25.27	701.59
Make-A-Video [56]	No	Yes	256×256	33.00	367.23
Align Latents [6]	No	Yes	480×480	36.12	324.14
Ours	No	Yes	480×480	**41.32**	**296.12**
Finetuning Setting					
TGANv2	No	No	128×128	26.60 ± 0.47	-
DIGAN [77]	No	No		32.70 ± 0.35	577 ± 22
MoCoGAN-HD [58]	No	No	256×256	33.95 ± 0.25	700 ± 24
CogVideo [33]	Yes	Yes	160×160	50.46	626
VDM [30]	No	No	64×64	57.80 ± 1.3	-
TATS-base [21]	No	Yes	128×128	79.28 ± 0.38	278 ± 11
Make-A-Video [56]	Yes	Yes	256×256	82.55	81.25
Align Latents [6]	Yes	Yes	256×256	73.17	113.47
Ours	Yes	Yes	256×256	$\mathbf{88.14 \pm 0.91}$	**74.62**

LLM and consequently escalate the query to a higher level of assistance automatically. Another limitation arises due to the restrictive context window sizes of the LLMs, which are typically 4k or 8k tokens. This constraint can result in the system's failure to adequately address complex queries that necessitate a broader understanding of the context or the synthesis of information from multiple documents. One way to mitigate this limitation is by increasing the context window size. To this end, ALiBi, a linear-biased attention mechanism, could be integrated into the system to allow for an adjustable maximum token length at the interface level. This proactive adaptation, through sentiment analysis and the incorporation of mechanisms like ALiBi, can potentially lead to a more fluid and effective interaction, enhancing both the system's capabilities and the user experience. Furthermore, these adaptations emphasize the importance of the system's ability to recognize its limitations and make automatic adjustments in real-time to meet the demands of complex queries.

6. Conclusion

In this study, we have developed a sophisticated system that elevates the creation of personalized videos to an unprecedented level of specificity and relevance, by leveraging the advanced cognitive capabilities of large language models (LLMs) and the nuanced temporal understanding of diffusion models. The system intelligently interprets user inputs, transforming them into expanded narratives that inform the video synthesis process. This is augmented by a state-of-the-art text-to-video retrieval transformer that infuses the generated content with coherent and contextually appropriate motion. The cornerstone of our approach is a pioneering temporal attention mechanism, which, in tandem with our robust multi-stage generative model, produces results that are not only technically superior but also deeply resonant with the user's original vision. This represents a quantum leap in the personalized video generation space, opening up new avenues for user-centric video content creation and marking a notable contribution to the field of visual intelligence systems.

7. Reflection

In this reflective section, we articulate our collective and individual learnings derived from our interdisciplinary group project, situated at the intersection of Natural Language Processing (NLP) and Computer Vision (CV).

The endeavor has culminated in the successful attainment of our initial objectives and has significantly broadened our technical acumen and collaborative capabilities.

Our foray into integrating NLP with CV necessitated a synergistic expansion of our expertise to engineer a sophisticated text-to-video generation system. This project has catalyzed the potential for future joint ventures amongst team members, given our newfound cross-disciplinary proficiencies.

Furthermore, this large-scale project underscored the criticality of clear communication across diverse academic backgrounds, fostering an environment where idea exchange, constructive feedback, and the simplification of complex technical concepts were paramount.

Additionally, the project reinforced the value of teamwork and meticulous time management. Our team honed the ability to delegate tasks effectively, harness individual strengths, and adhere to stringent timelines, ultimately achieving a balance between academic responsibilities and research imperatives. These skills proved instrumental in meeting deadlines and ensuring the timely fruition of our project.

7.1. Akide Liu

In this project, we set out to create a sophisticated system that marries the complexities of vision and language, aiming to produce personalized videos tailored to user specifications. This multi-modal system bridges Natural Language Processing (NLP) and Computer Vision (CV), areas pivotal to my current research trajectory in large generative vision-language models. As the project lead, my role was integral in navigating the project from conceptualization to execution. We have conceptualized, developed, and evaluated a state-of-the-art text-to-video generation system, which has significantly deepened my comprehension of diffusion models and fortified my foundational knowledge in NLP.

My contributions to the project were extensive and lead the project development. I spearheaded the initial design, meticulously crafting the system architecture to ensure seamless integration of NLP and CV components. I actively engaged in developing the algorithm that underpins our text-to-video synthesis, meticulously fine-tuning the diffusion model parameters to ensure high fidelity in video generation. My technical acumen was also instrumental in enhancing the system's language understanding capabilities, contributing to the refinement of the LLM components.

The collaboration within our team was the epitome of interdisciplinary synergy. Regular strategy sessions were held, fostering a fertile ground for innovative ideas and allowing for robust peer review of the evolving system. Each member's specialized knowledge was harnessed to address specific challenges, whether it was improving the system's natural language comprehension or enhancing the visual quality of generated content.

In summary, this project was not only a testament to the power of interdisciplinary collaboration but also a personal journey of professional growth, enhancing both my leadership abilities and my technical proficiencies in cutting-edge AI technologies.

7.2. Hanwen Wang

My previous research experience was primarily focused on Virtual Reality, which did not align with the field of this project, Computer Vision. However, during the initial stages of the project, I gained an understanding that Computer Vision and Virtual Reality share intersections. This not only significantly piqued my interest in the project but also expanded my awareness of the Computer Vision field.

In this project, my primary responsibility was related to creating visual materials. This task required not only a deep understanding of the project's methods and setup but also a strong design skill to present model structures effectively. Therefore, I diligently studied the structure of the LDM model and the principles behind keyframe generation. Additionally, during the coding work, I utilized pre-existing diffusion models and fine-tuned their parameters, which enhanced my proficiency in using deep learning tools.

Regarding teamwork, our group conducted meetings during the project's preparation phase to identify a suitable project topic and outline a project plan. Following this, I rigorously adhered to the project plan, completed my individual tasks, and actively prepared for presentation work. Therefore, I believe my collaboration skills were effectively improved.

In summary, this project experience has been both novel and meaningful for me. It not only increased my expertise in the Computer Vision field but also improved my personal proficiency in utilizing deep learning tools and creating visual materials.

7.3. Mong Yuan Sim

While I primarily focused on Natural Language Processing (NLP), my expertise in computer vision was limited. This project has significantly improved my comprehension of the advancements in computer vision, specifically in video generation. I have delved into a new domain and broadened my technical skills to grasp and implement models in computer vision. I believe this experience is mutually beneficial, enhancing my understanding in both computer vision and NLP.

By bridging the gap between these two domains, I can create opportunities for interdisciplinary collaboration and innovative solutions. Furthermore, this project has notably allowed me to improve my coding and debugging skills, particularly when dealing with large models and datasets. I became more adept at crafting efficient, optimized code and have learned to diagnose and troubleshoot complex issues more effectively.

Working on this large-scale project has underscored the importance of collaboration and effective communication among team members to achieve our objectives. Transitioning from NLP to computer vision requires rapid learning, compelling me to acquire fundamental knowledge in computer vision swiftly and develop the necessary technical skills.

In summary, this project has not only expanded my technical knowledge in computer vision but has also enriched my soft skills, encompassing teamwork, time management, problem-solving, and effective communication. These skills are transferable and can benefit me in my future career within any areas.

8. Acknowledgement

This paper serves as a technical report for the final project conducted within the scope of the University of Adelaide Course, COMP SCI 4816 - AppliedMachine Learning Honours. We would like to express our gratitude to A/Prof. Qi Wu, Dr. Bernard Evans, Xueqian Li, course teaching team, for providing guidance and support throughout the duration of this course. Additionally, we extend our appreciation to my friend, Jinchao Ge, for valuable discussions and insights that contributed to the development of this work.

References

[1] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018. 5

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 7, 9

[3] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544, 2013. 5

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 9

[5] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. ipoke: Poking a still image for controlled stochastic video synthesis. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021. 5

[6] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, SeungWook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. Apr 2023. 3, 4, 7, 10, 11

[7] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A Efros, and Tero Karras. Generating long videos of dynamic scenes. *arXiv:2206.03429*, 2022. 5, 10

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[9] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 5

[10] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, HenriquePondedeOliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, FelipePetroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, ElizabethA. Barnes, Ariel Herbert-Voss, WilliamH. Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, AndrewN. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, MatthewM. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Samuel McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *Cornell University - arXiv,Cornell University - arXiv*, Jul 2021. 5

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022. 3

[12] Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022. 3

[13] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *Cornell University - arXiv,Cornell University - arXiv*, Oct 2021. 5

[14] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 2018. 5

[15] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G. Derpanis, and Björn Ommer. Stochastic image-to-video synthesis using cinns. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2021. 5

[16] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, and Jonathan Granskog. Structure and content-guided video synthesis with diffusion models. 4

[17] Fernando A Fardo, Victor H Conforto, Francisco C de Oliveira, and Paulo S Rodrigues. A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms. *arXiv preprint arXiv:1605.07116*, 2016. 9

[18] Gereon Fox, Ayush Tewari, Mohamed Elgharib, and Christian Theobalt. Stylevideogan: A temporal generative model using a pretrained stylegan. In *British Machine Vision Conference (BMVC)*, 2021. 5

[19] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 5

[20] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *ArXiv*, abs/2211.10435, 2022. 3

[21] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. *ECCV*, 2022. 11

[22] Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Jul 2019. 5

[23] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. Nov 2022. 4

[24] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Cornell University - arXiv,Cornell University - arXiv*, Sep 2020. 5

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 9, 10

[26] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, DiederikP Kingma, Ben Poole, Mohammad Norouzi, DavidJ Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. 4

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 8

[28] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and DavidJ. Fleet. Video diffusion models. Apr 2022. 2, 4

[29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 5

[30] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 11

[31] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego De, Las Casas, Lisa Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, GeorgeVanDen Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. 3

[32] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv:2205.15868*, 2022. 5

[33] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers, 2022. 11

[34] Srinivas Iyer, Xiaojuan Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Veselin Stoyanov. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *ArXiv*, abs/2212.12017, 2022. 3

[35] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 3

[36] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. 4

[37] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916, 2022. 3

[38] Alex X. Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018. 5

[39] Pauline Luc, Aidan Clark, Sander Dieleman, Diego de Las Casas, Yotam Doron, Albin Cassirer, and Karen Simonyan. Transformation-based adversarial video prediction on large-scale data. *ArXiv*, 2020. 5

[40] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and ‡Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. 4

[41] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Conference on Empirical Methods in Natural Language Processing*, 2022. 3

[42] Gaurav Mittal, Tanya Marwah, and Vineeth N Balasubramanian. Sync-draw: Automatic video generation using deep recurrent attentive architectures. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1096–1104, 2017. 5

[43] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171, 2021. 2

[44] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 4

[45] Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv preprint arXiv:2006.13846*, 2020. 9

[46] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 8

[47] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 2, 5

[48] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. 2, 5

[49] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022. 3

[50] Aditya Ramesh, Gabriel Goh, Mark Chen, Ilya Sutskever, and Anirudh Agarwal. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 4

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4, 8

[52] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 7

[53] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 7

[54] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. 9

[55] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021. 7

[56] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 4, 7, 11

[57] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *International Conference on Learning Representations*, 2021. 5

[58] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. *ICLR*, 2021. 11

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 5

[60] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023. 3

[61] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2, 5

[62] Thomas Unterthiner, Sjoerdvan Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *Cornell University - arXiv,Cornell University - arXiv*, Dec 2018. 9

[63] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. *ICLR*, 2017. 5

[64] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. *arXiv preprint arXiv:2205.09853*, 2022. 5

[65] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016. 5

[66] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171, 2022. 3

[67] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 7

[68] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, M. Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri,

Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddharth Deepak Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hanna Hajishirzi, and Daniel Khashabi. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Conference on Empirical Methods in Natural Language Processing*, 2022. 3

[69] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652, 2021. 3

[70] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. 5

[71] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022. 3

[72] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *International Conference on Learning Representations*, 2020. 5

[73] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *ArXiv*, abs/2111.02080, 2021. 3

[74] Tao Xu, Han Zhang, Xiaolei Zhang, Xiaogang Huang, Shaoting Zhang, and Dimitris Metaxas. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018. 4

[75] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers, 2021. 5

[76] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022. 5

[77] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022. 5, 11

[78] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, P. Zhang, Yuxiao Dong, and Jie Tang. Glm-130b: An open bilingual pre-trained model. *ArXiv*, abs/2210.02414, 2022. 3

[79] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5908–5916, 2017. 4

[80] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068, 2022. 3

[81] Zhenhai Zhang, Yuxin Wang, Jia Li, Han Zhang, Tao Xu, Xiaolei Huang, and Dimitris Metaxas. Internvid: A large-scale video-centric multimodal dataset for multimodal understanding and generation. *arXiv preprint arXiv:2112.10741*, 2021. 9

[82] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2023. 4

[83] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 2, 5

[84] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 4

Appendix

Our appendices are organized into two distinct sections. The first section provides a comprehensive algorithmic description of our framework, elucidating its formal underpinnings. The second section focuses on the evaluative assessment of Large Language Generative Models (LLMs), demonstrating their robust capabilities in both generating and refining structures.

Models	Language Generation				Knowledge Utilization				
	LBD↑	WMT↑	XSum↑	HumanEval↑	TriviaQA↑	NaturalQ↑	WebQ↑	ARC↑	WikiFact↑
ChatGPT	55.81	36.44	21.71	79.88	54.54	21.52	17.77	93.69	29.25
Vicuna (7B)	60.12	18.06	13.59	17.07	28.58	9.17	6.64	16.96	26.95
LLaMA (7B)	66.78	13.84	8.77	15.24	34.62	7.92	11.12	4.88	19.78

Models	Knowledge Reasoning			Symbolic Reasoning		Mathematical Reasoning		Interaction with Environment	
	OBQA↑	HellaSwag↑	SocialIQA↑	C-Objects↑	Penguins↑	GSM8k↑	MATH↑	ALFW↑	WebShop↑
ChatGPT	81.20	61.43	73.23	53.20	40.27	78.47	33.78	58.96	45.12/15.60
Vicuna (7B)	30.00	26.26	36.39	44.25	36.24	14.03	3.54	1.49	6.90/1.40
LLaMA (7B)	27.00	25.57	33.11	39.95	34.90	10.99	3.12	2.24	0.00/0.00

Models	Human Alignment					Tool Manipulation			
	TfQA↑	C-Pairs↑	WinoGender↑	RTP↓	HaluEval↑	HotpotQA↑	Gorilla-TH↑	Gorilla-TF↑	Gorilla-HF↑
ChatGPT	69.16	81.40	62.50/72.50/79.17	3.07	66.64	23.80	67.20	44.53	19.36
Vicuna (7B)	57.77	67.24	49.17/49.17/49.17	4.70	43.44	6.20	0.00	0.00	0.33
LLaMA (7B)	47.86	68.50	54.17/52.50/51.67	5.94	14.18	1.60	0.00	0.00	0.11

Table 4. Evaluation on the eight abilities of LLMs with specially selected tasks. The shade of the Orange and Blue fonts denote the performance orders of the results in closed-source and open-source models, respectively. This table will be continuously updated by incorporating the results of more models.

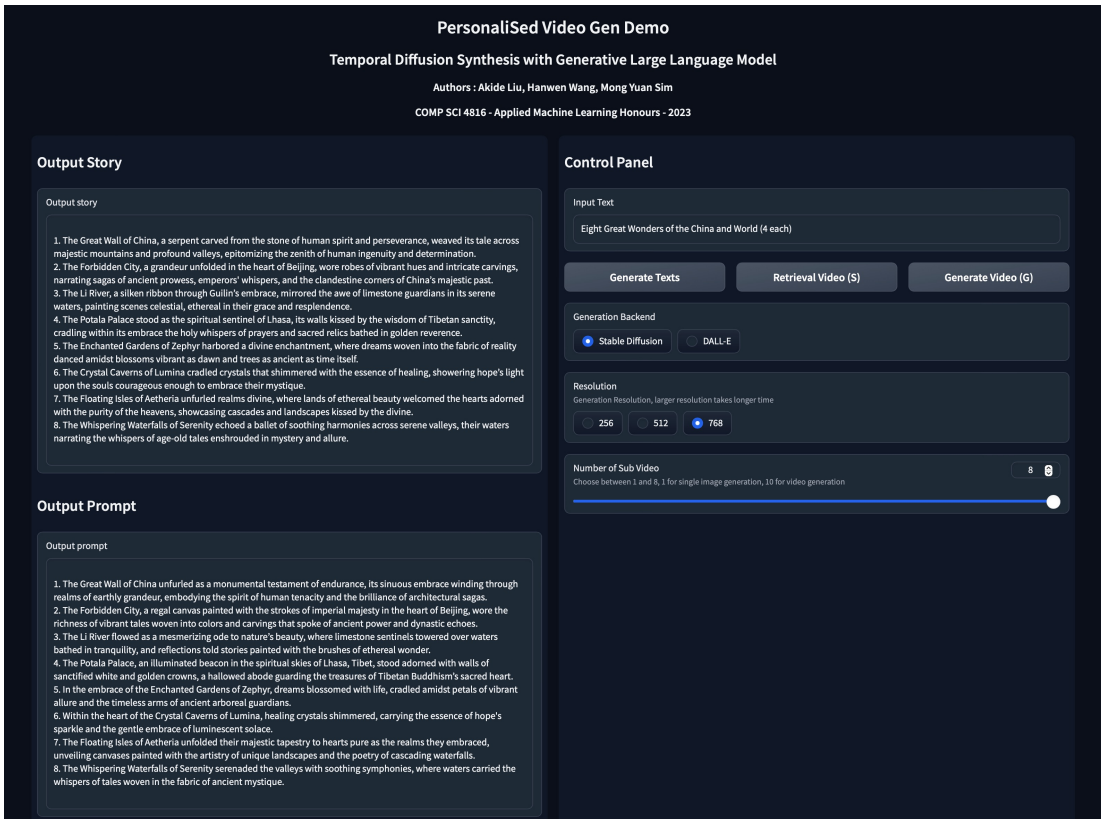

Figure 4. First Part of the UI.

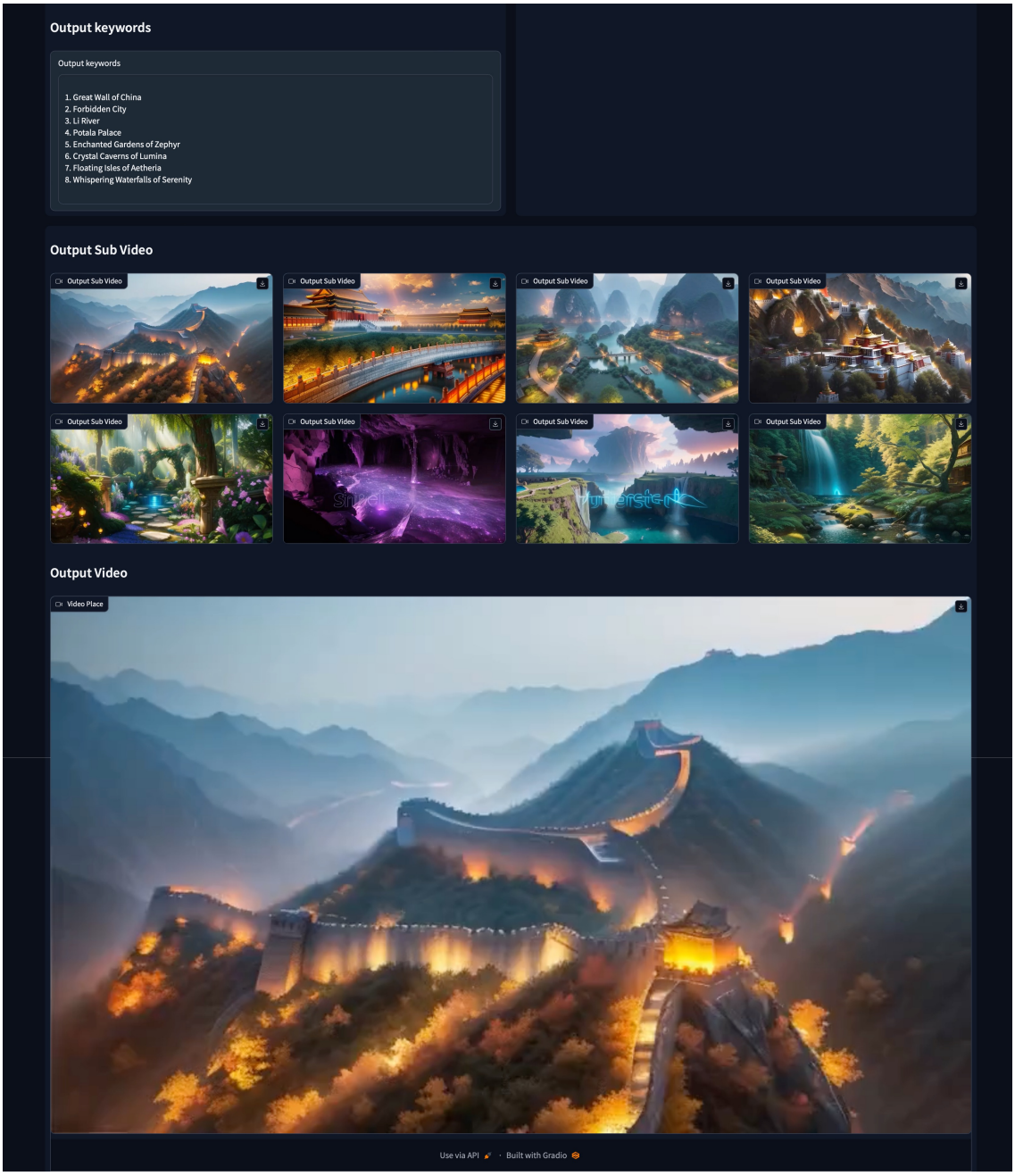

Figure 5. Second Part of the UI.

Table 5. Utilizing the advanced capabilities of Large Language Models (LLMs), this table shows the results of natural language prompts that intricately describe each scene for the video. These prompts are intricately transformed by the spatial layers of the video generator into realistic images, while the temporal layers ensure seamless transitions, culminating in videos that align with user-defined preferences and specifications, as exemplified in Fig. 1.

Content	Prompt
the Great Wall	The Great Wall, a serpent carved from the stone of human spirit and perseverance, weaved its tale across majestic mountains and profound valleys, epitomizing the zenith of human ingenuity and determination.
Li River	a silken ribbon through Guilin's embrace, mirrored the awe of limestone guardians in its serene waters, painting scenes celestial, ethereal in their grace and resplendence.
Potala Palace	The Potala Palace stood as the spiritual sentinel of Lhasa, its walls kissed by the wisdom of Tibetan sanctity, cradling within its embrace the holy whispers of prayers and sacred relics bathed in golden reverence.
Enchanted Gardens of Zephyr	The Enchanted Gardens of Zephyr harbored a divine enchantment, where dreams woven into the fabric of reality danced amidst blossoms vibrant as dawn and trees as ancient as time itself.

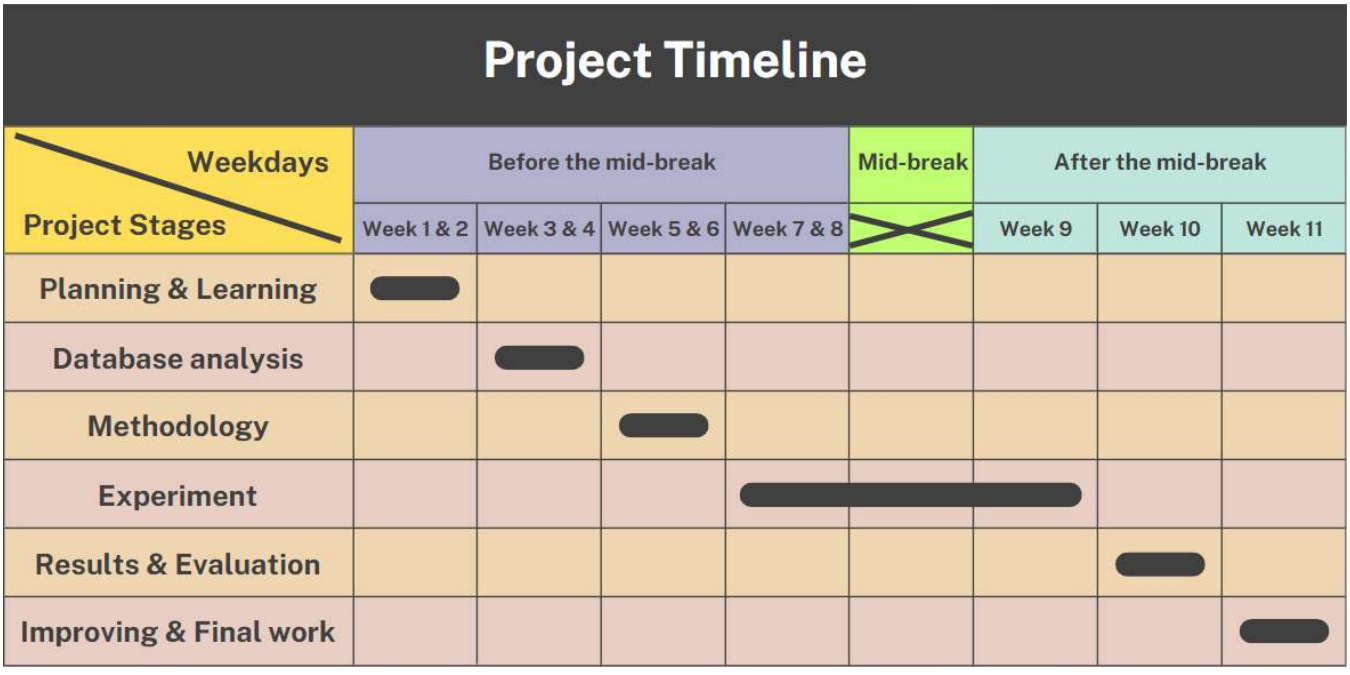

Figure 6. The Timeline of our project.