Automatic Generation of Multimedia Teaching Materials Based on Generative AI: Taking Tang Poetry as an Example

Xu Chen and Di Wu

***Abstract*—Generative artificial intelligence (AI) is widely recognized as one of the most influential technologies for the future, having sparked a paradigm shift in scientific research. The field of education has also been greatly impacted by this transformative technology, with researchers exploring the applications of generative AI, particularly ChatGPT, in education. However, existing research primarily focuses on generating text from text, and there remains a relative scarcity of studies on leveraging multimodal generation capabilities to address key challenges in multimodal data supported instruction. In this article, we present a technical framework for generating Tang poetry situational videos, emphasizing the utilization of generative AI to address the need for multimedia teaching resources. Our framework comprises three main modules: textual situational comprehension, image creation, and video generation. Moreover, we have developed a situational video generation system that incorporates various technologies, including text-to-text generation models, text-to-image generation models, image interpolation, text-to-speech synthesis, and video synthesis. To ascertain the efficacy of the modules within the Tang poetry situational video generation system, we undertook a comparative analysis utilizing the prevalent text-to-image and text-to-video generation models. The empirical findings indicate that our approach is capable of generating images that exhibit greater semantic similarity with the poems, thereby enabling a better comprehension of the poem's connotations and its key components. Concurrently, the Tang poetry videos generated can significantly contribute to the reduction of cognitive load and the enhancement of understanding during the learning process. Our research showcases the potential of generative AI in the education field, specifically in the domain of multimodal teaching resources.**

***Index Terms*—Automatic generation, generative artificial intelligence (AI), multimedia teaching materials, Tang poetry situational videos.**

I. Introduction

CHATGPT, launched at the end of 2022, has catalyzed a wave of generative artificial intelligence (AI) advancements, revolutionizing diverse fields. The joint support of large datasets, model parameters, and computing power has enabled the emergence of greater intelligence from large models (foundation models [1]). This intelligence includes in-context learning, instruction following, and step-by-step reasoning [2]. Thus, generative AI has emerged as a powerful accelerator for education in the process of digital transformation.

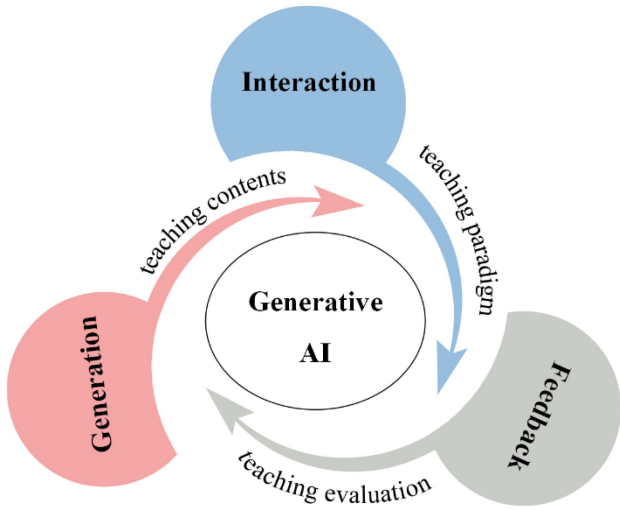

Fig. 1. Roles of generative AI in education.

The utilization of generative AI in the field of education can be categorized into three main aspects: generation, interaction, and feedback, as shown in Fig. 1. The "Generation" component aims to transform the conventional approach of teaching content production, which is primarily manual design conducted by educators. With the use of generative AI, an automatic generation process can take place for traditional teaching resources, such as syllabi [3], test questions [4], PPT courseware [5], and newer teaching resources, such as 3-D models and virtual teaching scenes [6]. The "Interaction" component revolutionizes the teaching interaction paradigm, placing machines in new roles, such as answering questions and providing speaking practice. Finally, the "Feedback" aspect modifies the means of teaching evaluation, as it replaces manual reviews with automatic machine corrections. Apart from merely evaluating content with standard answers, this technology can provide inspiring recommendations for relatively subjective tasks such as writing.

Despite the increasing popularity of generative AI in education, its landing remains limited. Schools have experimented

Manuscript received 10 July 2023; revised 1 December 2023 and 28 January 2024; accepted 13 March 2024. Date of publication 18 March 2024; date of current version 11 April 2024. This work was supported by the National Natural Science Foundation of China under Grant 42301496. *(Corresponding author: Di Wu.)*

The authors are with the Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China (e-mail: chenxu@ccnu.edu.cn; mr.wudi@163.com).

Digital Object Identifier 10.1109/TLT.2024.3378279

with generative AI in generating quizzes [7], practicing oral English [8], and reviewing essays [9], but these attempts have only utilized the capability of text generation. Research pertaining to the modality effect [10], [11] suggests that conveying information via multiple modals (for instance, video) can augment learning efficacy as opposed to utilizing a single modal (such as a picture or text). Therefore, there is a need to explore the integration of multimodal generation capabilities to help teachers and students provide multimedia teaching materials. According to a teacher–student survey report by McKinsey in multiple countries (Canada, Singapore, United Kingdom, United States), teachers spend 21% of their working time on preparation, which can be reduced by 47.6% using AI technology [12]. Retrieving and creating multimedia materials is a crucial aspect of teaching preparation. Thus, this article aims to explore the potential of generative AI in creating multimedia teaching resources using Chinese Tang poetry as an example. Regarded as a cornerstone of Chinese culture, Tang poetry is celebrated for its succinct diction. It frequently employs a myriad of expressive techniques, including metaphor, juxtaposition, synecdoche, and pun, thereby heightening the complexity of comprehending its textual content and forming mental imagery in comparison to other literary compositions. The development of multimedia instructional material designed for Tang poetry, which vividly illustrates the contextual scenarios of the poems, can facilitate students' understanding of Tang poetry by stimulating both their visual and auditory senses.

This article introduces an innovative system for the generation of situational videos correlating with Chinese Tang poetry, comprising three modules—text situational comprehension of Tang poetry, situational image creation, and situational video synthesis. Serving as an exploration of the application of multimodal generative AI within the context of "situational teaching," this system re-enacts the scenarios portrayed in Chinese Tang poems through video format. The primary objective of this system is to mitigate the cognitive burden experienced during the learning process and amplify learners' comprehension of Tang poetry.

II. Literature Review

A. Generative AI

Unlike discriminative AI, which is primarily dedicated to tasks such as image recognition, target detection, and trend prediction, generative AI is more oriented toward content generation as opposed to analysis. In the era of large language models, Generative AI can generate high-quality content, such as dialogues, images, codes, videos, and 3-D models. Currently, it has fulfilled the requirements for commercial use, particularly in the fields of text-to-text generation and text-to-image generation.

In the field of text generation, two mainstream technology routes based on the Transformer [13] architecture have evolved: Bert [14] and GPT [15]. OpenAI's GPT-3 [16], a milestone in large language models' development, demonstrated superior natural language capabilities with its ability of in-context learning and zero-shot prompting. InstructGPT [17] emerged as an improved version, using human feedback reinforcement learning to better align language models with human intention. As a result, ChatGPT has become a better conversational model. A series of models born in this period include LLaMA [18] and PaLM [19]. Furthermore, the multimodal model, GPT-4 [20], released in 2023 has expanded the boundaries of large language model capabilities even further.

The model architecture for generating images from text has shifted from generative adversarial networks [21] and variational autoencoders [22] to the simpler and better-performing Diffusion Model [23]. The Diffusion Model has a simpler training loss function, resulting in images of higher quality being produced. GLIDE [24] compared CLIP [25] guidance and classifier-free guidance strategies for generating images from text using Diffusion Models. Stable Diffusion [26] trains the diffusion model in the latent space, improving complexity reduction and detail preservation for the first time. The use of larger language models, such as T5 [27] as a text encoder significantly improves image fidelity and better aligns text and images, as demonstrated in Imagen [28]. DALL-E2 [29] uses the CLIP multimodal comparison model and introduces a two-stage text-to-image generation approach, explicitly generating image representations and training the prior with autoregressive methods or Diffusion models, which improve the diversity and realism of images.

B. Generative AI Applications for Education and Cognitive Load Theory

In the realm of AI-empowered education, numerous explorations have been conducted, such as teaching video analysis [30], MOOC learning effect prediction [31], and intelligent tutors [32]. However, the integration of generative AI is not yet common. Current research predominantly relies on OpenAI's API, and mainly caters to two types of teaching scenarios: self-directed learning and classroom instruction.

In terms of self-directed learning, Khan Academy has built a virtual teaching assistant as a trial based on GPT-4 to accommodate the diverse needs of students [33]. Duolingo employs GPT-4's role-playing capabilities to conduct conversations in different situations to aid learners in acquiring a new language [34].

For classroom instruction, most researchers use ChatGPT directly to generate interactive teaching materials, such as quizzes and flashcards [7], offer feedback [35], support literacy development [36], [37], and act as a pedagogical agent [38]. Few researchers have developed innovative educational applications with ChatGPT as a basis. Example of such an app is CGMap [9], offering an AI review of student works. iFLYTEK employs the Xinghuo cognitive model to facilitate rapid correction of Chinese and English compositions by teachers [39].

Overall, the utilization of generative AI in the field of education is still in its early stages and existing explorations mainly rely on the capacity for text generation. The successful application of multimodality, particularly in the realm of multimedia teaching materials, remains a salient requirement.

Mayer's [40] cognitive theory of multimedia learning based on the dual-channel assumption, the limited-capacity assumption, and the active-processing assumption, illustrates that multimedia extends beyond mere information transmission to offer

cognitive support for knowledge construction. Cognitive load theory [41] posits that the cognitive load engendered by illustrated narrations, such as videos, is less than that resulting from the utilization of illustrated texts. This is attributed to the simultaneous stimulation of verbal and visual working memory [42] when listening and observing concurrently, thereby mitigating the overloading of a singular visual channel caused by the retrieval of words and word patterns from long-term memory [43]. Variations in cognitive load subsequently give rise to modal effects within the learning process [44].

C. Tang Poetry Visualization

Numerous researchers have conducted research to automatically generate Tang poems from images using deep learning methods [45], [46], [47]. However, there is limited research on automatically visualizing Tang poems, such as generating situational images or videos based on the content of the poems. Li et al. [48] published the first Chinese poetry art visualization dataset and evaluated it using state-of-the-art text-to-image generation methods (AttnGAN [49] and MirrorGAN [50]). Although these methods can produce images with high visual fidelity, their expression of poetry semantics is limited.

The artistic conception expressed in ancient Chinese poetry differs significantly from the images typically found on the Internet. Pretrained text-to-image generation models often struggle to grasp the underlying meanings in ancient poems and the images they generate typically do not adhere to the traditional Chinese art style. Consequently, this article proposes a combination of a large language model and a text-to-image model to divide the visualization of Tang poetry into two phases: connotation understanding and image generation.

III. Tang Poetry Situational Video Generation System

While models such as Stable Diffusion [26] and Make a video [51] can create images and videos from textual inputs, they encounter two primary challenges when applied to Chinese Tang poetry: 1) the inability to accurately comprehend the connotation and artistic conception of Tang poetry and 2) the failure to directly generate imagery in the Chinese landscape painting style, which is essential for portraying and conveying ancient Chinese art. Consequently, this article presents a system for situational video generation that is tailored to Chinese Tang poetry, comprising three stages: textual situational comprehension, situational image creation, and situational video synthesis.

In this article, "situation" specifically denotes the artistic conception encapsulated in Tang poetry, the represented environment, and the corresponding traditional Chinese artistic style. "Connotation" explicitly denotes the author's emotions, vision, and the backdrop implicitly embedded in the poem. The textual situational comprehension primarily addresses the first challenge: accurately comprehending the connotation of Tang poetry. Conversely, the situation image creation module concentrates on the second challenge: generating images in the style of Chinese landscape painting.

The system's workflow shown in Fig. 2 unfolds as follows.

1) The input consists of the entire Tang poetry.
2) Each poem undergoes sentence decomposition, wherein punctuation marks facilitate the segmentation into individual sentences, serving text situational comprehension and situational video synthesis.
3) The text situational comprehension module comprehends the text's context, generates interpretations of single Tang poetry sentences using a large language model, and extracts keywords from twelve dimensions to generate prompts for Stable Diffusion.
4) The situational image creation module employs the Stable Diffusion model, producing situational images corresponding to each line of the Tang poetry. ControlNet is used to interpolate image key frames between distinct Tang poetry sentences.
5) In the situational video synthesis module, the original Tang poetry text is transformed into audio, which is then combined with the image generated in the fourth step to synthesize a comprehensive situational video.

Among all the steps, textual situational comprehension, situational image creation, and video synthesis are fundamental to the ultimate situational video outcome. The following sections provide an in-depth analysis of these three core modules.

A. Textual Situational Comprehension

The purpose of the Textual Situational Comprehension Module is to transform individual sentences from Tang poems into prompts suitable for Stable Diffusion. This entails addressing two primary challenges: analyzing the sentences of Tang poetry and generating prompts infused with an ancient Chinese artistic style. The analysis of Tang poetry can be complex due to its stylistic refinement and the prevalent use of emotional metaphors by the authors, leading to substantial loss of implicit information when directly translated from Chinese to English. As for the process of prompts generation, the content and quality of the image hinges on several key elements, such as characters, scenery, and objects, in addition to the artistic style.

To address the challenges previously discussed, we decompose textual situational comprehension into three distinct submodules: paraphrase, key element extraction, and prompts generation. These submodules are autonomously processed using the large language model, GPT-3.5-Turbo. Our primary contribution resides in enabling this language model to systematically break down the process and craft appropriate prompts to simultaneously extract the connotations and artistic conception inherent in Tang poetry.

For the paraphrase submodule, the large language model assumes the role of an author, tasked with the analysis of a specified poem. Beyond the poem's literal interpretation, it also endeavors to provide pertinent background context and elucidate the emotions conveyed by the original author.

Within the key element extraction submodule, a 12-dimensional framework has been devised to aid in the mapping of paraphrases and situational imageries derived from Tang poetry. These dimensions encompass "Subject," "Action," "Context," "Environment," "Lighting," "Artist," "Style," "Medium," "Type," "Color Scheme," "Computer Graphics," and "Quality."

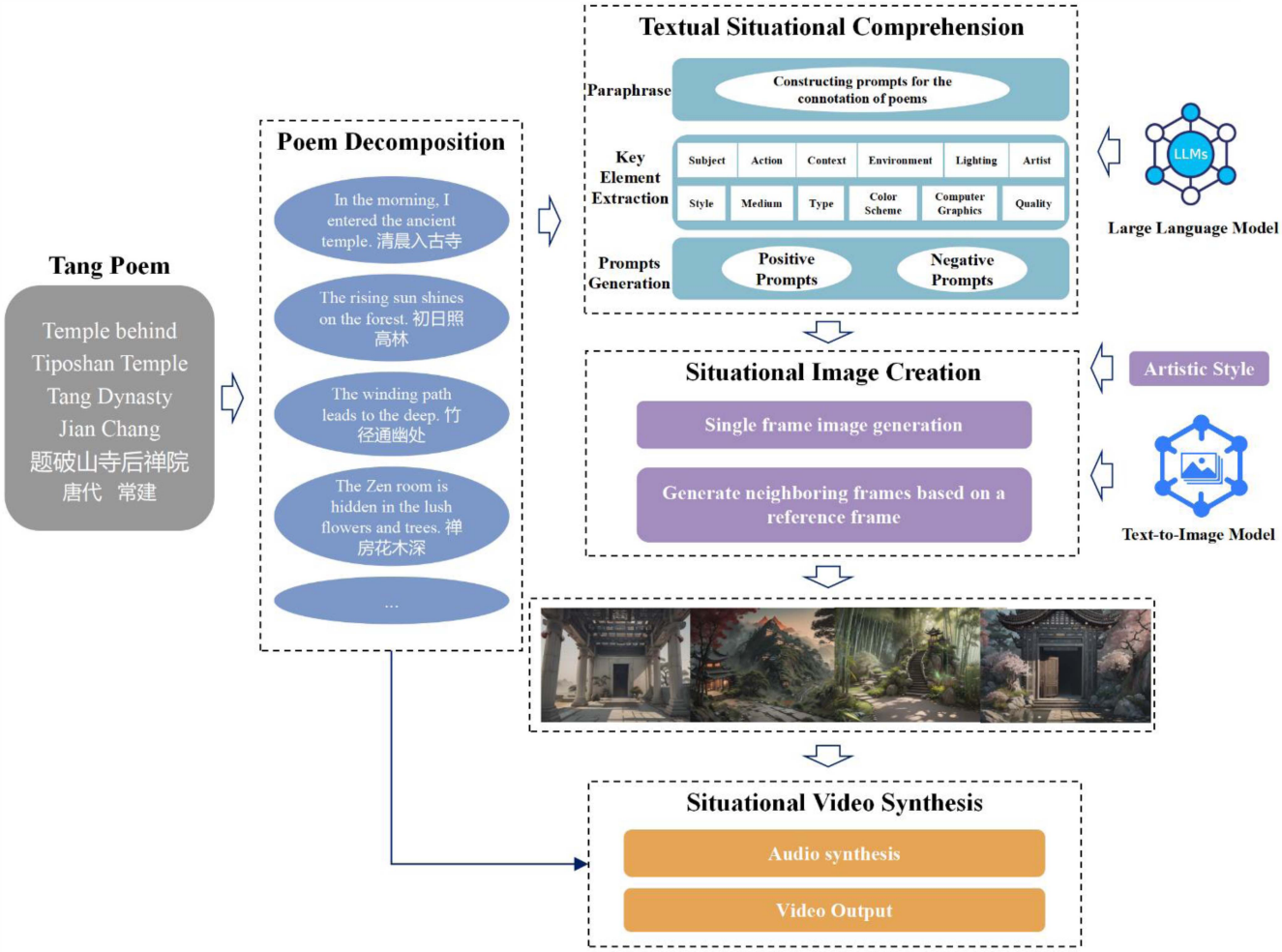

Fig. 2. Proposed Tang poetry situational video generation system, consisting of three main modules.

For instance, "Subject" incorporates elements such as people, animals, landscapes, and buildings. "Actions" span activities such as running, jumping, swimming, and falling. "Medium" includes forms such as oil on canvas, watercolor, sketch, photography, and Chinese ink painting with a rice paper texture. "Color Scheme" comprises variations such as pastel, vibrant, dynamic lighting, and colors such as green, orange, and red. "Computer Graphics" includes elements, such as 3-D, octane, and cycles, while "Quality" covers aspects, such as high definition, 4K, 8K, 64K, masterpiece, and superior quality. The extraction of keywords across these 12 dimensions is performed utilizing the output derived from the paraphrase submodule.

The prompts for Stable Diffusion comprise both positive and negative prompts. Positive prompts are generated through an amalgamation of the 12-dimensional keywords, extracted by the large language model, taking into account the provided examples as references. Conversely, the negative prompts employ a consistent set of predefined prompts, including "Low quality," "Logo," "Text," "Watermark," "Username," "Chinese text," and "Stamp mark." Furthermore, to enhance the representation of the ancient Chinese art style in the final result, we employed Lora models, integrating related Lora models and their proportions in the primary prompts. The complete prompts are given in Appendix A. The comparison of prompts obtained through the Textual Situational Comprehension module and direct translation is given in Table I. The comparative results indicate that the prompts generated by the textual situational comprehension module more aptly describe the circumstances in the poem, encapsulating elements such as the overall backdrop, positional relationships between subjects, color descriptions, and the painting style.

B. *Situational Image Creation*

The situational image creation module constitutes the most crucial aspect of the entire system, supplying visual resources for the ultimate situational video. This process involves two distinct submodules: single-frame image generation and interframe image interpolation. These submodules are tasked with addressing two significant challenges: generating images that align with various styles, and producing controllable generation grounded in reference frames.

TABLE I
COMPARISON OF PROMPTS GENERATED BY DIFFERENT METHODS

Method	**Positive Prompts for** "Beyond the bamboo grove, two or three branches of peach blossoms have started to unfold, heralding the advent of spring. Ducks frolic in the water, discerning the initial warmth returning to the river in early spring. 竹外桃花三两枝，春江水暖鸭先知"	**Positive Prompts for** "On an ancient trail, under the melancholic western wind, trots a gaunt horse, embodying a desolate and lonely tableau. 古道风瘦马"	**Positive Prompts for** "In the corner of a wall, a handful of plum branches demonstrate resilience, daring to bloom alone amidst the piercing winter chill. 墙角数枝梅，凌寒独自开"
Translation	Outside the bamboo, a few branches of peach blossoms, the duck is the first to know the river's warm in spring.	A skinny horse walks hard on the ancient road against the west wind.	There are a few plum blossoms in the corner, blooming alone in the cold weather.
Textual Situational Comprehension: key element extraction	Subject: Ducks Action: Playing in the water Context: In the early spring, outside the bamboo forest Environment: Near blooming peach trees Lighting: Soft sunlight filtering through the trees Artist: Xu Beihong Style: Traditional Chinese painting Medium: Digital painting Type: Landscape Color Scheme: Vibrant and lively Computer Graphics: 2D Quality: High definition, 8K	Subject: A gaunt horse Action: Trotting Context: On an ancient trail Environment: Outdoor, under the melancholic western wind Lighting: Soft, ambient Artist: Ziqian Zhan Style: Chinese ink painting with rice paper texture Medium: Digital painting Type: Landscape Color Scheme: Pastel, muted Computer graphics: 3D, Octane Quality: 8K	Subject: A handful of plum branches Action: Blooming Context: In the corner of a wall Environment: Outdoor, winter season Lighting: Soft, ambient Artist: Sixun Li Style: Chinese ink painting with rice paper texture Medium: Digital painting Type: Still life Color Scheme: Pastel, muted Computer graphics: 3D, Octane Quality: 8K
Textual Situational Comprehension: prompts generation	**Positive Prompt:** In the early spring, the earth is in recovery. A bamboo forest, painted in vibrant shades of green, signifies the growth of new leaves. Peach trees dot the forest, their branches adorned with early-blooming peach blossoms. The blossoms bloom brightly in hues of pink, creating a striking contrast against the verdant backdrop. In the midst of this scenic beauty, ducks frolic in the cool blue waters. Their joyous activity adds life to the serene environment. Inspired by the art style of Xu Beihong, known for his Chinese ink paintings with a rice paper texture, this masterpiece is created in a digital painting medium. The artwork is a landscape type, encapsulating the harmonious interaction of wildlife and nature. The color scheme is vibrant and lively, reflecting the energy of the early spring. Rendered in high-definition 2D, this piece captivates with its intricate details. The quality of the artwork is impeccable, presented in a stunning 8K resolution. **Negative Prompt:** Poorly drawn hands, low quality, distorted face, bad anatomy, beginner level, watermark, text, logo	**Positive Prompt:** (Best Masterpiece), 8K, (Chinese ancient painting style), (distant view shot), The sight passes through the barren landscape to reveal a gaunt horse trotting on an ancient trail, under the melancholic western wind, embodying a desolate and lonely tableau. The sky filled with layers of gray and white clouds, the soft ambient light casts long shadows, highlighting the horse's skeletal frame. **Negative Prompts:** worst quality, low quality, logo, watermark, username, Chinese text, stamp mark, 2 heads, 2 faces, cropped image, out of frame, draft, deformed hands, signatures, double image, long neck, multiple heads, missing limb, disfigured, grain, low-res, deformed, blurry, disfigured, mutation, mutated, long body, disgusting, poorly drawn, mutilated, mangled, surreal, extra fingers, duplicate artifacts, morbid, gross proportions, malformed, ugly, tiling, poorly drawn feet, out of frame, extra limbs, disfigured, deformed, body out of frame, bad anatomy, watermark, signature, cut off, low contrast, underexposed, overexposed, bad art, beginner, amateur, blurry, draft, grainy.	**Positive Prompt:** (Best Masterpiece), 8K, (Chinese ancient painting style), (close-up shot), A handful of plum branches demonstrate resilience, daring to bloom alone in the corner of a wall amidst the piercing winter chill. The branches, adorned with delicate plum blossoms, stand out starkly against the austere wall, their vibrant colors a stark contrast to the surrounding winter landscape. The image is a testament to resilience and beauty amidst harsh conditions, rendered in exquisite detail and beautiful color tones. **Negative Prompts:** worst quality, low quality, logo, text, watermark, username, Chinese text, stamp mark, cropped image, out of frame, draft, signatures, double image, disfigured, cut-off, grain, low-res, deformed, blurry, disfigured, mutation, mutated, disconnected limbs, long body, disgusting, poorly drawn, mutilated, mangled, surreal, duplicate artifacts, morbid, gross proportions, malformed, ugly, tiling, bad anatomy, watermark, signature, cut off, low contrast, underexposed, overexposed, bad art, beginner, amateur, blurry, draft, grainy.

Fig. 3. Different situational images were generated using fine-tuned models of various styles for the same tang poem "ducks frolic in the water, discerning the initial warmth returning to the river in early spring." Arranged from left to right and top to bottom, the fine-tuned models deployed for image generation correspond to models 4, 2, 3, and 7 outlined in Appendix B.

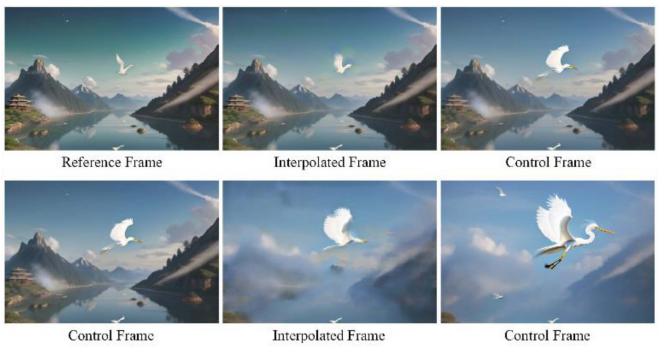

Fig. 4. Generate neighboring frames based on a reference frame. The reference frame represents the first frame image of the poem generated by stable diffusion. The control frame represents the key frame generated with ControlNet that change only the subject. The interpolation frame means an intermediate frame generated from the reference frame and the control frame by using RIFE. Refer to Appendix C for specific prompts.

Leveraging the pretrained Stable Diffusion model directly often results in the generation of images that are excessively realistic and incongruous with the scenes depicted in Chinese Tang poems. In order to address this issue, a drop-down list has been integrated into the system interface enabling the selection of various artistic styles. Each style corresponds to a distinct fine-tuned model based on the Stable Diffusion v1-5 Model. The renderings of these diverse artistic styles are illustrated in Fig. 3. For the corresponding model links, refer to Appendix B.

Upon generation of the initial frame image, we proceed to augment this image in a two-step process, as shown in Fig. 4. First, we employ ControlNet's ip2p model [52] in conjunction with two sets of prompts to guide Stable Diffusion in generating two distinct control frame images. The generated control frame image maintains a style and background akin to the reference frame image; however, alterations occur in the position and size of the subject within the image. Each set of prompts performs a specific role: one brings the subject in the poem closer to the lens and retains the background, while the other only retains the subject. Following this, we utilize the RIFE model [53] to execute two interpolations based on the reference frame and two control frames. During each interpolation process, four iterations are performed, yielding 17 images. Thus, after two interpolations, a cumulative total of 34 images are procured. Given that one image is duplicated and the frame rate of the synthesized video typically conforms to an even number, we ultimately retain 32 images for each Tang poem. The seamless generation of intermediary frames hinges primarily on the capacity of ControlNet's ip2p model to modify the subject's state, and action as per the prompts while preserving the style and remaining content intact.

C. Situational Video Synthesis

The multimedia presentation of Tang poetry comprises both auditory and visual elements. The auditory component derives from the original Tang poetry text, while the visual content emerges from the earlier phase of situational image creation. The complete procedure encompasses three key stages: speech synthesis, subtitle generation, and video synthesis. We employ the speech_sambert-hifigan_tts_zhiyan_emo_zh-cn_16k model [54], [55] from ModelScope for speech synthesis, transforming text into audio. Both the creation of subtitles and video synthesis rely on the moviepy [56] open-source library. The auditory and visual elements are synchronized based on audio duration, with subtitles centrally positioned at the video's right portion. To ensure seamless transitions between different poems, we utilize the RIFE model to interpolate the concluding frame of the preceding poem and the initial frame of the succeeding one, thereby crafting a 1-s transition.

IV. Experimental Results

To holistically evaluate our proposed Tang poetry situational video generation system, we utilized various text-to-image generation models to validate the effectiveness of the textual situational comprehension module in this section. Simultaneously, we contrasted the video content produced by mainstream text-to-video generation tools. Throughout all comparative experiments, we abstained from utilizing the conventional text-to-image quality evaluation metric, the Fréchet inception distance (FID) [57], owing to the absence of reference images for Tang poetry. Given that multimedia teaching materials for Tang poetry should effectively portray the content within the poetry, we designed an evaluation metric grounded in semantic similarity. This metric initially employs the visual language foundation model CogVLM [58] to generate descriptive text for the synthesized image.

Subsequently, the Sentence Transformers framework [59] is used to extract the semantics of both the poem's interpretation and the descriptive text in order to compute the cosine similarity for comparative purposes.

A. Effectiveness of Textual Situational Comprehension

Besides Stable Diffusion, other notable text-to-image generation models include DALL•E2 [60], Midjourney [61],

Fig. 5. Comparison of images generated by different methods for "Along the riverbank, peach blossoms are flourishing, accompanied by the gradual rise of the spring water. Within the water, the mandarin fish thrive, displaying their succulent and nourished state."

TABLE II
SEMANTIC SIMILARITY SCORES FOR DIFFERENT MODELS

Method	Semantic similarity score	
	W/O Textual Situational Comprehension	W/ Textual Situational Comprehension
DALL•E2	0.4436	**0.5566**
Midjourney	0.4159	**0.5153**
Tongyi	0.4803	**0.5741**
Stable Diffusion	0.3925	**0.5499**
locs-china-landscapes-v2	0.4043	**0.5189**

Fig. 6. Comparative results of three groups of prompts for "along the riverbank, peach blossoms are flourishing, accompanied by the gradual rise of the spring water. Within the water, the mandarin fish thrive, displaying their succulent and nourished state."

Wenxinyige [62], and Tongyi [63]. Given that Wenxinyige only accommodates Chinese text input, we refrained from conducting a comparative analysis with it. During the verification process, beyond employing mainstream text-to-image generation models, we also took into account fine-tuned models based on Stable Diffusion v1.5, such as locs-China-landscapes-v2 in Appendix B, which integrates Chinese paintings into its training data. A total of 20 Tang poems were randomly selected for comparison. To maintain evaluation input consistency, we supplied the same pair of prompts to all models: one set comprised the English translation of the poem's Chinese interpretation with a Chinese landscape style, and the other set was the output from the textual situational comprehension module. The results in Table II indicate that the prompts derived from the textual situational comprehension module can significantly enhance the semantic similarity of the images produced by various models, thereby more accurately representing the objects referenced in the poem.

As shown in Fig. 5, we arbitrarily selected one poem to juxtapose the images produced before and after the application of the textual situational comprehension module, aiming at more effectively illustrating the comparison results. The comparison results reveal that the image produced solely from the interpretation of the poem typically encompasses only a portion of the subject matter portrayed in the poem. The Textual Situational Comprehension module decomposes the elements in the poem across different dimensions, aiding the model in more effectively comprehending the embedded semantics. Consequently, the produced image more closely aligns with the poem's content.

To delve deeper into the influence of the 12 dimensions outlined in the Textual Situational Comprehension module on the resultant images, we categorized these dimensions into three distinct groups: subject, background, and style for further examination. Specifically, the "subject" encompasses the dimensions of "Subject" and "Action"; the "background" comprises "Context," "Environment," and "Lighting," whereas the "style" incorporates "Artist," "Style," "Medium," "Type," "Color Scheme," "Computer Graphics," and "Quality"—seven dimensions in total. Fig. 6 presents the analytical findings of a specific poem. The results indicate that insufficient analysis of the subject within the prompts may lead to a partial omission of the subjects in the generated image. A similar absence of background analysis in the prompts may yield generated images devoid of artistic

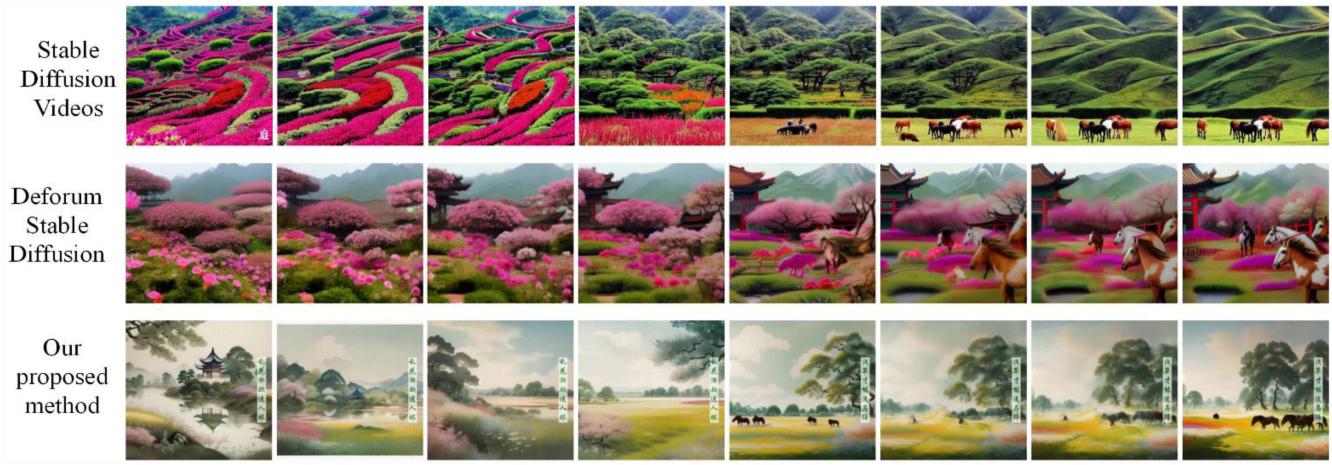

Fig. 7. Comparison of eight frames from the videos generated by different methods for "the numerous flowers are gradually opening to make people dazzled, and the shallow green grass is just enough to cover the hooves of the horses."

TABLE III
SEMANTIC SIMILARITY SCORES FOR DIFFERENT TEXT-TO-VIDEO GENERATION TOOLS

Method	Semantic similarity score
Stable Diffusion Videos	**0.5557**
Deforum Stable Diffusion	0.4283
Our	0.5543

backgrounds. Furthermore, the lack of stylistic analysis within the prompts can cause the produced images to appear overly realistic.

The comparative image generation experiment underscores that thorough situational comprehension is a prerequisite for effective creative situational image representation, particularly for Tang poetry, characterized by its refined language and profound implications. By constructing suitable prompts to facilitate the large language model's improved interpretation of Tang poetry, the quality of automated situational image creation of Tang poetry can be substantially enhanced.

B. Comparison With Text-to-Video Generation Tools Based on Stable Diffusion

In evaluating the quality of educational videos, Audio, Duration, Visual, Narration, and Content are typically employed as indicators [64], [65]. Given that audio and subtitles are auto-generated based on Tang poetry text, this study primarily concentrates on the visual and content parameters of the produced video. Recognizing that the two applications, Stable Diffusion Videos [66] and Deforum Stable Diffusion [67], utilize Stable Diffusion for video synthesis, a comparison was conducted with these two applications, focusing on visual and content metrics. With respect to content, semantic similarity is employed to evaluate comprehension. Specifically, the multimodal large model CogVLM is treated as a "student" tasked with providing descriptive text for the generated image. Subsequently, the semantic similarity between this descriptive text and the poem's interpretation is compared.

Throughout the evaluation procedure, each Tang poem was consistently set to generate 32 frames, with a uniform image size of 512 $\times$ 512 pixels for three methods. Both Stable Diffusion Videos and Deforum Stable Diffusion were given the English translation of the poem's Chinese interpretation as the prompt, supplemented with a Chinese painting style. In total, five poems, equating to 480 frames, are evaluated. As evidenced by the results presented in Table III, the comprehension score for Stable Diffusion Videos marginally surpasses ours, while Deforum Stable Diffusion records the lowest score. Upon analyzing the test data, it was noted that the interpolated frames introduced by our method contributed to a decline in the overall semantic similarity score. Contrasting with our approach of interpolating in the image's latent space, Stable Diffusion Videos interpolates within the text's latent space, thereby safeguarding the semantic comprehension of the intermediate frames.

However, as illustrated in Fig. 7, even though the Chinese painting style is incorporated into the Stable Diffusion Videos' prompt, the resultant image does not exhibit a distinct style and lacks background cohesion compared to the image latent space interpolation methods. In the context of personalized text-to-video generation tasks, determining how to differentiate the subjects from the background in the text latent space, interpolate the subject semantics while maintaining background consistency, and effectively infuse a personalized style is worthy of further research.

C. Evaluation Experiments on Cognitive Load and Learning Effectiveness

To assess the influence of Tang poetry situational videos, produced by our devised system, on cognitive load and learning outcomes, we executed comparative experiments employing two distinct learning conditions: generated videos and PPT. This study was conducted in collaboration with a primary school in a

city in China. The developed Tang poetry multimedia teaching resources were utilized to support the teaching of Tang poetry in its Chinese course. The experimental participants were sourced from a third-grade class by the teacher in the school, amounting to 20 participants (11 females and 9 males) with an average age of 8.85 years (SD = 0.37). Despite their diverse academic performances, the students had not previously encountered the Tang poetry assessed in this study as part of their formal classroom instruction. Participants were randomly allocated to either the video condition group or the PPT condition group. The allocated learning time for each group was 5 min, during which the participants were permitted to review the video or PPT repeatedly. Upon conclusion, all participants were required to complete a paper-based questionnaire. To mitigate the cognitive load imposed by the learning environment, both groups engaged in independent learning within familiar classrooms. The subject matter for learning was "Fishing Song 渔歌子," a piece by Zhang Zhihe, a poet from the Tang Dynasty. The instructional materials utilized by the video group comprised videos generated by our system, while the PPT group used materials that included text and image excerpts from the People's Education Press courseware in Chinese.

The comprehensive questionnaire comprised three sections: demographic data, cognitive load measurement, and knowledge assessment. Demographic data encompassed three inquiries concerning gender, age, and prior exposure to the poem in question. The cognitive load was gauged using a nine-point self-rating scale of mental effort [17] wherein students were required to indicate the level of exertion experienced during the learning process, with 0 signifying minimal effort and 9 denoting maximal effort. This self-report scale of mental effort enjoys widespread use and is deemed a reliable and valid estimator of cognitive load [68], [69]. The knowledge assessment encompassed both retention and comprehension (refer to Appendix D), procured from the Baidu question bank. Retention was evaluated through single-choice and fill-in-the-blank questions, while comprehension was assessed using fill-in-the-blank and open-ended questions. Each item was scored according to predefined guidelines, allotting one point for correct answers and zero points for incorrect responses, with a provision for partial scoring in certain items.

Survey results revealed that 80% of participants had prior exposure to the poem. The statistical results under the two conditions are presented in Table IV. The knowledge assessment Cronbach's alpha value of 0.966 denotes the reliability of the questionnaire. On average, the cognitive load score of participants during the learning process was 7.25 (SD = 1.48). This high score corresponds to a higher level of effort, suggesting that Tang poetry presents a considerable challenge for third-grade students to comprehend. The cognitive load experienced by students in the video condition group was lower than that of the PPT condition group, demonstrating the modality effect in the learning process. This suggests that multimodal teaching materials assist in reducing cognitive load. In terms of learning effectiveness, the video condition group demonstrated superior performance compared to the PPT condition group, particularly with regard to retention, indicating that the situational videos generated by our system facilitate a better understanding of the scenes portrayed in the poems among students.

TABLE IV
DESCRIPTIVE STATISTICS OF COGNITIVE LOAD AND LEARNING EFFECTIVENESS ACROSS CONDITIONS

		N of items	Video Condition	PPT Condition
Cronbach's alpha coefficient			0.966	
Cognitive load	Mean	1	6.5	8
	SD		1.27	1.33
Retention	Mean	3	91.9	60.9
	SD		0.16	0.04
	Min		61.9	57.1
	Max		100	66.7
Comprehension	Mean	2	40	25
	SD		0.29	0.26
	Min		0	0
	Max		75	50

Note. The scores for both Retention and Comprehension were respectively converted into percentages.

Based on the comprehensive analysis of the questionnaire results, it is evident that the proposed system for generating situational videos of Tang poetry has demonstrated the ability to accurately convey key elements of Tang poetry. This, in turn, contributes to reducing cognitive load and enhancing the learning effectiveness of Tang poetry. The comprehensive survey showcases the potential of generative AI in the creation of multimodal teaching resources.

V. DISCUSSION

A. General Discussion

This article presents the development of an automatic generation system for situational videos of Tang poems using generative AI. The primary objective is to investigate the potential application of generative AI in teaching and provide a viable pathway for harnessing the capabilities of multimodal generative AI.

The Textual Situational Comprehension module we developed aids in enhancing the semantic similarity between the generated images and the poems. This improvement is attributed to the analysis of the poems across 12 dimensions, thereby facilitating better alignment of the poem's content. In contrast to the method of text semantic space interpolation used for video generation, the developed Situational Image Creation module is capable of maintaining a balance between generating poem-aligned content and ensuring image continuity, achieved through the generation of key frames and the interpolation of image semantic space.

The findings from the questionnaire suggest that when engaging with Tang poetry, which necessitates the construction of a visual understanding, the situational videos produced by our system can have a beneficial impact on reducing students' cognitive load and enhancing learning outcomes.

In addition, our research uncovers the potential applications of multimodal generative AI within the educational field. For instructional content characterized by a high cognitive load, the development of multimodal teaching materials proves to be an effective strategy for reducing this load. The quality of such multimodal teaching resources, created using generative AI, is contingent upon tailored prompt engineering or text-to-image and text-to-video generation models.

B. Limitation and Future Direction

Despite the capability of the proposed Tang poetry situational video generation system to support the automatic generation of such videos, certain limitations remain.

First, the system's capacity of simulating dynamic effects is constrained. The current situational image creation module struggles to effectively simulate moving elements such as flowing rivers, fluttering leaves, or falling rain and snow, thus hindering the creation of realistic scenarios that can enhance user immersion during the viewing process. The integration of generative AI with 3-D game engines may represent a crucial avenue for automating multimedia teaching resource production in the future.

Second, simply generating situational videos that align with the content of the poems is insufficient to facilitate an understanding of Tang poetry. It is vital to augment the teaching materials with additional information, such as background details about the poem and explanations of the poem's content. This could further aid students in comprehending and memorizing Tang poetry. The challenge lies in designing an automated process that can effectively organize this information and create seamless visual connections, which remains a key consideration in the design of Tang poetry teaching resources.

Finally, atypical scenes sporadically emerge in the autogenerated Tang poetry situational images such as the amalgamation of multiple animal figures. This occurrence is typically attributed to the interplay among multiple subjects within the poem. To tackle this issue, it is imperative to construct an evaluation model, employing a vast array of Chinese-style artworks to fine-tune a large multimodal model. Throughout the image generation procedure, the produced images are assessed based on three criteria: semantic similarity, content comprehensibility, and human preference, followed by iterative optimization to guarantee the quality of the generation.

VI. Conclusion

The advanced capabilities showcased by large language models, such as GPT4, have led to a surge of interest in generative AI within academia and industry, with it being perceived as a pivotal transformative technology. Despite this, the applications of generative AI in the production of teaching resources remain limited, primarily relying on text-to-text generation models for tasks, such as the creation of conversational drills and quizzes. Furthermore, there is a notable scarcity of efforts toward the intelligent generation of multimodal teaching resources. Given the modality effect instigated by diverse modal teaching resources, especially for learning content that induces high cognitive load, there is an immediate need to employ multimodal generative AI to provide multimedia teaching resources for educators and learners. To address this gap, this study takes Chinese Tang poetry as an example and employs generative AI to construct an intelligent generation system of Tang poetry situational videos. The primary aim of this system is to equip educators and learners with Tang poetry situational videos, thereby aiding in the reduction of cognitive load during the learning process and enhancing their comprehension of Tang poetry.

The crucial aspect of replicating the content of Tang poetry lies in discerning the key elements, such as the environmental and subject-related aspects concealed within it. Comparative experiments demonstrate that the textual situational comprehension module, developed in this study, can enhance the generation quality of various text-to-image generation models in the context of Tang poetry situational images. In addition, the Tang poetry situational videos produced by our system play a significant role in alleviating cognitive load and augmenting understanding of Tang poetry content.

For generative AI-based systems to be broadly applied in the field of education, challenges such as the scientific evaluation of generated content and the procedural simulation of dynamic content must be addressed. As such, future research will concentrate on the amalgamation of generative AI and 3-D game engines to facilitate the intelligent generation of multimodal teaching materials that adhere to physical laws. Moreover, we will design an automated process and criteria to scientifically evaluate the content generated by generative AI.

Appendix

A. Prompts for Textual Situational Comprehension

The complete prompts utilized for the extraction of 12-dimensional keywords and the generation of Stable Diffusion prompts are as follows:

"Stable Diffusion is an AI art generation model similar to DALLE-2."

It can be used to create impressive artwork by using positive and negative prompts. Positive prompts describe what should be included in the image.

Very important is that the Positive Prompts are usually created in a specific structure.

(Subject), (Action), (Context), (Environment), (Lightning), (Artist), (Style), (Medium), (Type), (Color Scheme), (Computer graphics), (Quality), (etc.)

Subject: Person, animal, landscape

Action: Dancing, sitting, surveil

Verb: What the subject is doing, such as standing, sitting, eating, dancing, surveil

Adjectives: Beautiful, realistic, big, colorful

Context: Alien planet's pond, lots of details

Environment/Context: Outdoor, underwater, in the sky, at night

Lighting: Soft, ambient, neon, foggy, misty

Emotions: Cosy, energetic, romantic, grim, loneliness, fear

Chinese Artist: Ziqian Zhan, Baishi Qi, Sixun Li, Tong Guan

Art medium: Oil on canvas, watercolor, sketch, photography, Chinese ink painting with rice paper texture

Style: Polaroid, long exposure, monochrome, GoPro, fisheye, Bokeh, Photo, 8K UHD, DSLR, soft lighting, high quality, film grain, Fujifilm XT3

Art style: Landscape, painting of birds and flowers, minimalism, abstract, graffiti

Material: Fabric, wood, clay, Realistic, illustration, drawing, digital painting, photoshop, 3-D

Color scheme: Pastel, vibrant, dynamic lighting, Green, orange, red

Computer graphics: 3-D, octane, cycles

Illustrations: Isometric, pixar, scientific, comic

Quality: High definition, 4K, 8K, 64K

example Prompts:

1) (Best Masterpiece), 8K, (Chinese ancient painting style), (distant view shot), the sight passes through the bamboo forest and leaves to reveal a Chinese ancient garden, with multiple ancient Chinese buildings scattered among the mountains and forests, high mountain waterfalls, ancient Chinese pavilions and waterfalls on cliffs in the distance, ponds, small flowing bridges, lush trees, abundant branches and leaves, beautiful natural lighting, blue sky and clouds, beautiful color tone, exquisite composition, film composition, (depth of field effect), (atmospheric perspective), (exquisite and delicate details).
2) (Ultra wide-angle lens), (Best masterpiece), 8K, An ancient Chinese town nestled in mountains and hills, with lush trees, verdant leaves, murmuring streams. The visuals are exquisite and beautiful, brimming with sunshine (white background:1.5), white background, outside border.
3) This award-winning masterpiece is an incredibly detailed and photorealistic CG unity 8K wallpaper capturing the beauty of a classical Chinese garden. The stunning landscape features a serene lake and river surrounded by lush vegetation and majestic trees. The dramatic scenery is enhanced by the natural light and the blue sky, dotted with fluffy clouds. The waterfall is a focal point, adding depth and movement to the image. The use of Bokeh, Depth of Field, HDR, Bloom, Chromatic Aberration, and Intricate detail makes this an exceptional work of art. This piece is trending on both ArtStation and CGsociety, and has earned high praise for its unparalleled quality and intricate attention to detail.
4) (Masterpiece), (best quality), 8K, no humans, looking from afar, there are beautiful ancient towns in China, including the ancient capital of Chang'an, with quaint buildings and winding alleys. At night, lanterns are hung high and layers of buildings stand in the dim light. In the distance, there are mountains and flowing water, and nearby, there are tree branches and leaves. The small town is beautiful and lively, with fireworks lighting up the night sky, making it incredibly beautiful.
5) Highly detailed, majestic royal tall ship on a calm sea, realistic painting, by Charles Gregory Art station and Antonio Jacobsen and Edward Moran, (long shot), clear blue sky, intricate details, 4K.

Negative prompts are things you do not want to be included in the generated images, everything in one word divided by only commas not period.

Use this Negative Prompt and add some words what you think that match to Prompt: worst quality, low quality, logo, text, watermark, username, Chinese text, stamp mark, two faces, cropped image, draft, deformed hands, signatures, twisted fingers, double image, long neck, malformed hands, multiple heads, extra limb, poorly drawn hands, missing limb, disfigured, cut off, grain, low-res, blurry, disfigured, poorly drawn face, mutation, mutated, floating limbs, disconnected limbs, long body, disgusting, poorly drawn, mutilated, mangled, surreal, extra fingers, duplicate artifacts, morbid, gross proportions, missing arms, mutated hands, cloned face, malformed, ugly, tiling, poorly drawn hands, poorly drawn feet, poorly drawn face, disfigured, body out of frame, bad anatomy, watermark, signature, low contrast, underexposed, overexposed, bad art, beginner, amateur, distorted face, blurry, draft, grainy, etc.

Very important: Use an artist matching the art style, or do not write any artist if it is realistic style or some of that.

I want you to write me one full detailed prompt about the idea written from me, first in (Subject), (Action), (Context), (Environment), (Lightning), (Artist), (Style), (Medium), (Type), (Color Scheme), (Computer graphics), (Quality), (etc.). Then in Positive Prompt: write in next line for Positive Prompt: follow the structure of the example prompts; and Negative Prompts: write in next line for Negative Prompts about the idea written from me in words divided by only commas not period. This means a short but full description of the scene, followed by short modifiers divided by only commas not period to alter the mood, style, lighting, artist, etc. Write all prompts in English.

B. Fine-Tuned Model Links

The links to several fine-tuned models based on Stable Diffusion v1-5 Model utilized in the process of generating Tang poetry scenario videos are provided as follows:

1) https://civitai.com/models/16376/comi-noir-classic-v2;
2) https://civitai.com/models/43331/majicmix-realistic;
3) https://civitai.com/models/64228/linexdiffusion;
4) https://civitai.com/models/25086/locs-China-landscapes-v2;
5) https://civitai.com/models/7371/rev-animated;
6) https://civitai.com/models/27259/tmnd-mix;
7) https://civitai.com/models/73305/zyd232s-ink-style.

C. Prompts for Fig. 4

The prompts for the reference frame are as follows.

Positive Prompt: {{Extremely detailed CG}}, ((Masterpiece)), ((Best quality)), 32K UHD, Epic composition, Chinese ancient landscape style, under the radiant sun, a flock of egrets, painted in vibrant blues and whites, are soaring freely. They frolic in the clear sky in front of a mountain, rendered in deep green tones. The scene is set in a bright and clear lighting, capturing the essence of a beautiful day. This masterpiece is

inspired by the art style of Baishi Qi, known for his Chinese ink paintings with a rice paper texture. The artwork is a landscape type, capturing the interaction of wildlife with their natural environment. The color scheme is bright and vibrant, reflecting the energy of the scene. Rendered in high-definition 2-D, this piece captivates with its intricate details. The quality of the artwork is impeccable, presented in a stunning 8K resolution.

Negative prompt: Low quality, logo, text, watermark, username, Chinese text, stamp mark, cropped image, out of frame, draft, twisted fingers, double image, long neck, multiple heads, ugly, poorly drawn hands, missing limb, disfigured, cut off, grain, low-res, deformed, blurry, bad anatomy, disfigured, mutation, mutated, floating limbs, disconnected limbs, long body, disgusting, poorly drawn, mangled, surreal, extra fingers, duplicate artifacts, morbid, gross proportions, malformed, ugly, tiling, poorly drawn hands, poorly drawn feet, poorly drawn face, body out of frame, bad anatomy, watermark, signature, cut off, low contrast, underexposed, overexposed, bad art, beginner, amateur, draft, grainy.

The Positive prompts for the first control frame are as follows:

"The egret is flying towards the camera."

The Positive prompts for the second control frame are as follows:

"The egret is flying towards very far away mountains."

D. Knowledge Assessment Section of the Questionnaire (Translated to English)

1) This poem depicts an image ().
   a) Image of a rainy scene in a water town
   b) Image of fishing with Jiangnan characteristics
   c) Image of fishing with Saibei style

   Correct answer: B
2) In the initial two verses of the poem, the geographical locale depicted is _____, the temporal setting is and the scenery portrayed comprises _____, _____, _____, _____, _____.

   Correct answer: Mount Xisai, spring, Mount Xisai, egrets, peach blossoms, flowing water, mandarin fish
3) Hues such as _____, _____, _____, _____ are prevalent throughout the entire poem.

   Correct answer: white, pink, cyan, green
4) The entire poem articulates the lifestyle fascination of a fisherman.

   Correct answer example: leisurely
5) In the poem, "The fisherman does not need to return despite the slanting wind and drizzle," what is the literal interpretation of "does not need to return," and what is the concealed rationale behind it?

Correct answer example: The rainfall was moderate, the fisherman donned a raincoat, and the fish inhabiting the river were notably plump; the author no longer wishes to engage in the political turmoil within the court.

ACKNOWLEDGMENT

The authors would like to thank the contributors of the Stable Diffusion, GPT, and RIFE.

REFERENCES

[1] R. Bommasani et al., "On the opportunities and risks of foundation models," Jul. 2022, Accessed: Feb. 17, 2023. [Online]. Available: http://arxiv.org/abs/2108.07258
[2] W. X. Zhao et al., "A survey of large language models," May 2023, Accessed: May 29, 2023. [Online]. Available: http://arxiv.org/abs/2303.18223
[3] The New York Times, "Don't ban ChatGPT in schools. Teach with it," Accessed: Oct. 23, 2023. [Online]. Available: https://www.nytimes.com/2023/01/12/technology/chatgpt-schools-teachers.html
[4] A. Tlili et al., "What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education," *Smart Learn. Environ.*, vol. 10, no. 1, Feb. 2023, Art. no. 15, doi: 10.1186/s40561-023-00237-x.
[5] SlideUpLift, "How to use ChatGPT to make a PowerPoint presentation?" Accessed: Oct. 23, 2023. [Online]. Available: https://slideuplift.com/blog/how-to-use-chatgpt-to-make-a-presentation/
[6] J. Song, B. Wang, Z. Wang, and D. K.-M. Yip, "From expanded cinema to extended reality: How AI can expand and extend cinematic experiences," in *Proc. 16th Int. Symp. Vis. Inf. Commun. Interaction*, 2023, pp. 1–5. [Online]. Available: https://api.semanticscholar.org/CorpusID:264350330
[7] R. Dijkstra, Z. Genc, S. Kayal, and J. Kamps, "Reading comprehension quiz generation using generative pre-trained transformers," in *Proc. 4th Int. Workshop Intell. Textbooks*, 2022, pp. 4–17.
[8] Teacher Joe, "How to use ChatGPT to improve speaking and writing in English." [Online]. Available: https://www.teacher-joe.com/chat-gpt-for-language-learning/how-to-use-chatgpt-to-improve-speaking-and-writing-in-english
[9] A. Olga et al., "Generative AI: Implications and applications for education," 2023, *arXiv.2305.07605*.
[10] R. Mayer, *Multimedia Learning*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 2020, doi: 10.1017/9781316941355.
[11] R. Low and J. Sweller, "The modality principle in multimedia learning," in *The Cambridge Handbook of Multimedia Learning*, 2nd ed., R. E. Mayer, Ed., Cambridge, U.K.: Cambridge Univ. Press, 2014, pp. 227–246, doi: 10.1017/CBO9781139547369.012.
[12] J. Bryant, C. Heitz, S. Sanghvi, and D. Wagle, "How artificial intelligence will impact K-12 teachers," McKinsey, Jan. 2020. [Online]. Available: https://www.mckinsey.com/industries/education/our-insights/how-artificial-intelligence-will-impact-k-12-teachers
[13] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Am. Ch. Assoc. Comput. Linguistics: Human Lang. Technol.,* in Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/n19-1423.
[15] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018.
[16] T. B. Brown et al., "Language models are few-shot learners," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, Jul. 2020, pp. 1877–1901, Accessed: May 26, 2023. [Online]. Available: http://arxiv.org/abs/2005.14165
[17] L. Ouyang et al., "Training language models to follow instructions with human feedback," Mar. 2022, Accessed: May 24, 2023. [Online]. Available: http://arxiv.org/abs/2203.02155
[18] H. Touvron et al., "LLaMA: Open and efficient foundation language models," Feb. 2023, Accessed: May 26, 2023. [Online]. Available: http://arxiv.org/abs/2302.13971
[19] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," Oct. 2022, Accessed: May 26, 2023. [Online]. Available: http://arxiv.org/abs/2204.02311
[20] OpenAI, "GPT-4 technical report," Mar. 2023, Accessed: May 24, 2023. [Online]. Available: http://arxiv.org/abs/2303.08774
[21] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.
[22] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Representations*, 2014, Accessed: May 26, 2023. [Online]. Available: http://arxiv.org/abs/1312.6114

[23] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., New York, NY, USA: Curran Associates, 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[24] A. Q. Nichol, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proc. 39th Int. Conf. Mach. Learn.*, vol. 162, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., Jul. 2022, pp. 16784–16804. [Online]. Available: https://proceedings.mlr.press/v162/nichol22a.html

[25] A. Radford et al., "Learning transferable visual models from natural language supervision," Feb. 2021, Accessed: May 26, 2023. [Online]. Available: http://arxiv.org/abs/2103.00020

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10674–10685, doi: 10.1109/CVPR52688.2022.01042.

[27] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," Jul. 2020, Accessed: May 26, 2023. [Online]. Available: http://arxiv.org/abs/1910.10683

[28] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," May 2022, Accessed: May 26, 2023. [Online]. Available: http://arxiv.org/abs/2205.11487

[29] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," Apr. 2022, Accessed: May 26, 2023. [Online]. Available: http://arxiv.org/abs/2204.06125

[30] R. Hasan, S. Palaniappan, S. Mahmood, A. Abbas, K. U. Sarker, and M. U. Sattar, "Predicting student performance in higher educational institutions using video learning analytics and data mining techniques," *Appl. Sci.*, vol. 10, no. 11, Jun. 2020, Art. no. 3894, doi: 10.3390/app10113894.

[31] A. A. Mubarak, H. Cao, and S. A. M. Ahmed, "Predictive learning analytics using deep learning model in MOOCs' courses videos," *Educ. Inf. Technol.*, vol. 26, no. 1, pp. 371–392, Jan. 2021, doi: 10.1007/s10639-020-10273-6.

[32] T. Crow, A. Luxton-Reilly, and B. Wuensche, "Intelligent tutoring systems for programming education: A systematic review," in *Proc. 20th Australas. Comput. Educ. Conf.*, 2018, pp. 53–62, doi: 10.1145/3160489.3160492.

[33] OpenAI, "Khan academy explores the potential for GPT-4 in a limited pilot program," Mar. 4, 2023. [Online]. Available: https://openai.com/customer-stories/khan-academy

[34] OpenAI, "GPT-4 deepens the conversation on Duolingo," Mar. 4, 2023. [Online]. Available: https://openai.com/customer-stories/duolingo

[35] W. Dai et al., "Can large language models provide feedback to students? A case study on ChatGPT," *2023 IEEE Int. Conf. Adv. Learn. Technol.*, Apr. 2023, doi: 10.35542/osf.io/hcgzj.

[36] Z. Li and Y. Xu, "Designing a realistic peer-like embodied conversational agent for supporting children storytelling," May 2023, Accessed: May 24, 2023. [Online]. Available: http://arxiv.org/abs/2304.09399

[37] O. "Oz" Buruk, "Academic writing with GPT-3.5: Reflections on practices, efficacy and transparency," in *Proc. 26th Int. Acad. Mindtrek Conf.*, Oct. 2023, pp. 144–153, doi: 10.31224/2861.

[38] R. Abdelghani et al., "GPT-3-driven pedagogical agents for training children's curious question-asking skills," Mar. 2023, Accessed: May 26, 2023. [Online]. Available: http://arxiv.org/abs/2211.14228

[39] iFLYTEK, "Summary of the iFLYTEK Xinghuo large model Q&A," May 8, 2023. [Online]. Available: http://www.iflytek.com/news/2651

[40] R. E. Mayer, "Cognitive theory of multimedia learning," in *The Cambridge Handbook of Multimedia Learning*, 2nd ed., R. E. Mayer, Ed., Cambridge, U.K.: Cambridge Univ. Press, 2014, pp. 43–71, doi: 10.1017/CBO9781139547369.005.

[41] S. Kalyuga, "Cognitive load theory: How many types of load does it really need?," *Educ. Psychol. Rev.*, vol. 23, no. 1, pp. 1–19, Mar. 2011, doi: 10.1007/s10648-010-9150-7.

[42] A. Baddeley, "Working memory," *Science*, vol. 255, no. 5044, pp. 556–559, Jan. 1992, doi: 10.1126/science.1736359.

[43] F. Paas and J. Sweller, "Implications of cognitive load theory for multimedia learning," in *The Cambridge Handbook of Multimedia Learning*, 2nd ed., R. E. Mayer, Ed., Cambridge, U.K.: Cambridge Univ. Press, 2014, pp. 27–42, doi: 10.1017/CBO9781139547369.004.

[44] W. Leahy and J. Sweller, "Cognitive load theory and the effects of transient information on the modality effect," *Instructional Sci.*, vol. 44, no. 1, pp. 107–123, Feb. 2016, doi: 10.1007/s11251-015-9362-9.

[45] B. Liu, J. Fu, M. P. Kato, and M. Yoshikawa, "Beyond narrative description: Generating poetry from images by multi-adversarial training," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 783–791, doi: 10.1145/3240508.3240587.

[46] B. Wang, R. Hu, and L. Yang, "Constructing the image graph of tang poetry," in *Natural Language Processing and Chinese Computing* (Lecture Notes in Computer Science, vol. 11839), J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, Eds., Berlin, Germany: Springer, 2019, pp. 426–434, doi: 10.1007/978-3-030-32236-6_38.

[47] L. Xu, L. Jiang, C. Qin, Z. Wang, and D. Du, "How images inspire poems: Generating classical Chinese poetry from images with memory networks," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. 5618–5625, Apr. 2018, doi: 10.1609/aaai.v32i1.12001.

[48] D. Li et al., "Paint4Poem: A dataset for artistic visualization of classical Chinese poems," Sep. 2021, Accessed: May 26, 2023. [Online]. Available: http://arxiv.org/abs/2109.11682

[49] T. Xu et al., "AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1316–1324, doi: 10.1109/CVPR.2018.00143.

[50] T. Qiao, J. Zhang, D. Xu, and D. Tao, "MirrorGAN: Learning text-to-image generation by redescription," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1505–1514, doi: 10.1109/CVPR.2019.00160.

[51] U. Singer et al., "Make-a-video: Text-to-video generation without text-video data," 2022, *arXiv:2209.14792*.

[52] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," Feb. 2023, Accessed: Jun. 12, 2023. [Online]. Available: http://arxiv.org/abs/2302.05543

[53] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-time intermediate flow estimation for video frame interpolation," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 624–642, doi: 10.1007/978-3-031-19781-9_36.

[54] N. Li, Y. Liu, Y. Wu, S. Liu, S. Zhao, and M. Liu, "RobuTrans: A robust transformer-based text-to-speech model," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8228–8235.

[55] J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 17022–17033.

[56] Zulko, "Moviepy," *GitHub Repository*, GitHub, 2020. [Online]. Available: https://github.com/Zulko/moviepy

[57] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in Neural Information Processing Systems*, I. Guyon, Eds., New York. NY, USA: Curran Associates, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

[58] W. Wang et al., "CogVLM: Visual expert for large language models," 2023, *arXiv:abs/2311.03079*.

[59] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," Aug. 2019, Accessed: Nov. 11, 2023. [Online]. Available: http://arxiv.org/abs/1908.10084

[60] OpenAI, "Dall·e 2," Accessed: Jun. 15, 2023. [Online]. Available: https://openai.com/dall-e-2

[61] Midjourney, "Midjourney," Accessed: Jun. 15, 2023. [Online]. Available: https://www.midjourney.com/home/

[62] Baidu, "Wen xin yi ge," Accessed: Jun. 15, 2023. [Online]. Available: https://yige.baidu.com/?source=33257731

[63] Alibaba DAMO Academy for Discovery, "cv_diffusion_text-to-image-synthesis," Accessed: Jun. 15, 2023. [Online]. Available: https://www.modelscope.cn/models/damo/cv_diffusion_text-to-image-synthesis/summary

[64] N. Zulkarnain, H. Prabowo, F. L. Gaol, and S. M. Isa, "Video quality indicators for video-based learning system in higher education," in *Proc. 9th Int. Conf. Front. Educ. Technol.*, 2023, pp. 24–27. [Online]. Available: https://api.semanticscholar.org/CorpusID:260900705

[65] G. Zhao, B. Wang, J. Liu, and J. Wang, "Microlessons in Chinese universities: Concepts, technology, and case analyses," in *Proc. Int. Conf. Blended Learn.*, 2016, pp. 73–84. [Online]. Available: https://api.semanticscholar.org/CorpusID:19096606

[66] Nateraw, "Stable-diffusion-videos," *GitHub Repository*. GitHub, 2023. [Online]. Available: https://github.com/nateraw/stable-diffusion-videos

[67] Deforum, "Deforum-stable-diffusion," *GitHub Repository*, GitHub, 2023. [Online]. Available: https://github.com/deforum-art/deforum-stable-diffusion

[68] Ø. Anmarkrud, A. Andresen, and I. Bråten, "Cognitive load and working memory in multimedia learning: Conceptual and measurement issues," *Educ. Psychol.*, vol. 54, no. 2, pp. 61–83, Apr. 2019, doi: 10.1080/00461520.2018.1554484.

[69] F. Chen et al., *Robust Multimodal Cognitive Load Measurement* (Human–Computer Interaction Ser.). Berlin, Germany: Springer, 2016, doi: 10.1007/978-3-319-31700-7.

Xu Chen received the M.S. degree in cartography and geographical information engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2012 and 2021, respectively.

She is currently an Assistant Research Fellow with the Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan. Her major research interests include 3-D virtual learning resource reconstruction and educational AI agents.

Di Wu received the B.S. degree in computer science and technology and the Ph.D. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2000 and 2006, respectively.

He is a Professor and a Doctoral Advisor with the Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan. He is an ISO/IEC JTC1 SC36 WG6 Co-Convener and Project Editor. His main research interest focuses on educational information technology standards and applications.