Digital Human Technology in E-Learning: Custom Content Solutions

Chen, Sinan ; Yang, Liuyi ; Zhang, Yue ; Zhang, Miao ; Xie, Yangmei ; Zhu, Zhiyi ; Li, Jialong

(Citation)
Applied Sciences,15(7):3807

(Issue Date)
2025-04

(Resource Type)
journal article

(Version)
Version of Record

applied sciences

MDPI

Article

Digital Human Technology in E-Learning: Custom Content Solutions

Sinan Chen [1,2,*], Liuyi Yang [3,*], Yue Zhang [2], Miao Zhang [4], Yangmei Xie [5], Zhiyi Zhu [3] and Jialong Li [6]

[1] Center of Mathematical and Data Sciences, Kobe University, 1-1 Rokkodai-cho, Nada, Kobe 657-8501, Japan
[2] Graduate School of Engineering, Faculty of Engineering, Kobe University, 1-1 Rokkodai-cho, Nada, Kobe 657-8501, Japan; 228t802t@gsuite.kobe-u.ac.jp
[3] Graduate School of System Informatics, Kobe University, 1-1 Rokkodai-cho, Nada, Kobe 657-8501, Japan; 194x606x@gsuite.kobe-u.ac.jp
[4] Lab of Sediment Hazards and Disaster Risk, Graduate School of Maritime Sciences, Kobe University, 5-1-1 Fukae Minamicho, Higashinada-ku, Kobe 658-0022, Japan; 218w402w@gsuite.kobe-u.ac.jp
[5] Graduate School of Human Development and Environment, Kobe University, 1-1 Rokkodai-cho, Nada, Kobe 657-8501, Japan; 247d605d@gsuite.kobe-u.ac.jp
[6] Department of Computer Science and Engineering, Waseda University, 1-104 Totsukamachi, Shinjuku-ku, Tokyo 169-8050, Japan; lijialong@fuji.waseda.jp
* Correspondence: chensinan@gold.kobe-u.ac.jp (S.C.); 211x508x@gsuite.kobe-u.ac.jp (L.Y.)

Abstract: With advances in digital transformation (DX) in education and digital technologies becoming more deeply integrated into educational settings, global demand for video-based learning materials continues to rise, resulting in substantial effort being required from teachers to create e-learning videos. Furthermore, while many existing services offer visual content, they primarily rely on templates, making it challenging to design custom content that addresses specific needs. In this study, we develop a web service that facilitates e-learning video creation through integrated artificial intelligence (AI) and digital human technology. This service enhances educational content by integrating digital human characters and voice synthesis technologies, aiming to create comprehensive e-learning videos by incorporating visual motion and synchronized audio into educational content. In addition, this service also aims to enable the creation of engaging content through advanced visuals and animations, effectively maintaining learner interest.

Keywords: E-Learning; digital human; educational content; web service

Academic Editor: Martin Ebner

Received: 18 February 2025
Revised: 22 March 2025
Accepted: 27 March 2025
Published: 31 March 2025

Citation: Chen, S.; Yang, L.; Zhang, Y.; Zhang, M.; Xie, Y.; Zhu, Z.; Li, J. Digital Human Technology in E-Learning: Custom Content Solutions. *Appl. Sci.* **2025**, *15*, 3807. https://doi.org/10.3390/app15073807

1. Introduction

The education sector is undergoing a rapid digital transformation (DX) in which the adoption of digital technologies enhances instructional efficiency and broadens access to educational resources [1]. To advance digital education, the design of tools must consider the perspectives and requirements of teachers and students. Tools that prioritize efficiency and address individual learner needs can foster a more seamless integration of digital technologies into educational settings. Studies have demonstrated a positive correlation between high academic performance and the use of digital technologies, which are particularly suited to addressing students' personalized learning requirements [2]. Against this backdrop, video-based instructional materials have emerged as an essential component of education due to their inherent flexibility and the intuitive, visual manner in which they present content. Moreover, digital education can reduce the workload of teaching staff while enhancing the quality of instruction through greater individualization and clearer educational objectives [3]. The influence of digital education now extends beyond

traditional subjects such as language and programming, reaching into specialized areas like medicine [4].

Despite their growing importance, current processes for producing instructional videos exhibit several notable shortcomings. These issues include excessive dependence on educators and production teams and overall reliance on time-consuming and labor-intensive workflows. Moreover, most existing tools employ rigid templates, making it challenging to meet the diverse requirements of various teaching scenarios, resulting in limiting personalization and innovative potential [5]. Accordingly, numerous studies highlight the potential of emerging digital technologies in education, but their practical implementation often depends on the guidance and involvement of educators or supervisors. In addition, negative attitudes among some teachers toward new educational paradigms can impede the effective integration of these technologies [6].

Artificial intelligence (AI) agents, autonomous systems using natural language processing, machine learning, and computer vision, have revolutionized various industries through their ability to interpret, decide, and respond to environmental inputs [7]. In educational contexts, these agents show promise for personalizing learning experiences and supporting teaching tasks. Despite this potential, developing fully autonomous educational AI agents requires addressing several challenges, particularly in balancing the diverse needs of teachers and students [8]. Nevertheless, there are no effective and comprehensive AI-agent solutions in education within current research. It is necessary to investigate the application scenarios, adoptions, and pedagogical considerations of digital learning tools in education to comprehensively understand the effective integration of such AI-driven systems into education [9].

The digitalization of education is already underway, and the potential value of AI technologies in educational services is well recognized. In this study, we focus on utilizing AI agents to streamline the process of producing e-learning videos and enable AI technologies to play a pivotal role in advancing efficiency and personalization in education. Our objective is to design and develop an AI-driven web service that caters to a broader range of learner needs and enhances the overall learning experience. This service autonomously generates comprehensive, high-quality e-learning videos with educational content.

Therefore, the main contributions of our paper are as follows:

- Development of an AI-driven digital education platform that automates the generation of microlecture videos;
- Integration of style-diverse digital human teachers capable of delivering customized course content presentations.

In the rest of the article, Section 2 describes the existing work and the technical challenges; Section 3 describes the proposed methodology; Section 4 describes the details of the methodology implementation; Section 5 introduces a use case; Section 6 describes the evaluation and discussion of the proposed framework; Section 7 summarizes the article.

2. Related Work

2.1. Preliminaries

Considerable research has investigated various dimensions of digital learning across subjects and educational levels. For example, studies have examined the impact of text-to-speech (TTS) technologies on enhancing learning opportunities and academic performance of students with disabilities, illustrating how emerging tools can significantly improve learning experiences [10]. At the same time, digital education research has expanded from basic K-12 instruction to specialized domains, including science, technology, engineering, arts, and mathematics, where it informs conceptual understanding and dialogue-based learning strategies that promote deep engagement and critical thinking skills [11,12].

Multimodal learning in DX education leverages diverse instructional stimuli, learning environments, and data sources to create more engaging and personalized learning experiences. According to Luo (2023) [13], advancements in digital technologies and analytics have significantly enhanced the integration of multimodal approaches, allowing for a more holistic understanding of learners' needs and behaviors. AI-driven digital human teaching systems are advancing education by utilizing multimodal approaches and replicating human teaching techniques. Immadisetty et al. (2023) [14] emphasize the importance of integrating cues like posture, gestures, facial expressions, and verbal interactions to enhance affect recognition and improve student engagement in online classrooms. Similarly, Mulian et al. (2023) [15] explore the efficacy of virtual AI teachers in fine motor skill training, demonstrating their ability to emulate human instructors and effectively support skill acquisition through precise feedback and guidance.

In addition, research has explored the integration of AI-driven learning platforms that foster collaboration and interaction between students and educators [16]. This body of work underscores the importance of aligning technological development with pedagogical goals, ensuring that tools support cognitive, emotional, and social dimensions of learning.

Building on these findings, the present study aims to develop digital educational tools tailored for teachers and students alike. By leveraging adaptive, personality-sensitive approaches, these tools aim not only to reduce the burden on educators but also to meet the diverse needs of learners. Ultimately, this research seeks to create an educational experience that is both dynamic and efficient, advancing the vision set forth in the introduction by providing a concrete pathway for the practical implementation and utilization of AI-driven digital human technologies in education.

2.2. Comparison with Commercial Tools

In comparison with commercial AI-powered e-learning video generation tools such as Elai.io [17], Vyond [18], and AI Studios [19], our proposed method demonstrates distinct advantages in automation, customization, visual presentation, and user experience, as shown in Table 1. Unlike Vyond, which relies heavily on templates and manual adjustments, our method offers full automation with a modular design, significantly reducing the effort required for content creation. While both Elai.io and AI Studios offer high customization capabilities, our system further excels by supporting multiple presentation styles and multilanguage content generation. Additionally, our method ensures natural and fluent virtual lecturer animations with precise lip synchronization, outperforming the somewhat rigid lip movements observed in Elai.io. In terms of user experience, our platform features a user-friendly interface, making it more accessible compared to Vyond's complex operation.

Table 1. Comparison with commercial tools.

Comparison Dimension	Our Method	Elai.io	Vyond	AI Studios
Automation Level	High: Fully automated process with flexible modular design	Medium: Semi-automated, requires manual configuration of some parameters	Low: Highly template-driven, requiring significant manual adjustments	Medium: High level of automation but limited customization

Table 1. *Cont.*

Comparison Dimension	Our Method	Elai.io	Vyond	AI Studios
Customization Ability	High: Supports multiple styles (formal, casual, key points) and multi-language support	High: Supports personalized digital human characters and emotional voice expression	Medium: Fixed styles with limited customization	High: Can choose virtual character appearance and emotional voice expression
Visual Presentation	High: Natural and fluent virtual lecturer movements with high lip synchronization accuracy	Medium: Good virtual human effect, but lip synchronization is somewhat rigid	Low: Simple animations with limited virtual human capabilities	High: Good synchronization between animation and speech with realistic virtual humans
User Experience	User-friendly: Simple operation with clear interface	User-friendly: Simple operation and intuitive interface	Average: Complex operations with a high learning curve	User-friendly: Fast generation with diverse customization options

2.3. Technical Challenges

2.3.1. Cross-Modal Generation and Temporal Synchronization

Producing microlecture videos requires the simultaneous processing of multiple data modalities, including text, speech, images, and motion. Ensuring that these data are both contextually and temporally aligned is critical for maintaining the coherence and authenticity of the digital human's performance. As highlighted in a study by Haji-Ali et al. [20], achieving temporal alignment across modalities is essential for generating synchronized and high-quality audiovisual content. Achieving this goal demands highly efficient algorithms capable of coordinating diverse data streams and precisely correcting any delays or asynchronies that may arise.

2.3.2. Emotional Expression in Synthetic Speech

Incorporating emotional expression into synthetic speech is essential for creating engaging and effective teaching experiences. Recent advancements in deep learning have enabled speech synthesis systems to convey a wide range of emotions, enhancing the adaptability of digital educators to various instructional contexts. For instance, a study by Tits et al. (2021) [21] introduced a methodology for controlling the expressiveness of synthetic speech via visualization and interpretation of a learned latent space, allowing for nuanced emotional modulation in educational content. Additionally, maintaining high audio quality while minimizing computational complexity remains a priority to support real-time applications in educational settings.

2.3.3. Automated Customization of Educational Content

The educational needs of different courses and learners can vary significantly, making it essential for the system to automatically analyze teaching materials, understand semantic content, and generate e-learning videos that meet specific learning objectives. This capability involves automated customization of PowerPoint slides, textual elements, and teaching resources featured in the video. By analyzing the requirements of the audience, the system can automatically produce diversified instructional content tailored to specific learning goals and preferences. Achieving this level of personalization poses significant challenges,

as it requires advanced algorithms capable of semantic understanding, context recognition, and adaptive content generation across diverse learning scenarios [22].

2.4. Theoretical Foundations

To better understand how the proposed system affects learning outcomes, we anchor our approach in Mayer's cognitive theory of multimedia learning [23]. This theory suggests that meaningful learning occurs when learners can build connections between verbal and visual representations. Features such as high-quality synthesized audio and accurate lip synchronization align with the modality principle and temporal contiguity principle, which state that learning is enhanced when words and pictures are presented simultaneously rather than separately. Our system's design—especially its focus on synchronized speech and expressive digital humans—aims to reduce extraneous cognitive load and increase engagement, which are key factors in effective multimedia learning.

3. Methodology

This section presents the methodology employed to develop a system that integrates educational content with digital human technology. Our proposed framework is structured into three distinct stages: content input, multimodal processing, and digital human animation.

3.1. Step 1: Content Input

The first step involves the input of educational content, which may include text, audio, video, or multimodal data. The system is designed to accommodate a wide variety of educational materials, ensuring flexibility and adaptability across different subject domains. The user uploads the content through a user-friendly interface. Text content, for instance, can include lecture notes or scripted dialogues, while multimodal data might involve presentations containing both text and images.

3.2. Step 2: Multimodal Processing

Once the content is inputted, it undergoes a series of processing steps facilitated by specialized modules:

Step 2-1: Text Processing

The text processing module leverages natural language processing (NLP) techniques to analyze, structure, and enhance the input text. Key steps include:

- Tokenization and Parsing—Breaking the input text into meaningful units for easier manipulation;
- Semantic Analysis—Ensuring the coherence of the content by examining its semantic structure, enabling the generation of appropriate speech and visuals;
- Customization—Incorporating user-defined preferences such as tone and language style.

Step 2-2: Speech Synthesis

Using state-of-the-art text-to-speech (TTS) technology, the processed text is transformed into high-quality, human-like speech. Deep learning models, such as Tacotron or WaveNet, are employed to achieve natural intonation and rhythm.

Step 2-3: Image and Scene Generation

Complementary visuals are generated to enrich the educational content. For instance, diagrams, animations, or relevant background scenes are created based on the input text's context. Advanced computer vision and generative models, such as Stable Diffusion, are utilized for this purpose.

3.3. Step 3: Digital Human Animation

The final stage integrates the processed content to generate a synchronized digital human animation. This step combines the following components:

Step 3-1: Facial Animation

The digital human animation of the digital human is generated frame by frame using generative AI. The model takes a reference facial photograph and the previously generated speech as inputs, and uses the speech features to drive the digital human's movements.

Step 3-2: Body Gestures

Body movements and gestures are generated based on the contextual cues derived from the input text. Machine learning models trained on datasets of human movements ensure natural and expressive animations.

Step 3-3: Integration and Rendering

All elements, including speech, visuals, and animations, are seamlessly integrated into a single video output. Rendering engines optimize the final output for smooth playback, ensuring that the video is suitable for diverse educational settings.

3.4. Step 4: Evaluation

To ensure the effectiveness of the system, iterative testing and refinement are performed. User feedback is collected to assess:

- The clarity and engagement level of the digital human;
- The accuracy of synchronization between speech and animation;
- The educational value of the content delivery.

Based on the feedback, the underlying models and algorithms are continuously optimized to enhance performance and usability. This iterative cycle ensures that the system remains at the forefront of digital education technology. By following this methodology, the system provides an innovative approach to delivering educational content, leveraging advanced AI technologies to create engaging and impactful learning experiences.

Key principles informing this approach include:

- Usability: Ensuring that the system is accessible to teachers with varying technical skills;
- Personalization: Tailoring content and presentation styles to meet the needs of different learners;
- Efficiency: Reducing the time and effort required to produce engaging instructional videos.

The overall framework is as shown in Figure 1. First, the user inputs educational content (such as text or multimodal data); next, the content is processed through modules such as text processing, speech synthesis, and image generation; then, a digital human animation synchronized with the speech is generated and integrated into a complete video. An iterative evaluation loop finally analyzes user feedback to optimize the generated content.

3.5. Discussion

This methodology exhibits several notable advantages while also presenting certain limitations. A primary strength lies in its modular design, in which content generation, speech synthesis, and animation are distinctly separated to enhance flexibility and scalability. Nevertheless, potential enhancements can be pursued to further improve the system's overall efficacy.

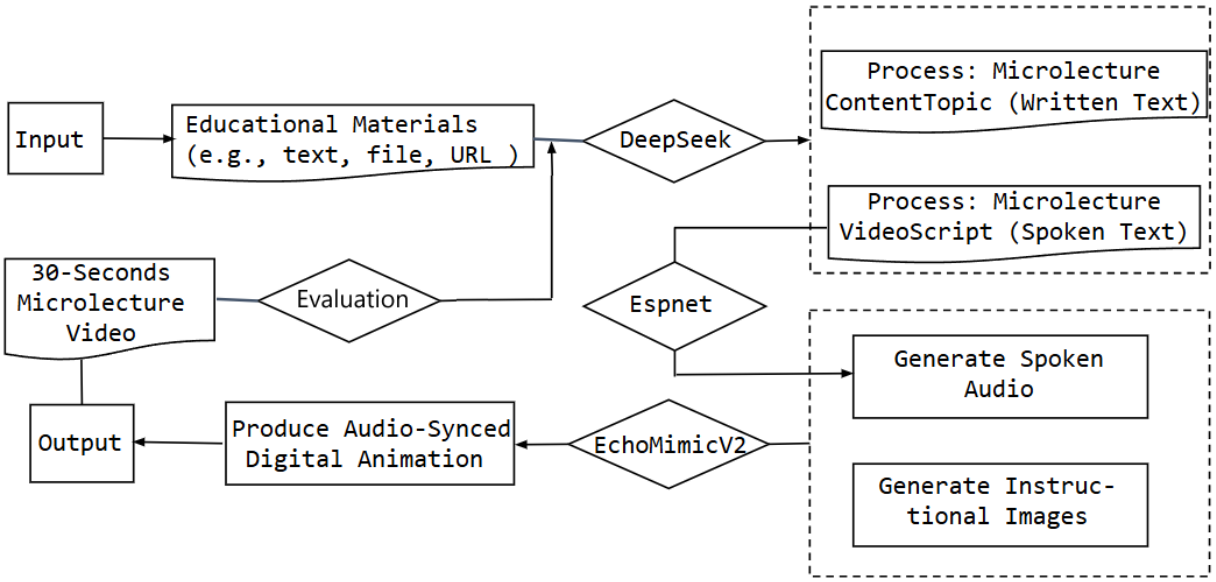

Figure 1. The overall flowchart of proposed method.

For instance, integrating adaptive learning algorithms could enable real-time adjustments to the generated content, thereby refining alignment with learner feedback and engagement metrics. Likewise, leveraging lightweight models or hybrid approaches that combine AI-driven automation with human oversight may strike a more balanced compromise between efficiency and quality, thus delivering more accurate and specialized educational content.

4. Implementation

4.1. Microlecture Generation

We used Deepseek [24] to generate the content. The process begins with inputting a piece of material, which the generative pre-trained transformer (GPT) summarizes. Then, the summarized content is formatted into a general-to-specific structure and exported as an image. In this study, we chose to generate video segments approximately 30 s long. This length was selected as a practical balance between our current hardware capabilities and the volume of data needed for evaluation. However, it is important to note that the system itself can generate videos that are several minutes long and supports concatenation of multiple segments. The generation time is closely tied to hardware performance, and longer videos require proportionally more processing resources. In this case, the material used focuses on the life of Newton [25].

4.1.1. Written Text Generation

To enable GPT to generate summaries more efficiently, users can assign it a specific role tailored to particular needs. This approach helps GPT better understand your expectations, such as professionalism, depth, and the intended target audience. For example, in our experiment, we first specified the role of GPT as follows: "You are a language education expert who helps learners distill the main content of textbooks or articles". Next, we instructed GPT to produce summaries in three distinct styles: formal, casual, and key points, as illustrated in Listing 1.

Listing 1. Summarization prompts for style variations.

```
summarization_prompts = [
    {
        "style_name": "formal",
        "prompt": "Summarize the content in a formal academic
            style
                with a general-to-specific structure within
                    200 characters in English."
    },
    {
        "style_name": "casual",
        "prompt": "Summarize the content in a casual tone
                with a general-to-specific structure within
                    200 characters in English."
    },
    {
        "style_name": "keypoints",
        "prompt": "Summarize the content with key points
            highlighted,
                with a general-to-specific structure within
                    200 characters in English."
    }
]
```

4.1.2. Spoken Text Generation

Next, we generate spoken content based on the summary. At this stage, we prompt GPT with the instruction: "Please rewrite the following summary in a spoken, storytelling style, in English".

4.1.3. Content Structuring

Next, we formatted the summaries in the three styles, providing GPT with specific instructions for structuring the output, as shown in Listing 2.

It is important to note that these three styles—formal, casual, and key points—are illustrative examples rather than fixed options. The system allows users to define custom summarization styles via prompt engineering. This flexibility supports the personalization of content presentation based on students' preferences and learning needs, enhancing engagement and comprehension.

Listing 2. Formatting guidelines prompt.

```
common_structure = (
    "Please structure the summary as follows:\n"
    "1) One overall statement on the first line.\n"
    "2) Then several bullet points (each under 40 characters)
        with key details.\n"
    "Use point at the start of each bullet."
)
```

4.2. Instructional Images Generation

To enhance the presentation of structured educational content, we implemented a Python (version 3.10)-based method for converting text into images. The text_to_image function dynamically formats input text by wrapping lines, customizing fonts, and adjusting layouts to ensure visual clarity.

4.3. Audio Generation

Transform the prepared script into naturally sounding speech aligned with the instructional content. Based on the script generated in the previous stage, this component produces voice output for the digital instructor. Teachers and students can adjust parameters such as voice tone, pitch, and emotional quality.

By customizing factors such as voice style and mood, the system can create a more appealing and learner-focused auditory experience. This adaptability allows the generated audio to better resonate with diverse learner preferences, thereby increasing engagement and knowledge retention.

Using the spoken content, we employ TTS technology to produce audio output. We evaluated several TTS methods, as summarized in Table 2. Considering factors such as generation speed, output quality, and ease of deployment, we selected Espnet for use in our experiments [26–36].

Table 2. Comparison of different TTS technologies.

Technology	Number of Supported Languages	Generation Speed	Voice Quality
Espnet	5	Fast	High
Mozilla TTS	20	Moderate	High
Fish-Speech	8	Fast	High
Coqui TTS	13	Fast	High
Google TTS	Over 50	Fast	High
VoiceVOX	Only Japanese	Fast	High
Vertex AI Studio	Over 50	Fast	High

4.4. Digital Human Generation

In the digital human generation stage, we tested four open-source digital human models locally. We used the voice data generated in the previous step 2-1, which consists of a 9 s audio clip, to drive the pictures automatically generated by getimg.ai [37]. These pictures, with a resolution of 1024 × 1024, were used to generate a 10 s video. According to the time taken by the four models to generate videos of the same length, resolution, and size, we finally decided to use the EchoMimicV2 model as the digital human generation model in this study.

4.5. Video Generation

Combine presentation materials and the synthesized audio with a digital instructor's animated movements and expressions to create the final video. The final stage involves creating the animated digital instructor who delivers the lesson content. This includes generating facial expressions, head movements, and hand gestures that correspond to the instructional script and highlight key teaching points. Initially, the focus is on animating the digital instructor's upper body, ensuring that facial cues and gestures align with the audio track to create a coherent, lifelike presentation. In future iterations, the approach could be extended to enable dynamic, real-time interactions between the digital instructor and students, further personalizing the learning experience.

4.6. Combination

The final step in producing the e-learning video involves integrating the various components—namely, the slide content, the synthesized speech, and the animated digital instructor—into a cohesive multimedia presentation. First, the text generated for the slides is placed into a presentation format. Simultaneously, the digital human animation script is converted into audio output using TTS technology.

Achieving seamless alignment between the PPT slides, digital human animation, and audio presents several challenges. Timing is critical: each slide transition must match the instructor's vocal cues, while the avatar's gestures and facial expressions should align with the content's emphasis and pacing.

All facial images used in this study were generated synthetically using the text-to-image functionality of getimg.ai. According to the platform's terms of use, such generated images are permitted for academic research purposes. Similarly, all audio data were generated using open-source models provided by ESPnet, without involving any real human voice recordings.

5. Case Study

In this study, we selected the topic of Sir Isaac Newton as a representative example to test and demonstrate the overall workflow of our system. The focus of this paper is primarily on presenting the methodology and technical integration, rather than evaluating subject-specific performance across disciplines. Due to the rapid evolution of AI models and potential variability in generation accuracy, broader testing across multiple domains is planned for future work.

5.1. Different Styles of Microlecture

Our method supports the generation of microlectures in different styles, including casual, formal, and key points summaries. As shown in Figure 2, the three styles demonstrate the system's capacity for stylistic variation through prompt-based control. Meanwhile, Table 3 presents the generated spoken texts in three different styles, which are the contents that the virtual teacher will deliver. For instance, the formal style is suited for academic lectures, the casual style appeals to informal or motivational learning contexts, and the key points style aids quick revision or review. Users can further define new styles depending on their pedagogical goals or learner preferences.

```
Sir Isaac Newton was a genius who changed science forever.
· Formulated laws of motion & gravity
· Co-invented calculus
· Built 1st reflecting telescope
· Worked on optics & alchemy
```

(**a**) Casual summary

```
Sir Isaac Newton was a pivotal figure in the Scientific
Revolution, contributing to mathematics, physics, and
beyond.
· Formulated laws of motion & gravitation.
· Co-developed calculus & optics theories.
· Built first reflecting telescope.
· Influential in alchemy & biblical studies.
```

(**b**) Formal summary

Figure 2. *Cont.*

Sir Isaac Newton was a pivotal figure in the Scientific Revolution and modern science.
- Formulated laws of motion & gravitation
- Established classical mechanics
- Contributed to optics & calculus
- Built first reflecting telescope
- Influential in physics & mathematics

(**c**) Key points summary

Figure 2. Three styles of summarization: casual, formal, and key points.

Table 3. Different styles of generated spoken text.

Style	Spoken Text
Casual	Let me tell you about Isaac Newton, a true giant of science. Picture this: a man who unlocked the secrets of motion and gravity, forever changing how we see the universe. He didn't stop there—Newton co-created calculus, a math so powerful it shapes our world today. Imagine him crafting the first reflecting telescope, peering into the cosmos, and writing masterpieces like "Principia" and "Opticks". His ideas laid the foundation for modern science, influencing everything from physics to optics. Newton wasn't just a scientist; he was a visionary who reshaped our understanding of the natural world.
Formal	Let me tell you the story of Sir Isaac Newton, a brilliant mind who shaped the Scientific Revolution and Enlightenment. He unveiled the laws of motion and universal gravitation, explaining how the universe works. Newton also invented calculus and expanded the binomial theorem, revolutionizing mathematics. Ever curious, he built the first reflecting telescope and explored light in his groundbreaking book "Opticks". His masterpiece, "Principia Mathematica", laid the foundation for classical mechanics. Beyond science, Newton served as Master of the Royal Mint, ensuring the accuracy of British coinage. His legacy? A world forever changed by his genius and relentless pursuit of knowledge.
Key points	Let me tell you the story of Isaac Newton, a man who changed the way we see the world. Born in 1642, Newton was a brilliant scientist and mathematician. He unlocked the secrets of motion and gravity, writing them down in his famous book, "Principia". He also co-created calculus, though he and Leibniz had a bit of a rivalry over it. Newton didn't stop there—he built the first reflecting telescope and explored light in his book "Opticks". Later in life, he even ran the Royal Mint, making British coins more secure. Newton's ideas shaped modern science and still inspire us today.

5.2. Different Digital Human

We evaluated the performance of multiple digital human generation models and selected the EchoMimicV2 model for its superior balance between animation quality and computational efficiency. Figure 3 illustrates examples of digital human animations generated using different platforms, demonstrating the adaptability of our system in creating lifelike, expressive avatars for educational purposes.

(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 3. Digital human generated by different platforms. (**a**) Input image [37]. (**b**) FirstOrder-Model [38]. (**c**) AniPortrait [39]. (**d**) EchoMimicV1 [40]. (**e**) EchoMimicV2 [41].

5.3. Result of Combination

By integrating text summaries, synthesized audio, and digital human animations, our system generates comprehensive e-learning videos. An example of such a combination is shown in Figure 4, highlighting the seamless alignment between the visual, auditory, and instructional components. This integration ensures coherence and enhances learner engagement by delivering content in an intuitive and visually appealing manner.

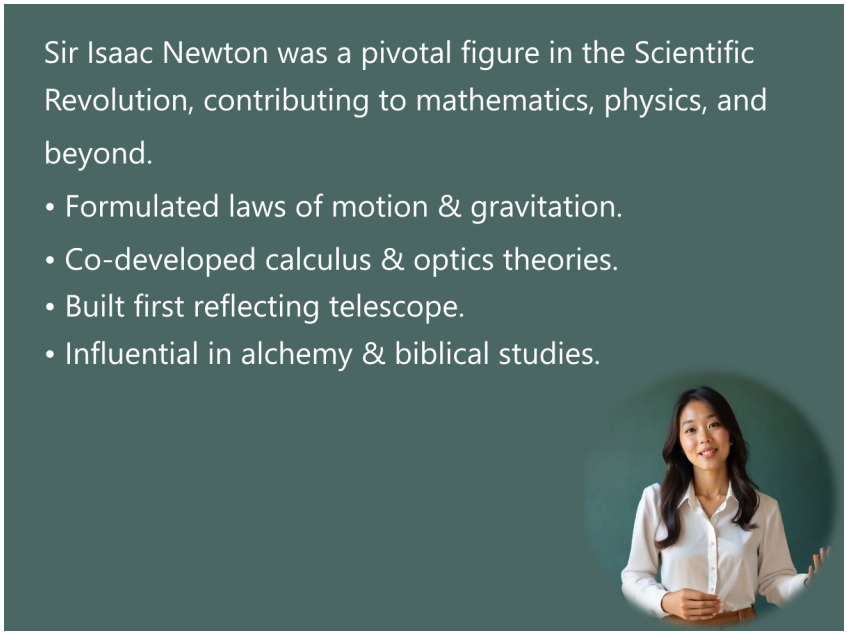

Figure 4. An example of one combination.

6. Evaluation and Discussion

We constructed a comprehensive evaluation framework for the generated e-learning videos, encompassing both objective metrics and a subjective user survey. The objective parameters are assessed from three distinct perspectives: the textual content, the audio quality, and the alignment between video and content.

6.1. Objective Evaluation

6.1.1. Content

In evaluating textual accuracy, we will employ a semantic matching score to verify how closely the generated text aligns with the original teaching materials. Beyond basic fidelity, we will also examine concept contradictions by comparing statements throughout the text, thereby identifying any logical or factual inconsistencies. To ensure a natural progression, we will measure sentence coherence by computing similarity scores between adjacent sentences, revealing whether the content flows smoothly. The calculation methods

and meanings of the three metrics are shown in Table 4. As the test results indicate, the three text styles exhibit comparable performance scores. We employed the semantic matching score to evaluate the coverage of the original text in the generated outputs. Apart from the key point style, which achieved a coverage rate of 85.71%, the other styles fully covered the original content without introducing extraneous information. The concept contradictions test confirms that no internal inconsistencies appear in the generated texts. In addition, all three styles received the same sentence coherence score, indicating fluent and logically connected sentences.

Table 4. Semantic text evaluation metrics.

Metric Name	Calculation Method	Meaning
Semantic Matching Score [42]	Compare the semantics of the original text with the test text, typically using text classification methods to verify whether the test text remains faithful to the original meaning.	Indicates how closely the test text aligns with the source semantics; higher scores suggest stronger consistency.
Concept Contradictions [43]	Divide the test text into multiple sentences and compare each pair to detect semantic contradictions. Identifying contradictory relationships reveals logical conflicts or semantic inconsistencies.	Helps discover potential logical or conceptual conflicts within the test text, thereby improving overall accuracy and coherence.
Sentence Coherence [44]	Extract semantic vectors for each sentence in the test text, compute adjacency (e.g., cosine) similarity between neighboring sentences, and derive an average coherence score.	Evaluates whether sentence-to-sentence logical and semantic transitions are smooth. Higher scores indicate more natural flow and coherence throughout the text.

6.1.2. Audio Quality

We evaluate the generated speech using four metrics: (1) the automatic speech recognition (ASR) score, which measures intelligibility by comparing recognized text with the original text; (2) pitch, which influences the perception of highness or lowness of a tone and is determined by the fundamental frequency of the sound wave; (3) speech rate (also referred to as tempo), which quantifies the rhythm or speed of speech in terms of words or syllables per minute; and (4) short-time energy (STE), which captures temporal variations in speech loudness. Table 5 details these metrics and their computational procedures.

Pitch provides insights into the quality of a sound, aids in differentiating tones and accents, and highlights speaker-specific characteristics [45]. Speech rate directly affects segmentation, word alignment, and overall recognition accuracy; accurate tempo analysis enables ASR systems to adapt to varying speaking styles. Lastly, short-time energy is commonly used to segment a voice signal by detecting word boundaries, and by analyzing energy variations over short time frames, it improves transcription precision [46].

From the results shown in Table 6, the “Casual” style achieves the highest ASR Score, 66.25, whereas the “Formal” and “Keypoints” styles score slightly lower at 62.20 and 61.25, respectively. Since the ASR score, pitch, and short-time energy values are normalized on a 0–100 scale, higher values correspond to better performance. All three styles share the same pitch value, 100, suggesting comparable tonal ranges.

Table 5. Audio quality evaluation metrics.

Parameter Name	Calculation Method	Meaning
ASR Score [47]	Convert the audio into text using an ASR system, then compare the recognized text with the reference script (matched words/total words).	Evaluates how accurately the spoken content conveys the intended meaning (speech intelligibility).
Short-Time Energy [48]	Use librosa.feature.rms to calculate the mean of the short-time energy.	Reflects loudness variation, indicating how dynamic or stable the audio signal is.
Pitch [49]	Use librosa.piptrack to extract pitch data and compute the average pitch value.	Indicates the stability and appropriateness of vocal intonation, contributing to the naturalness of speech.
Tempo [50]	Use librosa.beat.beat_track to estimate beats per minute (BPM).	Represents the speech rate, indicating how fast or slow the narration is delivered.

Tempo, measured in words or syllables per minute, does not have a strictly linear relationship with quality. Typical teacher speaking speeds range from 50 to 200 words per minute, and all three styles fall within this interval. The "Keypoints" style is delivered fastest at 123.05, followed by "Formal" at 117.45, and "Casual" at 86.13. In terms of short-time energy, the "Casual" style exhibits the highest value, 40.445, indicating stronger loudness variation compared to "Formal" at 35.91 and "Keypoints" at 39.73.

Table 6. Evaluation of audio quality.

Evaluation Metric	Casual	Formal	Keypoints	Expected Range
ASR Score	66.25	62.20	61.25	0–100
Pitch	100.00	100.00	100.00	0–100
Tempo	86.13	117.45	123.05	50–200
Short-Time Energy	40.45	35.91	39.73	0–100

6.1.3. Digital Human Performance

To evaluate the alignment between synthesized speech and the digital instructor's on-screen presentation, we will examine lip synchronization accuracy. By quantifying the timing offset between the audio and the avatar's mouth movements, we can ensure that the digital human component appears natural, coherent, and closely tied to the narrated content.

We employed the evaluation module from the Wav2Lip method to generate two metrics [51], LSE-D and LSE-C, for each of the three video styles. Specifically, LSE-D (lip sync error distance) quantifies the temporal and spatial discrepancy between the synthesized lip movements and the corresponding audio, thereby indicating the overall precision of lip synchronization. In contrast, LSE-C (lip sync error confidence) represents the confidence level or coherence of the lip movements relative to the audio track, reflecting the degree of alignment between the visual and auditory components.

As shown in Table 7, the casual style achieves the lowest LSE-D (8.923) and the highest LSE-C (6.650), indicating more precise and confident lip synchronization. In contrast, the formal and key point styles exhibit slightly higher LSE-D values, suggesting a marginal decrease in synchronization accuracy. Their corresponding LSE-C values are also lower than

that of the casual style, reflecting reduced coherence between audio and lip movements. Overall, these findings imply that the casual style may offer a stronger balance of accuracy and confidence in lip sync performance.

Table 7. LSE-D and LSE-C scores for three video styles.

Style	LSE-D	LSE-C
Casual	8.923	6.650
Formal	8.968	5.703
Keypoint	9.059	5.865

6.2. User Survey

The user survey involved 25 participants with diverse educational and professional backgrounds, all above high school level. The questionnaire was designed to evaluate six key dimensions of the generated educational videos: content consistency, expression style, audio quality, visual quality, attraction, and overall experience. For each dimension, participants were asked 6 questions using a 4-point Likert scale, and each parameter has been normalized to a score of 0 to 10. The results, summarized in Table 8, indicate an average overall score of approximately 5.7, reflecting moderate satisfaction among respondents. While content consistency and overall experience received relatively higher ratings, visual quality, and attraction were noted for potential improvement. These findings provide actionable insights for enhancing the teaching materials to meet audience expectations better.

Table 8. Evaluation of survey results.

Question	Average Score
Content Consistency	6.2
Expression Style	5.7
Audio Quality	6.0
Visual Quality	5.5
Attraction	5.3
Overall Experience	5.8

6.3. Discussion

Both the generated audio and digital human animations aligned closely with the instructional content. Notably, the casual style achieved the highest ASR score, likely due to its linguistic simplicity, while the formal and key points styles scored slightly lower. The faster tempo of the key points style may have further reduced ASR accuracy. Additionally, the casual style exhibited the greatest short-time energy, suggesting a dynamic intonation that could enhance learner engagement.

A principal advantage of this system is its capacity to offer varied content delivery styles, enabling educators to tailor instruction to diverse pedagogical requirements. The automated generation process also substantially decreases the effort and time needed to develop e-learning materials.

Nevertheless, certain constraints persist. The generation speed remains suboptimal for high-resolution animations or complex scripts, which may impede real-time implementation. Moreover, customization is currently confined to predefined styles, and reliance on pre-trained models restricts adaptability to specialized domains or underrepresented languages. For contents, current AI systems have limitations in semantic abstraction and hierarchical structuring, particularly when expressing complex hierarchical relationships in a coherent and logical manner. For example, as shown in Figure 2c, “influential in physics

& mathematics" is a general statement, while "contributed to optics & calculus" represents specific subfields. Although optics is a branch of physics, the hierarchical distinction aims to emphasize the progression from overall impact to specific contributions. Future research should focus on optimizing computational efficiency, incorporating adaptive algorithms, and extending language support to broaden the system's global applicability.

7. Conclusions

In this study, the overall implementation process begins with segmenting educational content into suitable video sections to ensure clarity and coherence. Next, AI tools generate character animations and corresponding voiceovers, which are then integrated with the educational content and optimized to produce a polished final video. To support users, our team provides step-by-step instructions detailing key operations, accompanied by screenshots or diagrams illustrating the generation process and tool interfaces. Additionally, we address common challenges encountered during video creation and propose practical solutions to resolve these issues effectively.

In the future, we aim to expand our scope to include web-based services, designing applications across three key dimensions: teachers, students, and developers. This phased approach will guide the implementation of our application concepts. By establishing an objective and comprehensive evaluation system, we will continually refine our digital human-driven instructional video generation technology, exploring further possibilities with a teaching-centered focus. In the next phase, our primary focus will be on applying multimodal interaction in virtual courses. By integrating various sensory modalities, including speech, facial expressions, gestures, and haptic feedback, these technologies seek to create immersive, personalized, and engaging learning experiences. The ongoing convergence of multimodal interaction technologies is expected to redefine the boundaries of virtual education, bridging the gap between digital instruction and human-centered teaching methodologies and advancing the digitalization of education.

Author Contributions: Software, L.Y., Y.Z. and M.Z.; writing—original draft preparation, S.C., L.Y., Y.Z., Y.X., Z.Z. and J.L.; writing—review and editing, S.C., L.Y., Y.Z., Y.X., Z.Z. and J.L.; supervision, S.C. and J.L.; project administration, S.C. and J.L.; funding acquisition, S.C. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by CMDS Joint Project Promoting DX Inside & Outside the University Grant Number PJ2024-03.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Resendiz Calderon, C.; Farfan-Cabrera, L.; Cazares, I.; Najera, P.; Okoye, K. Assessing benefits of computer-based video training and tools on learning outcomes and motivation in mechanical engineering education: Digitalized intervention and approach. *Front. Educ.* **2024**, *9*, 1292405. [CrossRef]
2. Frolova, E.; Rogach, O.; Ryabova, T. Digitalization of education in modern scientific discourse: New trends and risks analysis. *Eur. J. Contemp. Educ.* **2020**, *9*, 331–336. [CrossRef]
3. Bygstad, B.; Øvrelid, E.; Ludvigsen, S.; Dæhlen, M. From dual digitalization to digital learning space: Exploring the digital transformation of higher education. *Comput. Educ.* **2022**, *182*, 104463. [CrossRef]
4. Kyaw, B.; Posadzki, P.; Paddock, S.; Tudor Car, L. Medical students' digital education of communication skills: Systematic review and meta-analysis by the Digital Health Education Collaboration (Preprint). *J. Med. Internet Res.* **2018**, *21*, e12967. [CrossRef]

5. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39.
6. Kalimullina, O.; Tarman, B.; Stepanova, I. Education in the Context of Digitalization and Culture: Evolution of the Teacher's Role, Pre-pandemic Overview. *J. Ethn. Cult. Stud.* **2020**, *8*, 226. [CrossRef]
7. Chan, A.; Ezell, C.; Kaufmann, M.; Wei, K.; Hammond, L.; Bradley, H.; Bluemke, E.; Rajkumar, N.; Krueger, D.; Kolt, N.; et al. Visibility into AI Agents. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, 3–6 June 2024; FAccT '24; pp. 958–973.
8. Chen, Y.; Jensen, S.; Albert, L.J.; Gupta, S.; Lee, T. Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Inf. Syst. Front.* **2023**, *25*, 161–182. [CrossRef]
9. Bozkurt, A.; Karadeniz, A.; Baneres, D.; Guerrero-Roldán, A.E.; Rodríguez, M.E. Artificial intelligence and reflections from educational landscape: A review of AI Studies in half a century. *Sustainability* **2021**, *13*, 800. [CrossRef]
10. Nur Fitria, T. Utilizing Text-to-Speech Technology: Natural Reader in Teaching Pronunciation. *JETLEE J. Engl. Lang. Teach. Linguist. Lit.* **2022**, *2*, 70–78. [CrossRef]
11. Dai, L.; Kritskaia, V.; Velden, E.; Jung, M.; Postma, M.; Louwerse, M. Evaluating the usage of Text-To-Speech in K12 education. In Proceedings of the 2022 6th International Conference on Education and E-Learning, Yamanashi, Japan, 21–23 November 2022; pp. 182–188.
12. Lin, C.J.; Wang, W.S.; Lee, H.Y.; Huang, Y.M.; Wu, T.T. Recognitions of image and speech to improve learning diagnosis on STEM collaborative activity for precision education. *Educ. Inf. Technol.* **2023**, *29*, 1–26. [CrossRef]
13. Luo, H. Advances in multimodal learning: Pedagogies, technologies, and analytics. *Front. Psychol.* **2023**, *14*, 1286092.
14. Immadisetty, P.; Rajesh, P.; Gupta, A.; Anala, M.R.; Soumya, A.; Subramanya, K.N. Multimodality in Online Education: A Comparative Study. *arXiv* **2023**, arXiv:2312.05797.
15. Mulian, H.; Shlomov, S.; Limonad, L.; Noccaro, A.; Buscaglione, S. Mimicking the Maestro: Exploring the Efficacy of a Virtual AI Teacher in Fine Motor Skill Acquisition. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 23224–23231.
16. Liu, R.; Sisman, B.; Li, J.; Bao, F.; Gao, G.; Li, H. Teacher-Student Training For Robust Tacotron-Based TTS. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6274–6278.
17. Elai.io. AI-Powered E-Learning Video Generation. 2025. Available online: https://elai.io/e-learning/ (accessed on 22 March 2025).
18. Vyond. Training and E-Learning Video Solutions. 2025. Available online: https://www.vyond.com/solutions/training-and-elearning-videos/ (accessed on 22 March 2025).
19. AI Studios. AI Studios: AI-Powered Virtual Humans for Video Creation. 2024. Available online: https://www.aistudios.com (accessed on 24 December 2024).
20. Haji-Ali, M.; Menapace, W.; Siarohin, A.; Skorokhodov, I.; Canberk, A.; Lee, K.S.; Ordonez, V.; Tulyakov, S. AV-Link: Temporally-Aligned Diffusion Features for Cross-Modal Audio-Video Generation. *arXiv* **2024**, arXiv:2412.15191.
21. Tits, N. *Controlling the Emotional Expressiveness of Synthetic Speech: A Deep Learning Approach*; Springer: Berlin/Heidelberg, Germany, 2022.
22. Stoyanova-Doycheva, A.; Ivanova, V.; Doukovska, L.; Tabakova, V.; Radeva, I.; Danailova, S. Architecture of a Knowledge Base in Smart Crop Production. In Proceedings of the 2021 International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, 30 September–2 October 2021; pp. 305–309. [CrossRef]
23. Mayer, R.E. *Multimedia Learning*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009. [CrossRef]
24. DeepSeek. DeepSeek: AI-powered Search Engine. 2024. Available online: https://www.deepseek.com (accessed on 24 December 2024).
25. Wikipedia Contributors. Isaac Newton—Wikipedia. 2024. Available online: https://en.wikipedia.org/wiki/Isaac_Newton (accessed on 24 December 2024).
26. Inaguma, H.; Kiyono, S.; Duh, K.; Karita, S.; Yalta, N.; Hayashi, T.; Watanabe, S. ESPnet-ST: All-in-One Speech Translation Toolkit. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5–10 July 2020; pp. 302–311.
27. Hayashi, T.; Yamamoto, R.; Yoshimura, T.; Wu, P.; Shi, J.; Saeki, T.; Ju, Y.; Yasuda, Y.; Takamichi, S.; Watanabe, S. ESPnet2-TTS: Extending the edge of TTS research. *arXiv* **2021**, arXiv:2110.07840.
28. Li, C.; Shi, J.; Zhang, W.; Subramanian, A.S.; Chang, X.; Kamo, N.; Hira, M.; Hayashi, T.; Boeddeker, C.; Chen, Z.; et al. ESPnet-SE: End-to-End Speech Enhancement and Separation Toolkit Designed for ASR Integration. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; IEEE: New York, NY, USA, 2021; pp. 785–792.
29. Arora, S.; Dalmia, S.; Denisov, P.; Chang, X.; Ueda, Y.; Peng, Y.; Zhang, Y.; Kumar, S.; Ganesan, K.; Yan, B.; et al. ESPnet-SLU: Advancing Spoken Language Understanding through ESPnet. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 7–13 May 2022; IEEE: New York, NY, USA, 2022; pp. 7167–7171.

30. Shi, J.; Guo, S.; Qian, T.; Huo, N.; Hayashi, T.; Wu, Y.; Xu, F.; Chang, X.; Li, H.; Wu, P.; et al. Muskits: An End-to-End Music Processing Toolkit for Singing Voice Synthesis. In Proceedings of the Interspeech, Incheon, Republic of Korea, 18–22 September 2022; pp. 4277–4281.
31. Lu, Y.J.; Chang, X.; Li, C.; Zhang, W.; Cornell, S.; Ni, Z.; Masuyama, Y.; Yan, B.; Scheibler, R.; Wang, Z.Q.; et al. ESPnet-SE++: Speech Enhancement for Robust Speech Recognition, Translation, and Understanding. In Proceedings of the Interspeech, Incheon, Republic of Korea, 18–22 September 2022; pp. 5458–5462.
32. Gao, D.; Shi, J.; Chuang, S.P.; Garcia, L.P.; Lee, H.y.; Watanabe, S.; Khudanpur, S. EURO: ESPnet unsupervised ASR open-source toolkit. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: New York, NY, USA, 2023; pp. 1–5.
33. Peng, Y.; Tian, J.; Yan, B.; Berrebbi, D.; Chang, X.; Li, X.; Shi, J.; Arora, S.; Chen, W.; Sharma, R.; et al. Reproducing Whisper-style training using an open-source toolkit and publicly available data. In Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan, 16–20 December 2023; IEEE: New York, NY, USA, 2023; pp. 1–8.
34. Sharma, R.; Chen, W.; Kano, T.; Sharma, R.; Arora, S.; Watanabe, S.; Ogawa, A.; Delcroix, M.; Singh, R.; Raj, B. ESPnet-SUMM: Introducing a novel large dataset, toolkit, and a cross-corpora evaluation of speech summarization systems. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan, 16–20 December 2023; pp. 1–8.
35. Jung, J.w.; Zhang, W.; Shi, J.; Aldeneh, Z.; Higuchi, T.; Theobald, B.J.; Abdelaziz, A.H.; Watanabe, S. ESPnet-SPK: Full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models. In Proceedings of the Interspeech, Kos Island, Greece, 1–5 September 2024.
36. Yan, B.; Shi, J.; Tang, Y.; Inaguma, H.; Peng, Y.; Dalmia, S.; Polák, P.; Fernandes, P.; Berrebbi, D.; Hayashi, T.; et al. ESPnet-ST-v2: Multipurpose Spoken Language Translation Toolkit. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), Toronto, ON, Canada, 10–12 July 2023; pp. 400–411.
37. Getimg.ai. Getimg.ai—AI-Powered Image Generation. 2024. Available online: https://getimg.ai/home (accessed on 26 December 2024).
38. Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; Sebe, N. First Order Motion Model for Image Animation. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019.
39. Wei, H.; Yang, Z.; Wang, Z. AniPortrait: Audio-Driven Synthesis of Photorealistic Portrait Animations. *arXiv* **2024**, arXiv:2403.17694.
40. Chen, Z.; Cao, J.; Chen, Z.; Li, Y.; Ma, C. EchoMimic: Lifelike Audio-Driven Portrait Animations through Editable Landmark Conditioning. *arXiv* **2024**, arXiv:2407.08136.
41. Meng, R.; Zhang, X.; Li, Y.; Ma, C. EchoMimicV2: Towards Striking, Simplified, and Semi-Body Human Animation. *arXiv* **2024**, arXiv:2411.10061.
42. Li, H.; Xu, J. Semantic matching in search. *Found. Trends® Inf. Retr.* **2014**, *7*, 343–469.
43. Priest, G. Contradictory concepts. *Logic Reason. Ration.* **2014**, 197–215. [CrossRef]
44. Deng, N.; Wang, Y.; Huang, G.; Zhou, Y.; Li, Y. Semantic Coherence Analysis of English Texts Based on Sentence Semantic Graphs. *Icst Trans. Scalable Inf. Syst.* **2023**, *10*. [CrossRef]
45. Hincks, R. Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System* **2005**, *33*, 575–591.
46. Kumar, S.; Phadikar, S.; Majumder, K. Modified segmentation algorithm based on Short Term Energy & Zero Crossing Rate for Maithili speech signal. In Proceedings of the 2016 International Conference on Accessibility to Digital World (ICADW), Guwahati, India, 16–18 December 2016; pp. 169–172. [CrossRef]
47. Tobin, J.; Li, Q.; Venugopalan, S.; Seaver, K.; Cave, R.; Tomanek, K. Assessing asr model quality on disordered speech using bertscore. *arXiv* **2022**, arXiv:2209.10591.
48. Lai, D.; Zhang, X.; Ma, K.; Chen, Z.; Chen, W.; Zhang, H.; Yuan, H.; Ding, L. Automated detection of high frequency oscillations in intracranial EEG using the combination of short-time energy and convolutional neural networks. *IEEE Access* **2019**, *7*, 82501–82511. [CrossRef]
49. Henton, C. Pitch dynamism in female and male speech. *Lang. Commun.* **1995**, *15*, 43–61. [CrossRef]
50. Trouvain, J. Tempo variation in speech production. *Implic. Speech Synthesis. Phonus* **2004**, *8*.
51. Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V.P.; Jawahar, C. A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, Virtual, 12–16 October 2020; pp. 484–492.