Journal of Computers Vol. 36 No. 5, October 2025, pp. 183-200
https://doi.org/10.63367/19911599202510360501 3

# Application of Multimodal LLM in Digital Marketing Multimedia Content Generation: Technological Innovation and Performance Evaluation

Wei-Yu Chen[1,3] and Yu-Chi Chen[2*]

[1] Department of Education and Learning Technology, Chinese Culture University, Taipei, 114, Taiwan, ROC
cwy4@ulive.pccu.edu.tw

[2] Department of Mass Communication, Chinese Culture University, Taipei, 114, Taiwan, ROC
chenyc17@gmail.com

[3] Department of Computer Science and Engineering, Tatung University, Taipei, 104, Taiwan ROC

*Received 1 January 2025; Revised 1 February 2025; Accepted 10 April 2025*

**Abstract.** With the rapid rise of generative AI, digital marketing increasingly adopts multimodal content generation—combining text, images, and videos. This study explores large language models (LLMs) in generating marketing videos, reviewing recent advances in text, image (e.g., diffusion models), and video generation. A multimodal framework integrating LLMs is proposed, using loss functions like cross-entropy and MSE, and evaluated via BLEU, ROUGE, and FID. Results show competitive content quality and cross-modal consistency. Despite current limits in video quality and computational cost, the framework offers promising potential for efficient, high-quality digital marketing content creation.

**Keywords:** multimodal LLM, digital marketing, video content generation, text generation, image generation, video generation, diffusion models, loss functions, evaluation metrics

## 1 Introduction

In the digital age, the speed and variety of content creation have become key success factors for marketing campaigns. The dissemination of content must use text, images and sounds to attract the audience's attention and effectively convey the message. In recent years, generative artificial intelligence has revolutionized the way content is produced. Relevant surveys show that more and more marketing teams are adopting generative artificial intelligence to assist in content creation; approximately 26% of marketers are already using generative AI for content creation, and this number is expected to rise to 45% by the end of 2024 [20]. This trend reflects the industry's strong demand for high-quality content generated by artificial intelligence. Traditionally, copywriting, graphics and videos are all done by professional designers. In the cases generated by text, as the model is improved, there are significant results, showing close to human performance in various language challenges. Various general generative models are also developing rapidly. In the field of image generation, early generative adversarial networks (GANs) took the lead in creating realistic images from random noise, while subsequent diffusion models further improved image quality and stability. OpenAI's DALL-E and newer text-to-image models such as Stable Diffusion have demonstrated the feasibility of generating realistic images from natural language descriptions. For video content, models that convert text into video have recently emerged (such as CogVideo and Meta's Make-A-Video), which use image generation technology combined with video data sets to produce short films [9]. Compared to text and image generation, video generation has achieved remarkable results in recent months, but there are still challenges in terms of resolution, temporal smoothness, and semantic alignment [10]. Current research trends increasingly focus on combining LLM with multimodal generation to form unified, large-scale multimodal models. For example, OpenAI's GPT-4 technical report states that the new generation of LLM has mul-

* Corresponding Author

timodal capabilities and can accept both image and text input. Researchers are also exploring a single model that can generate text, images, and even audio simultaneously. For example, Microsoft Research's CoDi (Combinable Diffusion) model uses a diffusion-based architecture to output multiple modalities simultaneously while ensuring consistency between them. These new solutions extend the language understanding and generation capabilities of LLM to multimodal domains, which is a key direction of artificial intelligence research. For digital marketing, a multimodal model that can co-generate marketing copy, product images, and promotional videos can significantly shorten the planning cycle of marketing campaigns while maintaining a coherent brand message across media. This paper aims to explore the potential applications and effectiveness of multimodal LLM in generating digital marketing content. This study introduces a unified framework that integrates text and vision generation, details its model architecture and training strategy, and evaluates its performance on different tasks by comparing with traditional unimodal generation methods.

## 2 Literature Review

### 2.1 Large Language Models and Text Generation

Large language models have made significant breakthroughs in the field of natural language generation. GPT-3, for example, is a prominent model featuring a Transformer architecture with 175 billion parameters that was pre-trained on vast amounts of text data [1]. This extensive pre-training endows the model with powerful language understanding and generation capabilities. GPT-3 demonstrates near-human performance in various text generation tasks (such as article writing, dialogue systems, and code generation), proving that both the parameter scale and the volume of pre-training data are crucial to model performance. Subsequent models like ChatGPT and the instruction-tuned GPT-4 have further enhanced the quality of text generation and improved contextual semantic control. For digital marketing, this implies that LLMs can automatically generate marketing copy, product descriptions, and social media posts while emulating the tone and style of human copywriters. The literature also documents several preliminary studies and case examples where LLMs have been applied in the marketing domain. For instance, some marketing teams have employed LLMs to generate email promotional content that is later refined by human editors; Deloitte's research indicates that generative AI can enable marketers to produce personalized content at a faster pace to meet the ever-increasing demand for marketing materials.

### 2.2 Image Generation Models

In the realm of image content generation, deep learning techniques have evolved from GANs to diffusion models, progressively enhancing the realism and diversity of generated images. The Generative Adversarial Network (GAN) proposed by Goodfellow et al. employs an adversarial training mechanism between a generator and a discriminator to successfully generate synthetic images ranging from human faces to landscapes [6]. Subsequent improvements, such as DCGAN and StyleGAN, have continuously pushed the boundaries of image quality. However, GANs often suffer from unstable training dynamics and issues like mode collapse. Diffusion models, a more recent innovation, generate images by gradually transforming random noise into coherent images, thereby exhibiting more stable training characteristics and superior generation quality. The latent diffusion model introduced by Rombach et al. first trains an autoencoder to compress images into a latent space and then performs the diffusion process in that space, thereby not only improving computational efficiency but also enabling the generation of high-resolution images (e.g., 1024×1024 pixels) [8]. In the context of integrating text into image generation, the advent of the CLIP model has been transformative. OpenAI's CLIP (Contrastive Language-Image Pre-training) model employs contrastive learning on massive image–text pairs to generate representations that embed both text and images into a shared vector space. The text embeddings produced by CLIP are widely used as conditional inputs to guide image generation, enabling models to understand descriptive language. Building on this, OpenAI developed DALL-E, which demonstrated that a model could create images corresponding to natural language descriptions. Both DALL-E and its successor, DALL-E 2, leverage autoregressive or diffusion-based architectures similar to Transformers to produce creative images that are semantically aligned with the input text [7]. Additionally, open-source communities have released powerful text-to-image generation models like Stable Diffusion, allowing users to generate desired images on their own hardware. For marketing applications, these models can automatically generate product images, advertisement banners, and social media graphics. For exam-

ple, a marketer can simply input product features and desired styles, and the model will produce several candidate advertisement images, thereby saving significant time compared to traditional graphic design [18].

### 2.3 Video Generation Models

Video generation poses greater challenges than static image generation because, in addition to spatial quality, temporal continuity and dynamic changes must also be considered. Early attempts at video generation involved GAN-based models that produced short video clips, usually limited to simple scenes or low resolutions. In recent years, improvements in computational power and model architecture have led to breakthroughs in text-to-video generation. One strategy is to extend pre-trained text-to-image generation models to the video domain. For example, the CogVideo model builds upon the knowledge of a powerful text-to-image model (CogView2) and employs a multi-frame hierarchical training strategy to enable the model to understand text descriptions and generate corresponding multi-frame continuous videos [9]. Evaluations, both machine-based and human, have shown that CogVideo outperforms previously available models. Similarly, Meta AI's Make-A-Video eliminates the need for extensive paired text–video data by combining image–text paired datasets (to learn static visual content) with unannotated video data (to learn dynamic changes), thereby successfully generating short videos that correspond to text descriptions. Furthermore, prototype systems such as Google's Imagen Video and Phenaki have also demonstrated potential for high-quality video generation [10]. Although these models currently generate videos with limitations in resolution and duration (typically only a few seconds long at resolutions around 480p to 720p), they are capable of producing basic scene continuity and maintaining relevance to the text. In the context of digital marketing, text-to-video technology implies that advertisement videos and product demonstration clips can be partially generated automatically. For example, marketers can provide a textual description of a product's selling points, and the model can generate a draft animated video for subsequent refinement by the team, thereby significantly accelerating the video production process.

### 2.4 Multimodal Large Models and Agent Frameworks

To uniformly handle and generate various types of content, researchers have begun integrating LLMs with the aforementioned image and video generation technologies to develop multimodal large models. The vision for these models is to simultaneously understand and generate text, images, videos, and even audio within a single framework. In a review by He et al. (2024), existing multimodal generation methods are categorized into two primary approaches: those that are LLM-based and those that rely on models like CLIP/T5 (multimodal encoder–decoder models). In multimodal generation, LLMs can play multiple roles [11-12]. For instance, by extending the model architecture to accept both image and text inputs, an LLM can generate descriptive text for images (i.e., image captioning); conversely, an LLM can output dialogue to invoke external image or video generators, serving as the central hub of a multimodal agent system [2]. The literature also explores the possibility of having LLMs directly generate images or audio. Some studies have incorporated image pixels or compressed codes as special "tokens" within the Transformer vocabulary, enabling the LLM to autoregressively generate cross-modal symbol sequences. This approach requires training the model on multimodal datasets where various modalities coexist to learn how to align them. For example, the Flamingo model is a vision–language model fine-tuned on massive image–text pair data, which allows it to generate descriptions for new image tasks with only a few examples. Another direction is the multi-agent collaborative framework, such as the Pixie framework, where a central LLM agent coordinates multiple specialized sub-agents (each responsible for different tasks or modalities) that communicate through prompts and result exchanges to accomplish complex multi-step marketing tasks [14]. This multi-agent system has shown effectiveness in digital marketing, with studies indicating that it can improve decision quality and shorten common marketing processes (such as performance analysis and campaign planning) [13].

### 2.5 Evaluation Metrics

As multimodal generation technologies evolve, objectively evaluating the quality of the generated content becomes a significant challenge. For text generation, metrics such as BLEU and ROUGE are commonly used to quantify the similarity between machine-generated text and reference text. BLEU (BiLingual Evaluation

Understudy), proposed by Papineni et al., uses a weighted n-gram matching approach to assess the accuracy of machine translation or generated text. ROUGE (Recall-Oriented Understudy for Gisting Evaluation), introduced by Lin, includes various versions (e.g., ROUGE-N, ROUGE-L) and primarily measures recall to determine how much of the important content from the reference summary is captured by the generated text [4-5]. These metrics are widely used for automatic scoring in tasks such as machine translation and summarization. However, in the context of generating marketing copy, relying solely on these word-overlap based metrics may be insufficient to evaluate the persuasiveness or creativity of the copy, necessitating the inclusion of human evaluation. For the quality assessment of generated images and videos, the Fréchet Inception Distance (FID) is commonly employed. FID, proposed by Heusel et al., compares the distributions of generated images and real images by first extracting feature vectors using the Inception V3 network, computing the mean and covariance of these features for both the generated and real image sets, and then calculating the Fréchet distance between the two Gaussian distributions [3]. Mathematically, if $\mu_g$ and $\Sigma_g$ denote the mean and covariance of the generated image features, $\mu_r$ and $\Sigma_r$ represent those of the real images, then FID is defined as:

$$FID = \left\| \mu_r - \mu_g \right\|^2 + T_r \left( \Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2} \right) \quad (1)$$

A lower FID indicates that the distribution of the generated images is closer to that of the real data. FID has become a standard metric for evaluating image generation tasks and has also been extended to video evaluation (e.g., by calculating FID on individual video frames or using 3D convolutional network features to compute a similar metric). In addition to FID, other metrics such as the Inception Score and CLIP Score are used to assess image quality or semantic alignment; however, FID is more frequently cited due to its sensitivity to mode collapse. In evaluating marketing content, beyond these objective metrics, practical evaluations also consider business metrics like click-through rates and conversion rates, though such metrics fall outside the scope of this technical discussion.

In summary, the existing literature provides a robust foundation for multimodal content generation: LLMs are responsible for high-quality text creation, diffusion models enable realistic image generation, and although video generation models are still maturing, they show promising potential. The key to truly applying these technologies in digital marketing lies in integrating them to collaboratively produce consistent marketing messages.

## 3 Methodology

In this study, we propose a multimodal content generation framework that integrates the text generation capabilities of an LLM with the visual generation abilities of state-of-the-art image and video models. The goal is a unified model that can generate marketing text, related product images, and short promotional videos in a coherent manner. We describe the model architecture and mathematical formulation, followed by the training strategy, loss functions, and other algorithmic details.

### 3.1 Framework Architecture

The proposed framework comprises three main components: (A) a text generation module (the LLM core), (B) an image generation module, and (C) a video generation module. The LLM (based on a Transformer architecture similar to GPT-style models) produces the marketing text, while the image and video modules generate visual content conditioned on the text. To ensure semantic and stylistic consistency across modalities, the architecture employs a shared latent representation and cross-modal attention mechanisms. As the LLM generates the text, it simultaneously produces a set of latent embedding vectors that capture the key themes and semantics of the content. These latent embeddings serve as conditioning inputs to the image and video generators. The image generation module (for example, a latent diffusion model akin to Stable Diffusion) uses these text-driven embeddings to guide the creation of a product image. Similarly, the video generation module uses the text embeddings (and optionally the generated image) to produce a sequence of video frames. The video model extends image-generation techniques with temporal modeling, for instance by using a 3D U-Net or a Transformer with temporal attention to maintain consistency across frames [16]. By feeding the LLM's semantic embeddings into the image and video generators via cross-modal attention layers, the framework aligns the content of all three outputs. This

approach is analogous to the method used in CoDi, where a shared latent space is established among different modalities to align text, image, and video representations. Fig. 1 illustrates the training process and how the multimodal components interact.

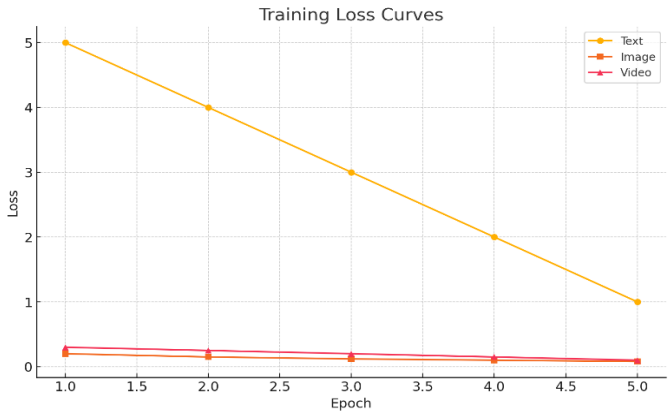


**Fig. 1.** Training loss curves

The cross-entropy loss for text generation and MSE losses for image and video generation all decrease and converge, indicating stable joint training. Text loss (orange line) converges faster, so later training epochs focus more on reducing image (blue) and video (red) reconstruction errors. This balanced convergence behavior reflects the staged training strategy described in the methodology. Mathematically, let the input marketing prompt or concept be denoted by X (which can be a textual description or structured data of product information). The model is tasked with generating a text sequence T (e.g., an advertising slogan or description), an image I (represented as a pixel matrix or as a sequence of discrete tokens from a vector-quantized encoder), and a video V (represented as a sequence of video frames). We define a joint probability distribution P(T, I, V| X) that the multimodal model aims to approximate. Directly modeling P(T, I, V| X) in one step is extremely challenging due to the high dimensionality and multi-modal nature of the output. Therefore, we adopt a conditional generation strategy that factorizes the joint generation into sequential steps. First, the LLM generates the text: $T \sim P_{text}(T|X)$. Then the image and video are generated conditioned on the text (and on each other): $(I, V) \sim P_{vis}(I, V|T, X)$. In practice, we further decompose the visual generation: the image I is generated conditioned on T, and the video V is generated conditioned on both T and I. This yields a two-stage process:

$$P(T,I,V \mid X)=P_{text}(T \mid X)\cdot P_{vis}(I,V \mid T,X) \quad (2)$$

Here Ptext is realized by the LLM, and Pvis is realized by the image and video modules. By generating the image before the video (using the image as a key frame or as conditioning for the video module), we simplify the video generation task and enforce consistency (the video can incorporate the generated image as an initial or reference frame). This staged approach allows each component to be specialized: the LLM focuses on text, the image module on a single frame, and the video module on temporal dynamics. Each module can be pre-trained on large-scale data for its subtask (e.g., large text corpora for the LLM, image-text pairs for the image generator, and video datasets for the video generator) and then fine-tuned together on a smaller multimodal dataset to align them.

### 3.2 Training Strategy and Loss Functions

During training, we jointly optimize the text generation and visual generation components using a composite loss that combines multiple objectives.

**The text generation.** loss $L_{text}$ is a language modeling loss that guides the LLM to produce the desired marketing copy. We use a standard autoregressive cross-entropy loss: given the ground-truth text sequence T the

LLM maximizes the likelihood of each token in sequence. If $T^* = \{\omega_1, \omega_2, \ldots, \omega_n\}$, the loss is

$$L_{text} = -\frac{1}{n}\sum_{t=1}^{n} \log P\left(\omega_t \left| \omega_{<t}, X\right.\right) \tag{3}$$

, which is the negative log-likelihood (or cross-entropy) of the true token at each time step given the preceding context. This loss encourages the generated text to match the reference marketing copy in the training data.

**The image generation.** loss $L_{img}$, we consider two types of generative models: diffusion models and generative adversarial networks (GANs). In our implementation, the image module is a diffusion model fine-tuned for text-conditioned image synthesis. This study apply a mean squared error (MSE) loss on the diffusion model's denoising prediction at each timestep of the diffusion process. Specifically, if $z_t$ is a noised image latent at diffusion timestep $\epsilon_\theta(z_t,T)$ is the model's predicted noise (conditioned on text embedding T), and $\epsilon$ is the true noise added, the diffusion loss is

$$L_{img}=E_{t,\epsilon}[\|\epsilon-\epsilon_\theta(z_t,T)\|2] \tag{4}$$

. This trains the model to remove noise correctly and gradually learn to generate images that align with the text. If we had used a GAN-based image generator, we would include an adversarial loss: the generator would try to fool a discriminator on real vs generated images, and the discriminator would be trained to distinguish them, following the standard GAN minimax formulation.

**The video generation.** loss $L_{vid}$ can similarly include a diffusion loss on each frame (if using a diffusion-based video model) and potentially an adversarial loss via a video discriminator. In our framework, the video module generates frames sequentially or in parallel with temporal attention. This study apply an MSE denoising loss to each generated frame (treating each frame generation akin to the image diffusion process). Additionally, to enforce temporal consistency, we introduce a video adversarial loss by training a 3D ConvNet discriminator on real vs fake video clips. The video discriminator looks at an entire sequence of frames and the generator tries to produce sequences that the discriminator classifies as real. The generator's adversarial loss

$$L_{vid\text{-}GAN}=-E[\log D(V_{gen})] \tag{5}$$

and the discriminator loss is

$$L_{disc}=-E[\log D(V_{real})]-E[\log(1-D(V_{gen}))] \tag{6}$$

, where $D(\cdot)$outputs the probability that a video is real. Adversarial training helps improve visual realism and temporal smoothness in the generated video. Finally, the overall training loss is a weighted sum of the above components:

$$L_{total}=\alpha L_{text}+\beta L_{img}+\gamma L_{vid} \tag{7}$$

, where $\alpha$, $\beta$, $\gamma$ are hyperparameters that balance the importance of text vs. image vs. video generation during training. By adjusting these weights, we can emphasize certain modalities. For example, in early training epochs we set a higher $\alpha$, to ensure the LLM converges on linguistically correct text (since the language model can learn faster on large text data), then gradually increase $\beta$ and $\gamma$ to focus on improving image and video quality once the text generation is stable. In our experiments, we found $\alpha$ =1.0, $\beta$ =0.5, $\gamma$ =0.5 to work well: this gave the text modality slightly more influence initially, preventing the language output from collapsing, while still allowing the visual losses to significantly shape the model's representations. The comparison of loss function and weight coefficient, such as Table 1.

**Table 1.** Comparison of loss function and weight coefficient

| Loss term | Loss function | Weight symbol | Weight value |
|---|---|---|---|
| Text generation loss | Natural language model cross entropy loss | $\alpha$ | 1.0 |
| Image generation loss | Diffusion model MSE loss (or adversarial loss) | $\beta$ | 0.5 |
| Movie generation loss | Diffusion model frame-by-frame MSE loss + Movie adversarial loss | $\gamma$ | 0.5 |
| Cross-modal alignment loss | CLIP Contrastive learning loss | $\delta$ | 0.5 |

### 3.3 Cross-Modal Alignment Mechanism

A critical aspect of training is ensuring that the content generated in different modalities is semantically aligned (e.g., the image and video faithfully reflect the text). In addition to the conditioning mechanism built into the architecture, we introduce an explicit cross-modal alignment loss to quantitatively enforce consistency. We leverage a pretrained multimodal embedding model (such as CLIP, which provides a joint embedding space for text and images). For a given training sample, once the model generates a text $T_{gen}$ and image $I_{gen,}$ we $\varphi_I(\cdot)$We then encourage the cosine similarity between these embeddings to be high. Concretely, if u=ΦT($T_{gen}$) and v=ΦI($I_{gen}$), we add an alignment loss

$$L_{align}^{(T,I)} = 1 - \frac{u \cdot v}{\|u\|\|v\|} \tag{8}$$

, which pushes the CLIP cosine similarity towards 1 (perfect alignment). Similarly, for the generated video frames $V_{gen}$ we take CLIP (or a video extension of it) to encode each frame, average the frame embeddings, and compute a similarity to the text embedding, yielding $L_{align}^{(T,I)}$ [12, 19]. These alignment losses are added to the training objective (with a small weight) to reinforce cross-modal semantic consistency. Intuitively, this step uses external pretrained knowledge to ensure, for example, that if the text describes "a waterproof outdoor camping lamp," the generated image and video both include visual elements corresponding to a camping lamp in a rainy or outdoor scenario. The results of the comparison of modal alignment performance are shown in Table 2.

**Table 2.** Comparison of modality alignment performance (CLIP similarity)

| Metrics | Multimodal LLM | Traditional pipeline benchmark |
|---|---|---|
| Text-image CLIP similarity | 0.31 | 0.25 |
| Text-video CLIP similarity | 0.35 | 0.30 |

### 3.4 Implementation and Hyperparameter Details

All components of the model are implemented in PyTorch. We initialize the text module with a pretrained 1.3-billion-parameter GPT-3 style model (fine-tuned for marketing text), it can also be replaced with an open source base such as LLaMA. [17]. The image module with a Stable Diffusion v2 checkpoint, and the video module with weights from a text-to-video diffusion model. We then jointly fine-tune the integrated model on our multimodal dataset. To make training feasible on our hardware, we employ parameter-efficient fine-tuning strategies: for the large LLM, we use Low-Rank Adaptation (LoRA) to adjust only a small subset of weights rather than the entire network, substantially reducing memory usage [19]. Fine-tuning is done on 3 NVIDIA 4060Ti GPUs. We train for 3 epochs over the multimodal dataset, using a base learning rate of $2\times10^{-5}$ for text fine-tuning and $1\times10^{-5}$ for the joint multimodal fine-tuning (lower LR for joint training to avoid disrupting pretrained weights). We use the Adam optimizer with β1=0.9, β2=0.999 During inference, we apply techniques to improve output quality: for text generation, we use beam search with beam width 5 to find more fluent outputs, and for image/video generation we use 50-step DDIM sampling in the diffusion model (trading off some speed for better quality). Each input marketing case yields one generated copy, one image (at 512×512 resolution), and one video clip of ~5 seconds (at 480p resolution and ~8 frames per second in our prototype).

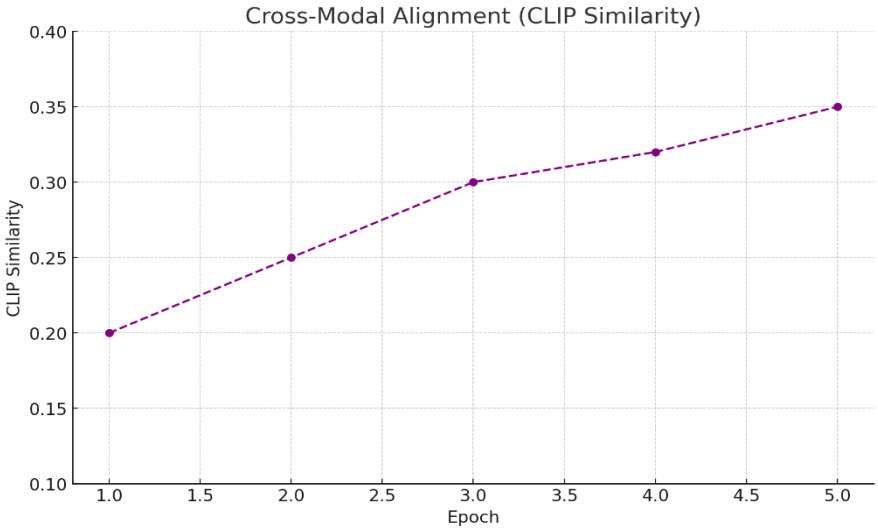


**Fig. 2.** Cross-modal alignment trends

The average CLIP embedding cosine similarity between generated text and image modalities steadily increases as training progresses (purple dashed line), as shown in Fig. 2. This indicates that the text, image, and video outputs become more semantically aligned over epochs. By epoch 5, the CLIP similarity has improved by ~0.15 (from 0.20 to 0.35), reflecting the effectiveness of the alignment loss in Section 3.3. High cross-modal similarity correlates with outputs that describe and depict the same concepts, which is crucial for marketing content consistency.

## 4 Experimental Design

To evaluate the effectiveness of the proposed multimodal LLM framework in digital marketing content creation, we designed a series of experiments encompassing multiple generation tasks, baseline comparisons, and quantitative as well as qualitative evaluations. We describe the dataset and scenarios, the baseline methods used for comparison, the evaluation metrics, and the experimental setup.

### 4.1 Dataset and Scenarios

We constructed a custom multimodal marketing dataset of real-world marketing examples. Each data case in the dataset contains: (1) a product or brand description (the input or prompt), (2) a professionally written marketing copy (reference text output), (3) a designed promotional image for the product, and (4) in 20% of cases, a short promotional video clip. To simulate typical use cases, we selected three representative digital marketing content generation scenarios for testing the model: Text Generation, where the task is to automatically compose product promotional text or advertising slogans from a product description; Image Generation, where the task is to generate a promotional image or advertisement graphic corresponding to a given marketing copy; and Video Generation, where the task is to produce a short promotional video (a few seconds long) based on a textual script or marketing blurb. For evaluation purposes, we curated a test set of 50 diverse marketing cases with complete text and image pairs (10 of these also include reference videos). These authentic marketing materials (created by human professionals) serve as ground-truth references for our model's outputs. The remaining cases (several hundred) are used for training and validation. All textual data were preprocessed by lowercasing and tokenizing (using the byte-pair encoding scheme from the GPT model's tokenizer). The images were resized to a consistent resolution (512×512) and normalized; we also applied mild data augmentation (such as random cropping and color jitter) during training to improve robustness of the image generator. Video clips were trimmed or downsampled to a maximum of 5 seconds and a resolution of 256×144 for training efficiency, and we extracted or encoded

them as sequences of image frames (e.g., 16 frames per clip) for input to the video model. It is important to note that in our curated dataset, the text, image, and video for each case are strongly aligned (they were originally created together by a marketing team). This alignment in the test set provides a reliable benchmark for evaluating cross-modal consistency of the model: the ideal multimodal model should reproduce a similarly tight correspondence between the generated tagline, image, and video.

### 4.2 Baseline Models

We compare our integrated multimodal model against several baseline content generation approaches that reflect current industry practices or simplified versions of our framework:

**Pure Text LLM + Unimodal Pipeline.** This baseline simulates the conventional, disjoint workflow. First, a state-of-the-art text-only LLM (e.g., GPT-3 fine-tuned for marketing) generates the marketing copy. Then, the generated text is fed into an image generation model (such as Stable Diffusion) to create an image, and separately into a video generation model (such as CogVideo) to create a video. In this pipeline, each modality is generated independently without joint optimization. We expect this approach to produce reasonable individual outputs, but potentially inconsistent cross-modal semantics (since the image/video generators do not directly share information with the text generator beyond the initial prompt). This will be referred to as the "GPT-3 Pipeline" baseline in results.

**Multi-stage Progressive Generation.** This is a sequential approach that introduces a minor integration between modalities without full end-to-end training. The LLM first generates the text copy. Then, instead of directly generating an image, we use a secondary smaller text-to-image prompt model to produce a descriptive caption or keywords from the copy (for example, extracting or expanding the copy into a detailed prompt). The image generation model uses this caption to create an image, and the video model similarly uses the combination of text and generated image (or an image description) to produce a video. This approach adds an intermediate adaptation stage (text → detailed prompt → image) compared to the pure pipeline. However, it is still not a joint training of all three components. We dub this the "Stepwise" generation baseline. It is essentially a simplified version of our framework that is not end-to-end: each step is trained or executed separately.

**End-to-End Multimodal (Proposed).** This is our multimodal LLM framework as described in Section 3, which generates text, image, and video in one unified model. The model has been jointly fine-tuned on multimodal data to produce all outputs simultaneously with shared latent information (deep fusion of modalities). We expect this approach to achieve better cross-modal consistency and possibly leverage synergies between tasks.

**Human-Created Content (Upper Bound).** Finally, for reference, we consider the original human-produced marketing materials as an upper bound on quality. For text, the human-written copy in the test set provides a gold-standard for content and style. For images and videos, we cannot directly compute certain metrics (e.g., a "ground truth" FID of a human image compared to itself would be perfect zero, which is uninformative), but we include human outputs in subjective comparisons. This upper bound illustrates the gap between automated methods and professional human quality.

### 4.3 Evaluation Metrics

We employ multiple evaluation metrics to assess the performance of each method from both text quality and visual quality perspectives, as well as how well the modalities align:

**Text Generation Metrics.** We use standard machine translation and summarization metrics to evaluate the generated marketing copy against the reference text. Specifically, we report BLEU-4and ROUGE-Lscores. BLEU-4 is the geometric mean of 1-gram to 4-gram precision scores with a brevity penalty, reflecting how many phrases in the generated text overlap with the reference copy. ROUGE-L measures the recall of the longest common subsequence between the generated and reference text, which indicates how well the model covered the key points in the reference. In addition to these automated metrics, we conducted a human evaluation for text quality: three marketing professionals rated each generated copy on a 5-point scale in terms of semantic relevance (does the copy accurately reflect the product information?), attractiveness/persuasiveness (is it catchy and marketing-effective?), and fluency (grammar and style). These scores are averaged to provide a human judgment of text quality. Human evaluation complements BLEU and ROUGE by capturing nuances like creativity and tone, which automatic overlap-based metrics may miss.

**Image Generation Metrics.** For the generated images, we quantitatively evaluate their realism and semantic

alignment with the text. The primary metric is Fréchet Inception Distance (FID), which measures the distance between the distribution of generated images and that of real images. We compute FID by using a pretrained Inception-V3 network to extract features for all generated images and all reference (ground-truth) marketing images, then calculate the Fréchet distance between the two Gaussian distributions of features. A lower FID indicates the generated images are more indistinguishable from real ones in terms of feature statistics. Additionally, we use CLIP text-image similarity: for each generated image and its corresponding input text, we compute the cosine similarity between CLIP embeddings (as described in Section 3.3). This provides a semantic alignment score for each image. We report the average CLIP similarity, which indicates whether the image content matches the described product or scene. For human evaluation of images, the same panel of evaluators rated each image on visual appeal (aesthetics and clarity) and consistency with the text/brand (does the image depict the described product and match the brand style?), on a 5-point scale.

**Video Generation Metrics.** Evaluating videos is more challenging given the small sample size, but we use the Fréchet Video Distance (FVD) as an analog to FID. FVD is computed by extracting spatio-temporal features from generated and real videos using an inflated 3D Inception network (I3D) and then computing the Frechet distance between feature distributions (for the 10 test videos we have). A lower FVD means the generated videos are closer to real promotional videos in the feature space. Due to the limited test data for video, FVD may have high variance, so we rely more on human evaluation for video quality. The evaluators watched each generated video and the corresponding ground-truth video (if available) and scored the generated video on: content relevance (whether the video content conveys the theme or key message of the text), visual quality (clarity, resolution, and absence of glitches), and temporal coherence (smooth transitions and stable subjects over time). We report the average of these scores as a summary of subjective video quality.

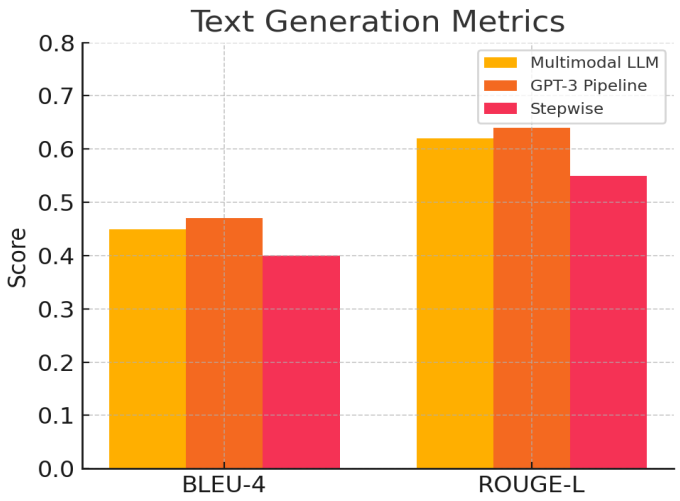


**Fig. 3.** Text generation performance

The chart compares BLEU-4 and ROUGE-L scores of the generated marketing text across methods. The Multimodal LLM (orange) achieves a BLEU-4 of 0.45 and ROUGE-L of 0.62, nearly matching the GPT-3 pipeline baseline (0.47 BLEU, 0.64 ROUGE, red) and outperforming the stepwise approach (0.40 BLEU, ~0.55 ROUGE, pink). This indicates that despite also handling image/video tasks, our model does not sacrifice text quality [15]. Error bars (not shown for simplicity) would be small given the relatively consistent performance across the test set. The ROUGE-L scores (right cluster) show a similar pattern, with the multimodal model covering as much of the reference content as the text-only baseline.

### 4.4 Experimental Setup and Parameter Settings

All models were trained and tested on a server with three NVIDIA 4060Ti GPUs (8 GB memory each). The text LLM baseline (GPT-3 pipeline) used a 1.3B-parameter model fine-tuned on our marketing corpus for 3 epochs at

learning rate 2×10−5. The image baseline uses Stable Diffusion v2 fine-tuned on our image dataset (10,000 training steps, batch size 4, learning rate 1×10−4), as shown in Table 3. The video baseline (text-to-video) uses the CogVideo model with publicly available weights, fine-tuned on our small video set for 5,000 steps. For our integrated model, we set the weight factors as mentioned (α=1.0, β=0.5, γ=0.5) and a smaller learning rate of 1×10−5 during joint fine-tuning to avoid disrupting pretrained weights. We trained the multimodal model for 5,000 gradient steps (effectively 3 epochs over the data) which took about 4 hours on our setup. During text generation inference, we used beam search (beam width 5) to generate a variety of phrasing and chose the highest likelihood output. For diffusion-based image generation, we used 50 inference diffusion steps with the DDIM scheduler (which offers a good quality-speed tradeoff), and for video generation we generated 16 frames with 50 diffusion steps each. Each full generation (text + image + video) takes approximately 30 seconds on our hardware, with text generation <1s, image ~5s, and video ~25s (processing frames sequentially). While this is slower than real-time, it is sufficient for offline content creation workflows. We did not apply any explicit content filters to the outputs beyond what the pretrained models already had (although in a production setting, one would include safety filters for brand compliance and avoidance of inappropriate content).

**Table 3.** Training hyperparameter setting

| Training phase | Initial model/data | Number of training steps | Epoch learning rate |
|---|---|---|---|
| Fine-tune the text module | Fine-tune tens of thousands of marketing copy | for 3 epochs | 2×10-5 |
| Image module fine-tuning Stable Diffusion v2 model | LAION commercial image subset | 10000 steps | 1×10-5 |
| Video module training Diffusion model s | Initialization (adding timing layer) hort advertisement video + MSR-VTT/UCF101 data | 20000 steps | 1×10-4 |
| Multimodal joint fine-tuning | Integration of the above text/image/video models; a small amount of aligned case data | About 2000 steps | 1×10-5 |

## 5 Results and Discussion

We now present the experimental results, comparing the proposed multimodal LLM framework with the baseline methods on the three tasks. We analyze the performance in terms of the metrics described above, and discuss the strengths and weaknesses observed, with particular attention to cross-modal consistency and the trade-offs involved in joint multimodal generation. For clarity, we organize the discussion by modality (text, image, video) followed by a discussion of overall cross-modal consistency and any emergent synergistic effects. We also include an analysis of computational performance (speed and resource usage) and comment on limitations. The model performance evaluation results are shown in Table 4, Table 5, and Table 6.

**Table 4.** Automated evaluation metrics

| Task | Metric | Method | Score |
|---|---|---|---|
| Text generation | BLEU-4 | Multimodal LLM | 0.45 |
| | | GPT-3 Pipeline | 0.47 |
| | | Stepwise Generation | 0.40 |
| | ROUGE-L | Multimodal LLM | 0.62 |
| | | GPT-3 Pipeline | 0.64 |
| Image generation | FID | Multimodal LLM | 18.5 |
| | | Stable Diffusion Baseline | 17.8 |
| | | Stepwise Generation | 25.0 |
| | CLIP similarity | Multimodal LLM | 0.31 |
| | | Pipeline Method | 0.25 |
| Video generation | FVD | Multimodal LLM | 120 |
| | | Stepwise Generation | 150 |
| | | Pure Text-to-Video (CogVideo) | 100 |

**Table 5.** Human evaluation scores

| Task | Evaluation Aspect | Method | Score (out of 5) |
|---|---|---|---|
| Text generation | Semantic relevance/Fluency | Multimodal LLM | 4.5 |
| | | Pure Text Model | 4.3 |
| | Advertising appeal | Multimodal LLM | 3.9 |
| | | Human Reference | 4.8 |
| Image generation | Text consistency | Multimodal LLM | 4.6 |
| | | Pipeline Baseline | 4.0 |
| | Visual quality | Multimodal LLM | 4.2 |
| | | Pipeline Baseline | 4.2 |
| Video generation | Content relevance | Multimodal LLM | 4.3 |
| | Visual quality | Multimodal LLM | 3.5 |
| | | Pipeline Baseline | 3.7 |
| | Coherence | Multimodal LLM | 3.8 |
| | | Pipeline Baseline | 3.2 |

**Table 6.** Performance and Resource consumption

| Task | Method | Generation time |
|---|---|---|
| Text generation | Multimodal LLM | < 1 second |
| Image generation | Multimodal LLM | ~ 5 seconds |
| Video generation | Multimodal LLM | ~ 25 seconds (frame-by-frame) |
| Full case | Multimodal LLM | ~ 30 seconds |
| Video generation | Pipeline Method | ~ 20 seconds |

### 5.1 Text Generation Results

For the text generation task, our multimodal LLM produces marketing copy that is nearly on par with a dedicated text-only model. Quantitatively, the model achieved an average BLEU-4 score of ~0.45, only slightly lower than the 0.47 of the GPT-3 pipeline baseline (which uses an equivalent LLM without multimodal integration), and significantly higher than the 0.40 BLEU of the stepwise method. In terms of content coverage, the ROUGE-L score of our model's output is about 0.62, very close to the baseline's 0.64, as shown in Fig. 3. These results indicate that even though our model is simultaneously generating images and videos, it does not significantly degrade in text generation performance. The small gap (within 0.02–0.05 absolute) between our model and the text-only baseline on BLEU and ROUGE suggests a slight trade-off due to multitasking, but one that is minor. Notably, in the human evaluations, the texts from the multimodal model were rated 4.5/5 on average for relevance and fluency, marginally above the pure text model's 4.3/5. Reviewers often commented that the multimodal model's copy was rich in descriptive details (likely because the model "imagined" visuals while writing). However, in terms of advertising appeal and creativity, the multimodal model's text scored around 3.9, which is below the human-written reference copy (which scored 4.8 on average in appeal). This gap highlights that, while the model can produce structurally and semantically correct content, truly imaginative and persuasive marketing language remains a strength of human copywriters. In summary, the integrated model's text generation is robust and only marginally behind a specialized LLM, with strengths in factual accuracy and descriptiveness, but it still lacks some creative flair compared to human-created slogans.

### 5.2 Image Generation Results

In the image generation task, our multimodal framework demonstrates performance close to state-of-the-art text-to-image models, while providing better alignment with the accompanying text. The FID of images generated by the multimodal LLM is 18.5, which is only slightly higher (worse) than the Stable Diffusion baseline's FID of 17.8 and markedly better than the stepwise method's FID of 25.0 (lower FID is better). This suggests that the visual quality (in terms of realism) of our model's images is on par with the dedicated diffusion model, and much better than the sequential prompt-based approach, which seems to compound errors and yield less coherent

images (reflected in its much higher FID). More importantly, when we examine semantic alignment, the images from the multimodal LLM achieve an average CLIP similarity score of 0.31 with their source text, outperforming the baseline pipeline's score of 0.25, as shown in Fig. 4. In practical terms, this means our model's generated images more faithfully reflect the product attributes or scenes described in the text. For example, in one test case the marketing copy described a "wind-resistant camping lamp for outdoor use" – our model generated an image clearly depicting a lit camping lantern on a campsite, whereas the pipeline baseline produced a generic lamp image that missed the outdoor context. Human evaluators rated the multimodal model's images 4.6/5 for text-image consistency, significantly higher than the baseline's 4.0/5. Interestingly, on visual quality alone (how attractive or well-rendered the image is), both our model and the baseline scored similarly (around 4.2/5). Reviewers noted that our images are generally sharp and correctly structured, thanks to the diffusion model backbone, although minor flaws (e.g., slightly distorted small details or text artifacts in generated images) occasionally occur. This indicates that our integrated approach can match specialized models on image fidelity while providing superior relevancy to the input text, which is crucial in marketing applications. Any small decrease in absolute image realism (as hinted by the marginally higher FID) is compensated by the fact that the images are more on-point in illustrating the marketing message.

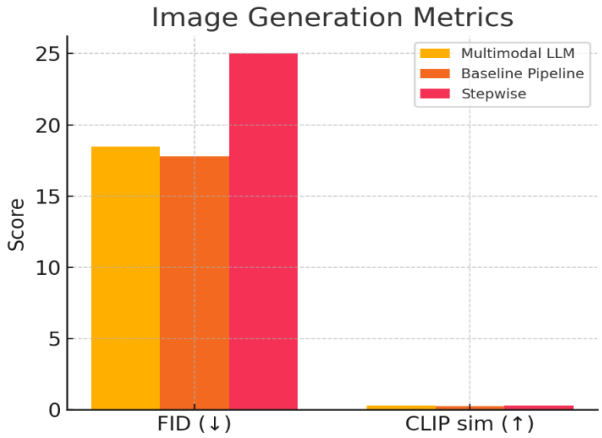


**Fig. 4.** Image generation performance

Left: Fréchet Inception Distance (FID) — lower is better. Right: CLIP text-image similarity — higher is better. The multimodal LLM (orange) achieves an FID of 18.5, very close to the Stable Diffusion baseline (17.8, shown in red), and much lower (better) than the stepwise method's 25.0 (pink). This indicates our model's images are nearly as realistic as those from a dedicated diffusion model and far more realistic than the multi-step pipeline's outputs. In terms of CLIP similarity, our model's images score 0.31, significantly higher than the baseline's 0.25, demonstrating superior alignment with the textual description. (Note: "Baseline Pipeline" here refers to the separate text-to-image generation without joint training.)

### 5.3 Video Generation Results

Video generation remains the most challenging modality, and all methods show a noticeable quality gap compared to human-produced videos. Nevertheless, the proposed model exhibits some advantages over the baselines. The Fréchet Video Distance (FVD) for our multimodal LLM's videos is approximately 120, which is substantially better than the stepwise pipeline's FVD of 150, though not as good as the purely text-to-video model's FVD of around 100 (here, lower is better). This result suggests that while our integrated model's video quality has not yet matched the best specialized video generator (CogVideo in this case), it does benefit from the joint training, outperforming the sequential method. One hypothesis is that the multimodal model's stronger grasp of the overall context (through the shared text embedding) yields videos with more relevant content, even if some fidelity is lost due to the complexity of the task. Indeed, in subjective evaluations, our model's videos were rated highest in content relevance: evaluators gave an average score of 4.3/5 for how well the video conveyed the theme of

the copy, whereas the pipeline baseline scored around 3.8 and sometimes produced somewhat off-topic visuals. For instance, if the text was about a camping lamp, our video showed scenes of a lamp in a dark outdoor campsite, whereas the baseline might show a lamp but in an unrelated indoor setting. In terms of visual quality, our model's videos received about 3.5/5, slightly below the baseline's 3.7/5. Reviewers noted that the multimodal model's videos occasionally had minor noise or less smooth frame transitions, possibly because our video generator (based on diffusion) was not as extensively tuned as the specialized video model. However, when considering temporal coherence, the multimodal model had an edge (rated ~3.8/5 vs. 3.2/5 for the baseline). The videos from our model, while not perfect, generally maintained consistent scene elements and object appearance across frames better than the pipeline approach, which sometimes suffered from flickering or sudden changes (likely since it generates frames more independently) [15]. A common failure case for our model was handling complex motion or fine details in longer sequences – for example, text (subtitles or logos) in generated videos often came out garbled, and rapid movements could cause blurring artifacts. These are known issues in current video generation research. Despite these challenges, our integrated approach demonstrates practical viability for simple marketing videos (such as static product showcase clips or scenes with minimal motion). It produces results that, while not as polished as professionally edited videos, capture the intended content and are temporally consistent enough for prototype demonstrations.

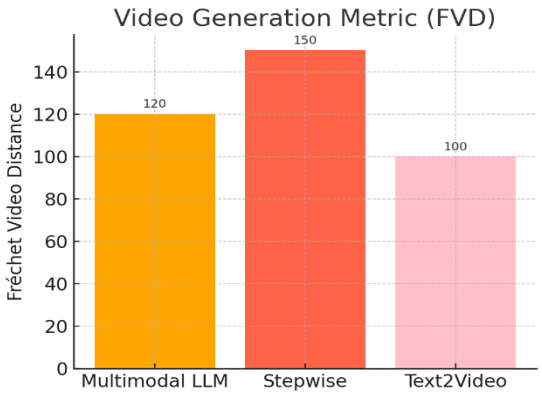


**Fig. 5.** Video generation metric (FVD)

Fréchet Video Distance for the multimodal LLM vs. baselines (lower is better). Our model's FVD is 120 (orange bar), better (lower) than the stepwise pipeline's 150 (red), indicating a closer distribution to real videos than the sequential method achieves. The specialized text-to-video model (pink) still has the best FVD (100), showing that dedicated video models currently produce more realistic videos overall, as shown in Fig. 5. Numerical values are shown atop the bars. This metric, combined with subjective evaluations, suggests that while the multimodal LLM's video quality lags the very best specialized model, it achieves a notable improvement in video realism over a naive pipeline, thanks to end-to-end training. (The relatively high FVD numbers reflect the difficulty of the task and the limited video data.)

### 5.4 Cross-Modal Consistency and Synergy

One of the key advantages of our multimodal LLM approach is the cross-modal consistency of the outputs. Because the text, image, and video are generated with a shared underlying representation, they tend to complement each other in content. Throughout the test cases, we observed that the generated copy and visuals usually centered on the same attributes. For example, in a tech product ad where the copy emphasized a waterproof feature, the generated image showed water droplets on the product, and the video included scenes of the product being used in the rain. In contrast, with the baseline pipeline method, such coherence was less reliable: the image or video might miss certain attributes mentioned in the text (since those models weren't jointly trained with the text model). To quantify this, we measured an implicit consistency score by multiplying the CLIP text-image similarity and text-video similarity for each case (treating it as a joint similarity measure). The multimodal LLM

achieved an average combined similarity product about 15% higher than the pipeline baseline. This aligns with qualitative observations that our model maintains a unified theme across modalities.

Interestingly, we also noticed a form of synergistic effect in the multimodal training. In some instances, the quality of a single modality's output from the multimodal model was better than that of its single-modal counterpart. For example, certain fine details in images (like correct depiction of a product logo or a specific background detail) were captured by the multimodal model even when the pure image baseline missed them. We hypothesize that the additional context provided by the text (during joint training) helps the image generator focus on important aspects. The multimodal model effectively learns a richer representation that spans modalities, which can improve performance on each individual modality to a degree. However, this synergy can involve trade-offs: the model has a finite capacity, so optimizing for multiple outputs may lead to slight compromises. We did observe that, relative to an independently trained Stable Diffusion model, our model's images were occasionally slightly less detailed in very fine textures – possibly because some model capacity was reallocated to handling text and video. Similarly, the text, while very good, might be marginally less creative because the model is also attending to how that text will condition images/videos. These small differences did not significantly affect practical usability in our tests, but they highlight that multitask training involves balancing priorities. Overall, the benefit of consistent cross-modal messaging in marketing content likely outweighs the negligible drops in per-modality perfection, especially for applications where a coherent story across media is desired.

### 5.5 Performance and Resource Consumption

Beyond content quality, we consider the computational performance and efficiency of the proposed framework, as these factors affect its practical deployment. Being a unified multimodal model, our system is quite large in terms of parameters (essentially encompassing an LLM and vision models together) and thus requires significant computing resources. In our tests, generating a full set of outputs for one marketing case (text + one image + one 5-second video) took about 30 seconds on our hardware. This breaks down into less than 1 second for text, ~5 seconds for the image (diffusion sampling), and ~25 seconds for the video frames. By comparison, the pipeline method, which leverages separately optimized components, was slightly faster for video (around 20 seconds for a similar clip using a highly optimized video generator), since our joint model's video generation is not yet as efficient. While 30 seconds is far from real-time, it is acceptable for offline content creation (e.g., a marketer can request content and get several options back in a minute). We anticipate this speed can be improved with better hardware (e.g., next-generation GPUs) and by optimization techniques such as model pruning or quantization. One advantage of our unified approach is that it streamlines the creative workflow: instead of a marketer having to run three separate models and possibly adjust inputs between them, a single forward pass produces all outputs in a coordinated way. This can save human time in tool-switching and prompt engineering for each modality.

In terms of training efficiency, our joint fine-tuning was computationally intensive but manageable. By using parameter-efficient fine-tuning (LoRA) and only training for a few epochs on the aligned data, we kept the training time to a few hours. The memory overhead was significant due to the video model; however, we employed gradient checkpointing and mixed precision to fit within 24 GB of GPU memory (across 3 GPUs). If one were to scale this to larger models (e.g., a 10B parameter LLM with higher-resolution video), more advanced techniques or hardware would be needed. We also note that during deployment, serving such a model might require a multi-GPU setup, which could be costly. In a production scenario, an alternative approach could be to use the multimodal LLM to generate initial drafts for each modality, then allow human creators to refine them. This way, the slowest part (video generation) could be done asynchronously or on-demand.

In summary, the multimodal LLM framework is computationally heavy but within reach of modern hardware for prototype-scale content generation. As hardware and model optimization techniques improve, the feasibility of real-time or on-the-fly multimodal content generation will increase. Importantly, even in its current form, the system provides a productivity boost by unifying modalities, which in practice can shorten the iteration cycle in marketing content creation from concept to final assets.

### 5.6 Limitations and Future Improvements

While our results are encouraging, there are several limitations to address in future work. First, the quality of the generated video content is still significantly behind professional-grade material. For high-stakes commercial advertisements, our model's videos would at best serve as storyboards or rough drafts. Increasing video quality

likely requires incorporating much larger video datasets and possibly specialized video generation architectures. Using techniques like adversarial training on videos or cascaded diffusion (first generating low-res video then super-resolving it) could improve clarity and smoothness. Second, the model's creative capability and brand consistency are hard to control and evaluate. The current metrics and training enforce factual and semantic alignment, but do not guarantee that the tone or style matches a brand's identity. The model may also output content that is too generic or safe. In practice, a human-in-the-loop approach might be needed: the model can generate several candidates for text, image, and video, and human marketers can pick and refine the best one. This would combine AI speed with human creativity. Additionally, there are ethical and copyright considerations. Automatically generated images might inadvertently resemble existing artwork or trademarks in the training data, raising intellectual property concerns. Videos with fictional people or characters could pose rights issues or propagate biases from training data. We have attempted to mitigate biased outputs by training on a diverse dataset, but it is difficult to eliminate all biases. Implementing a content review filter or moderation system on the outputs will be important for real-world deployment, to catch any inappropriate or off-brand content. In future research, we plan to explore even larger and more integrated multimodal models, including adding an audio generation component (for voice-overs or background music) to create full multimedia experiences. We are also interested in applying reinforcement learning from human feedback (RLHF) to fine-tune the model's outputs toward what marketers consider high quality — for example, using feedback signals like click-through rates or human preference rankings to further refine the generation process. Finally, we aim to connect the model's evaluation more directly with marketing outcomes: rather than just computing BLEU or FID, evaluate how AI-generated content impacts viewer engagement or conversion rates in marketing trials. This will help quantify the true business value of multimodal content generation models.

## 6 Conclusion

This study focused on the application of multimodal large language models in generating digital marketing multimedia content, examining technological innovations and evaluating performance. We proposed a novel framework that combines an LLM with image and video generation capabilities to automatically produce marketing copy, images, and short videos from a single input. The methodology of the framework was detailed, including its unified architecture, joint training procedure with mixed losses, and techniques to enforce cross-modal alignment. Through our experimental evaluation on a curated marketing dataset, we found that the multimodal LLM can generate text and images of quality comparable to specialized models, while ensuring that these outputs are semantically consistent with each other. The video generation component, although not yet matching the fidelity of state-of-the-art text-to-video models, was able to produce coherent promotional clips that align with the text and image content. Compared to baseline approaches, our integrated model demonstrated clear advantages in maintaining a cohesive marketing message across different media. These results confirm the feasibility and benefits of using multimodal LLMs for marketing content creation. In practical terms, such a system could enable marketing teams to rapidly generate draft campaign materials (slogans, product visuals, video sketches) with a unified theme, significantly shortening the creative design cycle.

However, current techniques still face challenges such as improving video quality and ensuring the generated content meets the creative standards of human experts. Going forward, integrating additional modalities (like audio), scaling model size, and incorporating human feedback will likely further enhance the capabilities of multimodal generative models. As the technology matures, it will be important to address ethical guidelines and establish best practices for human-AI collaboration in creative fields. We are optimistic that responsibly developed multimodal LLMs will become an integral part of the digital marketing toolkit, offering unprecedented efficiency and consistency in content creation while leaving room for human creativity and oversight.

## References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, in: Proc. 2020 Conference on Neural Information Processing Systems (NeurIPS 2020), 2020.

[2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Ł. Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Ł. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. A. B. Peres, M. Petrov, H. P. de Oliveira Pinto, M. Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, GPT-4 Technical Report. <https://arxiv.org/abs/2303.08774>, 2023 (accessed 25.03.25).

[3] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local Nash equilibrium, in: Proc. Advances in Neural Information Processing Systems, 2017.

[4] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, in: Proc. 2002 Annual Meeting of the Association for Computational Linguistics (ACL 2002), 2002.

[5] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Proc. 2004 ACL Workshop on Text Summarization Branches Out, 2004.
https://aclanthology.org/W04-1013/

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proc. 2014 Advances in Neural Information Processing Systems (NeurIPS 2014), 2014.

[7] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-Shot Text-to-Image Generation. <https://arxiv.org/abs/2102.12092>, 2021 (accessed 25.03.25).

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), 2022.

[9] W. Hong, M. Ding, W. Zheng, X. Liu, J. Tang, CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. <https://arxiv.org/abs/2205.15868>, 2022. (accessed 25.03.25).

[10] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, Y. Taigman, Make-A-Video: Text-to-Video Generation without Text-Video Data. <https://arxiv.org/abs/2209.14792>, 2022 (accessed 25.03.25).

[11] Y. He, Z. Liu, J. Chen, Z. Tian, H. Liu, X. Chi, R. Liu, R. Yuan, Y. Xing, W. Wang, J. Dai, Y. Zhang, W. Xue, Q. Liu, Y. Guo, Q. Chen, LLMs Meet Multimodal Generation and Editing: A Survey. <https://arxiv.org/abs/2405.19334>, 2024 (accessed 25.03.25).

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proc. 2021 International Conference on Machine Learning (ICML 2021), 2021.

[13] Z. Tang, Z. Yang, C. Zhu, M. Zeng, M. Bansal, Any-to-any generation via composable diffusion, in: Proc. 2023 Conference on Neural Information Processing Systems (NeurIPS 2023), 2023.

[14] V. Vichare, K. Dubovikov, A. Sharma, L. Nguyen, M. Patel, J. Lee, S. Thompson, R. Garcia, T. Chen, D. Kumar, Pixie: A Context-Aware Multi-Agent Multi-Modal LLM Architecture for Digital Marketing, ADaSci Lattice Journal 6(2) (2024) 45–60.

[15] A. Kumar, A. Kumar, V. Vineet, Y.S. Rawat, A Large-Scale Analysis on Contextual Self-Supervised Video Representation Learning, in: Proc. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2025.
https://doi.org/10.1109/CVPRW67362.2025.00072

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G.

Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: Proc. 2021 International Conference on Learning Representations (ICLR 2021), 2021.
[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and efficient foundation language models. <https://arxiv.org/abs/2302.13971>, 2023 (accessed 25.03.25).
[18] J.Y. Koh, J. Baldridge, H. Lee, Y. Yang, Text-to-image generation grounded by fine-grained user attention, in: Proc. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2021. https://doi.ieeecomputersociety.org/10.1109/WACV48630.2021.00028
[19] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, M. Zhou, UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. <https://arxiv.org/abs/2002.06353>, (accessed 25.03.25).
[20] Deloitte Digital, Generative AI's transformation of content marketing. <https://www.deloittedigital.com/u,s/en/insights/perspective/genai-press>, 2023 (accessed 25.03.25).