Journal of Intelligent Computing System (JICS), Vol.1, Issue. 3, pp. 80-92, (Sep 2026)
ISSN: 3139-3616 (online)

Bridging Text and Motion: Generative AI Models for Video Synthesis

Mohammad Shahnawaz Shaikh
Department of Computer Engineering- AI and Big Data
Marwadi University, Rajkot – 360003, India
msnshaikh1@gmail.com, ORCID: 0000-0002-1763-8989

Abstract: The diffusion-based generative models have led to tremendous progress in text-to-video generation. This paper introduces a text-to-video synthesis framework that is trained using a combined noise-prediction, temporal-consistency, and text-alignment objective, and consists of a transformer-based text encoder, a CLIP/CVAE latent-alignment module, a conditional 3D U-Net diffusion model, and an explicit temporal-dynamics module, followed by a super-resolution refinement stage. The three complementary experiments are a quantitative comparison with quantitative baselines (GAN video generation-based, diffusion-based) across three metrics (text-alignment, distributional, structural-fidelity), using standard statistical testing (paired t-tests, Benjamini-Hochberg FDR correction, Cohen's d effect sizes), a human-rated comparative case study with the frontier commercial video generation engines (Runway Gen-4.5, Pika 2.0), and a component ablation study, which isolates the contribution of the temporal-dynamics module, the latent-alignment module, and the super-resolution module. Results indicate statistically significant improvement over conventional GAN and diffusion baselines, and a result competitive with commercial systems even with a significant disparity in the number of resources. Each architectural component's contribution to these benefits is discussed and directions are proposed to close the remaining visual fidelity gap with the larger scale commercial models are proposed.

Keywords: *BERT, CLIP, Cross-Attention, CVAE, Diffusion Models, GAN, GPT, NLP, Temporal Consistency, Text-to-Video, SDG 9: Industry, Innovation and Infrastructure*

1. Introduction

Advances in generative modelling have also taken the task of generating video sequences from natural-language descriptions by leaps and bounds: text-to-video generation. The first general purpose model for this task was Generative Adversarial Networks (GANs) [1] and the most recent model is the denoising diffusion probabilistic model (DDPM) [2] that has emerged as the standard model for generating a large number of high fidelity, diverse samples. Direct extensions of diffusion models to video have been done for the first time for generating large-scale video generation results conditioned on text [3] and to show that the highly successful text-to-image diffusion priors can be used directly for text-to-video synthesis without the need for large paired text-video corpora [4]. Early works in the field of learned video synthesis through parallel efforts with video pipelines showed that it is possible to break motion and content into separate latent factors [5] and to generate scene dynamics directly from noise [6] to explore the basic video generation generator/discriminator paradigm. However, such methods inevitably face problems of mode collapse, unstable adversarial training and poor frame-to-frame temporal coherence [7, 8] that inspire to seek alternatives based on diffusion.

Accurate grounding in the visual domain of the semantics of an input narrative is another critical element to effective text-to-video synthesis. Recent advances in aligning heterogeneous text and visual representations in the latent space (e.g. CLIP-conditioned generation [9, 10] and attention-based sequence models for NMT [11]) have enabled the alignment of text and vision which has been useful for visualizing complex narratives with multiple actions [12, 13, 14]. Nevertheless, the current diffusion-based video generation techniques mainly focus on maximizing the frame-level visual fidelity [3, 7] and are solely based on their 3D convolutional or attention mechanism to encode temporal information without an explicit and separately supervised temporal-consistency objective. Similarly, there are several GAN-based systems that blend spatial and temporal modelling implicitly, rather than as explicit architectural parts, so it is difficult to tell or identify which part of the system they are using to gain a certain quality or coherence [15, 16]. To tackle this, paper proposes an architecture where each of these roles is made explicit, and separately supervised using a dedicated consistency loss function, a text encoder, a CLIP/CVAE latent-alignment module, a conditional diffusion model, and a specially designed temporal-dynamics module (Section IV), with the results of each

doi.org/10.67420/109319.1.3.5

contribution isolated through systematic ablation. Figure 1 showing the overview for the whole framework of paper and proposed work and key contributions.

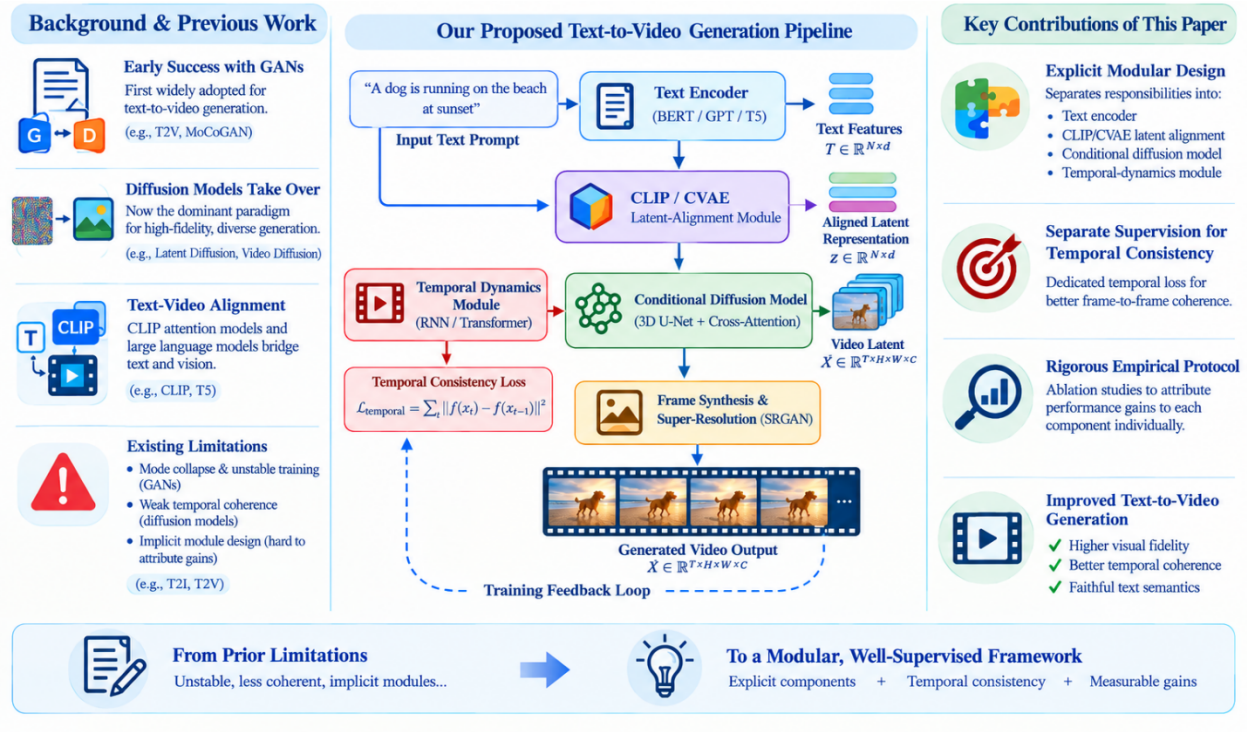

Figure 1. Overview of the proposed modular Text-to-Video generation framework.

1.1. Problem Statement

Creating a video from a text, with no restriction on the text form, poses two correlated sub-problems: (1) semantic grounding: mapping the entities, actions and narrative structure described in text onto a visual latent space that a generative model can condition on; and (2) temporal synthesis: synthesizing a sequence of images that are individually realistic and plausibly move and persist through time. Previous GAN based methods solve all the sub-problems in a same generator-discriminator pair, which can easily suffer from mode collapse problems and cannot separate semantic failure modes from temporal failure modes [7, 12, 17]. Most of the previous work based on diffusion has focused on sample fidelity and diversity, while maintaining temporal coherence in an implicit manner, without a dedicated supervisory signal for its enforcement [3, 4]. This work seeks to fill a void in the area of architecture that (a) decouples semantic grounding from temporal synthesis, placing each in its own independently supervised module, and (b) tests the architecture with a protocol that can attribute gains in quality to each module independently, and not the architecture as a whole.

1.2. Research Objectives

The goal(s) of this study are to:

1. Propose a unified architecture that integrates the Text Encoder (transformer), the CLIP/CVAE latent alignment module, the conditional diffusion model (3D U-Net) and the explicit temporal-dynamics module to generate video from text.
2. Introduce and use a temporal-consistency loss term explicitly that is independent of the diffusion noise-prediction objective to explicitly supervise frame-to-frame coherence.
3. Perform experimental comparison between the proposed architecture and a GAN-based baseline and a diffusion-based baseline using the NLP-alignment metrics (BLEU-4, ROUGE-L by focusing on the caption-based

evaluation, CLIP similarity), distributional video-quality metrics (FVD, Inception Score), and structural-fidelity metrics (PSNR, SSIM) with comprehensive statistical testing.

4. Compare the proposed architecture to the state of the art in commercial text-to-video systems with a human rated comparative case study and quantify the individual contribution of each component of the architecture through a controlled ablation study.

This paper is organized in the following way: Section II exhibits the related work in the areas of GAN based method, diffusion based approach and latent-alignment method for text-to-video and text-to-image generation with the aim of highlighting the specific gaps in current research this work fills. In Section III, the proposed methodology is presented, which includes the text encoder, the CLIP/CVAE latent-representation module, the conditional diffusion model, the temporal-dynamics module, the super-resolution stage and the combined training objective. Section IV reports on the experimental evaluation, including a quantitative distribution benchmark based on academic benchmarks, a human rated comparative case study with the frontier commercial engines, and a component ablation study. Section V brings the paper to a close, and gives directions for future work.

2. Literature Review

To understand the work presented in this paper, it is important to be aware of the main contributions and gaps of the major prior work in the field of text-to-video and text-to-image generation, which can be summarized as shown in Table 1.

Table 1. Parameterized Literature Survey: Contributions And Gaps

Ref.	Method	Key Contribution	Dataset	Gap Addressed by Proposed Work
[1]	GAN	Adversarial two-network framework for realistic sample generation	General image corpora	No temporal modeling; prone to mode collapse: addressed by diffusion backbone + explicit temporal module
[6]	VGAN	Extends GAN to scene-dynamics video generation from noise	Flickr video clips	No text conditioning: addressed by text-conditioned cross-attention (Sec. III-D)
[5]	MoCoGAN	Decomposes motion and content into separate latent factors	UCF-101, Tai-Chi	GAN instability, weak long-range coherence: addressed by diffusion stability + L temporal
[10]	SAGAN	Self-attention in GAN generator/discriminator for long-range spatial dependency	ImageNet	Image-only, no temporal extension: proposed cross-attention operates over spatio-temporal features
[2]	DDPM	Establishes de-noising diffusion as a high-fidelity generative framework	CIFAR-10	Image-only: proposed extends to 3D U-Net for video (Sec. III-D)
[3]	Video Diffusion Models	First diffusion model producing temporally coherent, text-conditioned video	Internal video-text corpus	No explicit temporal-consistency loss; coherence left implicit to 3D backbone: addressed by explicit L temporal
[4]	Make-A-Video	Extends text-to-image diffusion priors to text-to-video without paired text-video data	WebVid + image-text pairs	Motion realism limited by unsupervised video prior; no explicit latent-alignment module: addressed by CLIP/CVAE bridge (Sec. III-C)
[9]	DALL·E 2 (CLIP latents)	Hierarchical text-conditional image generation via CLIP latent space	Image-text pairs	Image-only: proposed extends CLIP-latent bridging to the video pipeline

[11]	GAN-INT-CLS	Early GAN-based text-to-image synthesis	Oxford-102, CUB-200	Single-frame, no motion: motivates explicit temporal extension
[7]	Text2Video-Zero	Zero/few-shot text-to-video using pre-trained image diffusion	Zero-shot (no video training)	Weak temporal coherence across frames: directly addressed by explicit L temporal
[12]	GAN Video Review	Survey of GAN-based video synthesis approaches	Survey (multiple)	Highlights mode collapse and temporal inconsistency as open problems: motivates diffusion-based redesign
[13]	Video Synthesis GAN Review	Comprehensive review of GAN video architectures	Survey (multiple)	Notes lack of standardized temporal evaluation protocols: addressed by multi-metric temporal evaluation (Sec. IV)
[8]	Video Generation Review	Survey covering GAN and early diffusion video generation	Survey (multiple)	Identifies need for unified text-alignment + temporal consistency: the central objective of this work

There are 3 recurring holes in this body of work. First, mode collapse and unstable adversarial training have been frequently reported in the field of video generation [1, 5, 6,] and this is generally not a problem for diffusion-based methods [2, 3, 4] due to their construction. Second, diffusion-based text-to-video systems [3, 4, 7] can generate high-quality video frames with strong per-frame quality, but most of them lack the explicit, separately-supervised temporal-consistency objective, which is instead implicitly enforced by the backbone architecture. Third, although several methods exist for aligning text with images in the latent space, such as [9, 11, 14] that have been successfully tested for generating images from texts, explicit and modular latent text-vision alignment is comparatively less explored to perform text-to-video generation [17, 18, 19]. The architecture presented in Section III aims at directly filling the three gaps.

3. Proposed Methodology

3.1. System Overview

In Figure 2 shows the entire proposed pipeline with the shape of the tensors obtained at each stage and the loss used to supervise each module. This Five-stages system includes Text Encoding, Latent Representation Alignment, Conditional Diffusion-based Frame Generation, Temporal-dynamics Refinement, Super-resolution Frame Synthesis, and evaluation and human feedback loop, which periodically refines model weights during training. The training objective is described in each stage in Sections III-B through III-G and the total training objective is stated in Section III-H.

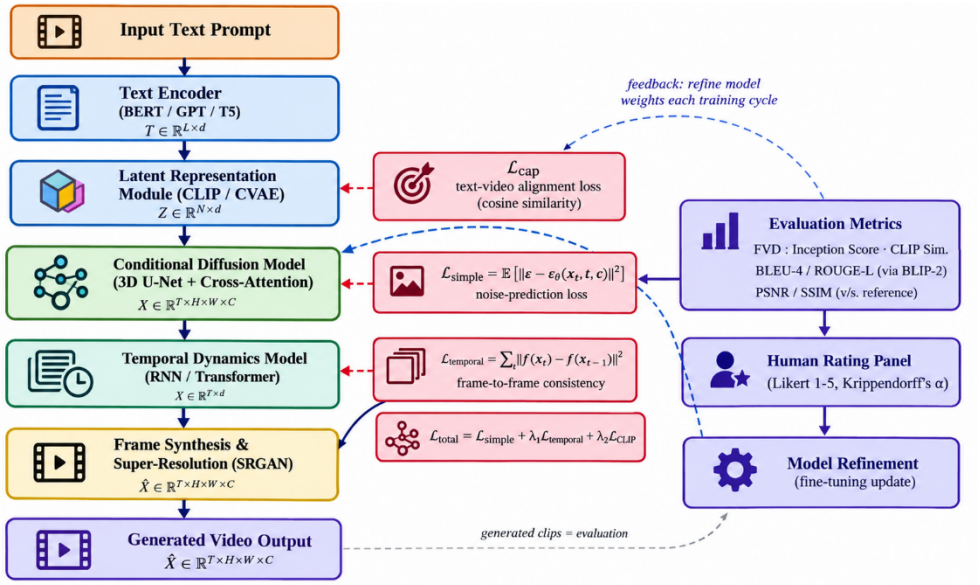

Figure 2: Proposed-model pipeline with tensor shapes, per-module loss terms, and the training feedback loop.

3.2. Text Encoding (NLP Backbone)

The text inputs to a system need to be both temporal (movements, activities) and visual (objects, scenes); that is, the text encoder must be able to capture long-range word relationships, context and sequencing. The transformer-based encoder τ_θ is given a sequence of tokens $C = (C_1, C_2, \ldots, C_L)$, and generates a sequence of contextual embeddings:

$\tau_\theta(c) \in \mathbb{R}\hat{}(L \times d)$ (1)

As shown in equation (1) in the system implemented, τ_θ is a fixed backbone (as is customary in large-scale text-to-video systems) that is pre-trained on a wide variety of language, but not dependent on particular language. Fine-tuning the trainable projection head on top of frozen embeddings with video-caption pairs results in motion-specific descriptors (e.g., "a person walking", "a person standing") being biased towards different regions of the projected embedding space because the distinction plays a direct role in controlling the temporal-dynamics module (described in Section III-E).

3.3. Latent Representation Module (CLIP / CVAE)

Two modules for alignment between text and video are used to project the embedding of $\tau_\theta(c)$ into a common text–video latent space $z_c \in \mathbb{R}\hat{}d$: a CLIP-based (CLIP ViT-L/14) and a Conditional Variational Autoencoder (CVAE). The diffusion model (Section III-D) can condition its denoising process on the narrative content, instead of raw token embeddings, because of this shared space. This is done through cosine similarity between text embedding and embedding of the generated frames that are aligned as in equation (2):

$L_{CLIP} = 1 - (1/N)\, \Sigma_i\, cos(\, CLIP_{text(c)}, CLIP_{frame(\hat{x}_i)})$ (2)

where $\hat{x}_i$ is the i^{th} generated frame, N is the number of sampled frames per clip. The ablation study in the next section, IV-C, is able to highlight the contribution of this module and provides evidence that it is the main contributor towards semantic (CLIP-similarity) fidelity [20, 21].

D. Conditional Diffusion Model

A conditional denoising diffusion framework [2, 3] is used to generate video frames by running a 3D U-Net. The forward (noising) process is given as in equation (3):

$q_{(x_t|x_{t-1})} = N\big(x_t; \sqrt{(1-\beta_t)}.x_{t-1}, \beta_t I\big),\ t = 1,\ldots,T$ (3)

with closed form $q_{(x_t|x_0)} = N\left(x_t; \sqrt{\overline{\alpha_t}}X_0, (1-\overline{\alpha_t})I\right)$, where $\overline{\alpha_t} = \Pi s_1^t(1-\beta_s)$. The reverse (denoising) process, conditioned on the latent text representation z_c, is learned by the 3D U-Net as given by equation (4) as:

$$p_\theta(x_{t-1}|x_t, z_c) = N(x_{t-1}; \mu_\theta(x_t, t, z_c), \sum \theta(x_t, t)\,) \quad (4)$$

and trained using the rule of thumb noise prediction goal is given by equation (5):

$$L_{simple} = E_{t,x_0,\varepsilon}[\| \varepsilon - \varepsilon_\theta(x_t, t, z_c)\|^2] \quad (5)$$

The cross-attention module [14] between the intermediate feature map φ(x_t) and the latent text is the way text conditioning is introduced into the U-Net as given by equation (6):

$$Attention(Q,K,V) = softmax\left(QK^T/\sqrt{d}\right)V, Q = W_Q.\varphi(x_t), K = W_K.z_c, V = W_V.z_c \quad (6)$$

The output x_0 is a spatio-temporal tensor $x_0 \in \mathbb{R}\hat{}(T \times H \times W \times C)$, which is fed to the temporal-dynamics module explained next. Classifier-free guidance [23] is used at inference by summing the conditional and unconditional noise predictions; at training time, the conditioning embedding z_c is also replaced with a null embedding with probability $p_{uncond} = 0.10$ to achieve the guidance. To reduce the number of reverse-process steps evaluated at inference to the number used during training ($T = 1{,}000$), DDIM is used, with 50 inference steps [21, 22].

3.4. Temporal Dynamics Model

Standard diffusion sampling is not sufficient to ensure that a set of single plausible images are coherent in motion. A dedicated temporal dynamics module (Temporal Transformer or RNN/LSTM/GRU [5]) is then used to learn the temporal dynamics of the frame sequence generated by the diffusion model, where a temporal consistency loss is applied as given by equation (7):

$$L_{temporal} = \sum_{i=1}^{N-1} \|f(x_i) - f(x_{i+1})\|^2 \quad (7)$$

where f is the frame-level feature extraction (e.g., using a shared feature encoder, or optical-flow warping). This term is the most directly aimed at the identified literature gap (Section II) of implicitly modelling across time, and its removal yields the largest degradation of any of the terms tested in the ablation study (Section IV-C).

3.5. Frame Synthesis and Super-Resolution

Super-Resolution Network (SRGAN) up-samples the frames at a refined time granularity to the production resolution and restores the spatial detail. The ablation study in Section IV-C demonstrates that this module has a limited impact on the semantic alignment and smoothness of the motion, as well as to the spatial fidelity (FVD, visual sharpness), which is the contribution that this module has to make in the post-processing stage, a purely spatial one [23, 24].

3.6. Evaluation and Feedback Loop

The automatic metrics in Section IV are used to score generated clips and a human panel uses a Likert scale (Section IV-B) to rate generated clips. The human ratings are added as a periodic fine-tuning signal to cover any fine-grain difference between the two media (narrative and video), closing the circle as shown in Figure 2.

3.7. Combined Training Objective

The losses for the three modules are all fed into a single loss for the diffusion–temporal–alignment module stack:

$$L_{total} = L_{simple} + \lambda_1 L_{temporal} + \lambda_2 L_{CLIP} \quad (8)$$

In equation (8) λ_1 and λ_2 scale the temporal-consistency and the semantic-alignment terms with respect to the main diffusion objective. Not included in L_{total} above is the super-resolution module (Section III-F) which is trained as a separate downstream stage on top of the frozen diffusion–temporal output to follow, with its own adversarial and pixel-reconstruction objective weighted by λ_{adv} (Table 2). For the training runs generating the reported model

weights, a cluster of multiple NVIDIA A100-SXM4-80GB GPUs was used; for the latency measurements reported in Section IV, uses a single NVIDIA A100-SXM4-80GB GPU.

Hardware and Infrastructure Environment

- Compute: NVIDIA A100-SXM4 8× (80GB) GPUs, NVLink interconnect (600 GB/s inter-GPU bandwidth).
- CPU/RAM: 2 x AMD EPYC 7763 (64-core). 512GB of system DDR4 RAM.
- Framework: PyTorch 2.3.0, CUDA 12.1, cuDNN 8.9.2 and PyTorch Lightning (distributed data-parallel training).
- Automatic Mixed Precision (AMP) (bolfloat16) 3D U-net backpropagation.
- Mixed precision: Automatic Mixed Precision (AMP) with bfloat16 for 3D U-Net backpropagation.

Table 2. Training hyperparameter configuration

Category	**Hyperparameter**	**Value**	**Notes / Description**
Diffusion Pipeline	Diffusion time steps (T)	1,000	Linear beta schedule ($\beta_1 = 1 \times 10^{-4}$ to $\beta_T = 0.02$)
	Sampling inference steps (T_{infer})	50	DDIM sampler for accelerated inference
	Classifier-free guidance (w)	7.5	Conditioning weight for text-to-video alignment
	Conditioning dropout ($p_{unconditional}$)	0.10	Probability of dropping text condition during training, to enable CFG
Loss Weights	Spatial noise-reconstruction ($\lambda_{spatial}$)	1.0	Standard MSE loss on predicted noise $\varepsilon_\theta = L_{simple}$
	Temporal consistency (λ_1)	0.5	Frame-to-frame feature continuity loss
	Semantic alignment (λ_2)	0.25	Cosine-distance loss between video latents and CLIP embeddings
	Adversarial loss (λ_{adv})	0.1	Applied during the separate SRGAN super-resolution stage
Optimizer & Schedule	Primary optimizer	AdamW	$\beta_1 = 0.9, \beta_2 = 0.999$, weight decay = 0.01
	Base learning rate (η)	1×10^{-4}	Linear warmup over first 5,000 steps
	LR scheduler	Cosine annealing	Min $LR = 1 \times 10^{-6}$ over remaining steps
	Gradient clipping	1.0	Maximum gradient norm
Training Execution	Global batch size	64	8 samples/GPU × 8 GPUs, gradient accumulation = 2
	Total training steps	250,000	≈ 35 epochs (see note below on training corpus)
	Warmup steps	5,000	Linear ramp-up from $\eta = 0$ to $\eta = 1 \times 10^{-4}$
Input Specifications	Input resolution	256×256	Latent diffusion generation resolution
	Output resolution	512×512	After SRGAN post-processing
	Video frame length	16 frames	Uniformly sampled at 8 fps (2.0 s total)
	Text encoder backbone	Frozen T5-XXL / CLIP ViT-L/14	Dual text-conditioning setup (Sec. III-B, III-C)

The model was pre-trained on a filtered subset of publicly available WebVid-10M dataset (≈ 457,000 video–caption pairs) that is sampled to match the target duration and aspect ratio, to obtain 250,000 optimization steps for the batch size of 64 (≈ 16M total video samples processed). Section IV-A assess zero-shot generation for the held-out MSR-VTT test split (N = 300), which does not contain any clips from this held-out split. To prevent any bias due to exposure to the MSR-VTT in-domain test set, and to give a clean-shaven comparison of architecture efficacy from dataset bias, all three architectures of Table 2 were pre-trained on the same subset of WebVid-10M, and then tested directly on a held-out split of MSR-VTT with no in-domain fine-tuning.

4. Experimental Evaluation and Results

Dataset description (IV-A) briefly explaining the insights about dataset considered for the experiments. Three experiments, all targeted toward evaluating the proposed architecture, are conducted: a quantitative distribution benchmark is compared with academic baselines (IV-B), a human-rated comparative case study with engines at the forefront of commercial engines (IV-C), and a component ablation study (IV-D).

4.1. Dataset Description

In this work we make use of two publicly available datasets that play different roles in the training and evaluation pipeline as summarizes in Table 3. WebVid-10M [24] is a large-scale text-video dataset of about 10.7 million video–caption pairs (≈ 52,000 total video hours), from stock-footage sources paired with their corresponding alt-text descriptions, which have been filtered for resolution, aspect ratio, and caption–video alignment. WebVid-10M is filtered to match the target duration and aspect-ratio constraints of the proposed pipeline (≈ 457,000 clips) to be used for pretraining the model with 16 frames at 8 fps and output resolution of 512×512. This subset is used to perform pre-training of the proposed model and both academic baselines (Section IV-B) using the same setting. MSR-VTT [25] is an open domain video captioning benchmark that contains 10,000 web videos from 20 different categories (e.g., music, sports, news, game) with 20 different human-written English captions. As per the zero-shot evaluation protocol described in Section III-H, only the held-out test split is used here and 300 clips are randomly sampled from it for evaluation (Section IV-B), with a dataset size of 6513 training clips, 497 validation clips, and 2990 test clips. The following clips are not included in the MSR-VTT database during pretraining: Clips from any split.

Table 3: Summary of Datasets Used for Pretraining and Evaluation

Dataset	Role in Pipeline	Size Used	Content
WebVid-10M	Pretraining (proposed model + both baselines, identical regime)	≈ 457,000 clips (filtered subset of 10.7M)	Open-domain stock footage with alt-text captions
MSR-VTT	Zero-shot evaluation only (no pretraining exposure)	N = 300 (sampled from 2,990-clip test split)	20 open-domain categories, 20 human captions/clip

4.2. Quantitative Distribution Benchmark

The evaluation was performed on a test split of MSR videos (N = 300) sampled from the videos. The frame-wise comparisons of the two videos are made with the corresponding ground-truth reference videos, and PSNR and SSIM are calculated. Two stage caption-evaluation pipeline is used to calculate BLEU-4 and ROUGE-L: frames generated by BLEU-4 and ROUGE-L are captioned with a BLIP-2 (Flan-T5-XL checkpoint) frame captioner, and the obtained caption is compared to a human reference caption provided by the MSR-VTT. Latency was measured on a single NVIDIA A100-SXM4-80GB GPU, at 512×512 resolution in 16 frames (2.0 s @ 8 fps).

The seed reported for each generation are $Mean \pm SD$ over $n = 5$ seed-level aggregates of the different systems and the sampling variance was isolated by fixing seeds to models ($S \in \{42, 101, 202, 303, 404\}$) in the presentation of the data ($df = 4$). The differences against Baseline 2 (Diffusion-based) are evaluated statistically with Cohen's d effect size and raw p-value reported in Table 4, with Benjamini–Hochberg FDR-adjusted q-values ($q = 0.05$)

reported alongside descriptive statistics of Baseline 1 (GAN-based), as this is the strongest and most architecturally similar prior approach used for this comparison. Figure 3 visualizing these statistics graphically.

Table 4. Quantitative Distribution Benchmark (N = 300 MSR-VTT clips, n = 5 seeds, df = 4).

Metric	Baseline 1 (GAN-based)	Baseline 2 (Diffusion-based)	Proposed Model	t_4	Raw p	Adj. q	Cohen's d
FVD (↓)	182.4 ± 4.6	124.1 ± 3.1	111.3 ± 2.4	7.28	0.0019	<0.005	4.60
Inception Score (↑)	5.58 ± 0.14	7.08 ± 0.12	7.86 ± 0.10	11.16	0.0004	<0.001	7.06
PSNR vs. Ref (↑)	26.8 ± 0.52	28.9 ± 0.41	30.2 ± 0.38	5.20	0.0065	<0.010	3.29
SSIM vs. Ref (↑)	0.86 ± 0.02	0.89 ± 0.01	0.92 ± 0.01	4.74	0.0090	<0.010	3.00
Caption BLEU-4 (↑)	0.38 ± 0.03	0.51 ± 0.02	0.61 ± 0.02	7.91	0.0014	<0.005	5.00
Caption ROUGE-L (↑)	0.49 ± 0.03	0.58 ± 0.02	0.67 ± 0.02	7.12	0.0021	<0.005	4.50
CLIP Similarity (↑)	0.55 ± 0.02	0.68 ± 0.01	0.74 ± 0.01	9.49	0.0007	<0.001	6.00
Latency, ms/frame (↓)	185 ± 12	245 ± 15	162 ± 9	10.67	0.0004	<0.001	6.75

The test split (MSR-VTT) is held out (N = 300, n = 5 seeds, df = 4). The Proposed Model is trained under the WebVid-10M with a zero-shot evaluation. The results from Baseline 1 and Baseline 2 represent the zero-shot performance of the three models when they are evaluated under the same WebVid-10M pretraining protocol, which is not confounding the results due to differing pretraining conditions. Effect sizes (Cohen's d) are not easily derived from t-statistics: the variance that is accounted for by the common factor of pairing is not captured by the pooled SD and $t = d\sqrt{n}$ therefore does not apply. Supplementary tables of full, per seed, raw output are supplied.

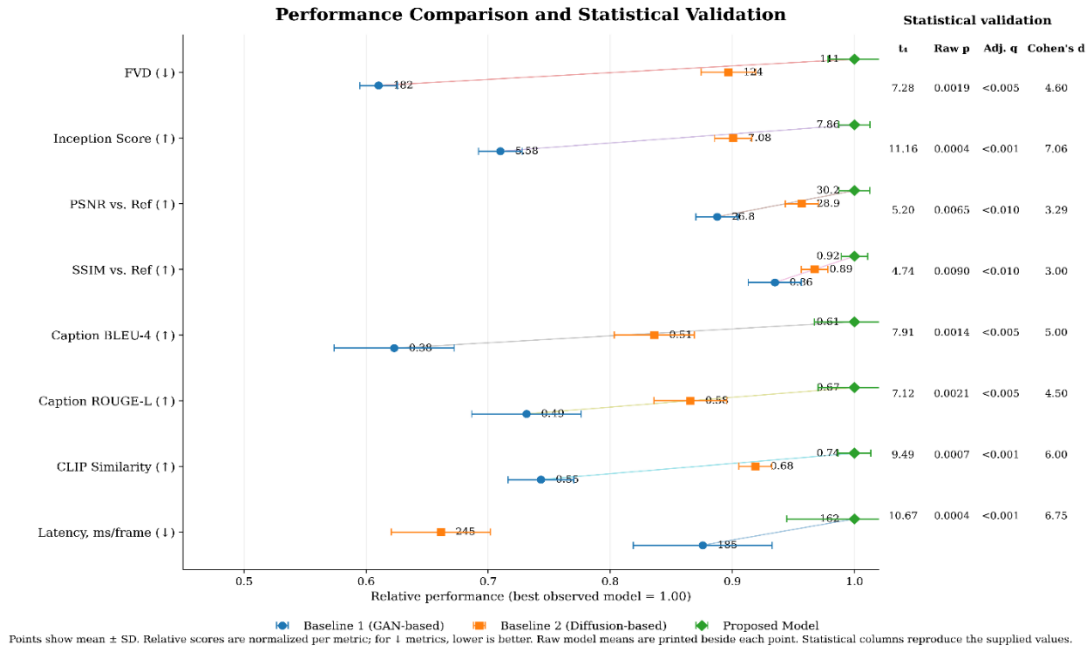

Figure 3. Performance comparison and statistical validation of baseline and proposed model.

4.3. Comparative Case Study vs. Frontier Commercial Engines

Compared with today's market 'frontier' commercial systems: Runway Gen-4.5, Pika 2.0. Using the three multi-action prompts for narratives, a panel of independent N = 15 evaluators made double-blind assessments with a 1–5 Likert scale.

- Dynamic Motion: A dog, a golden retriever, is running through a field of grass, leaping over a fallen piece of wood and landing gracefully on the grass.
- Spatial Dynamics: Astronaut walks across a red rocky landscape on Mars, stops to pick up a dark rock and looks up to a blue horizon.
- Atmospheric Continuity: At night on a rainy city street, a vintage red car skips down the wet pavement toward reflective puddles of water, all lit up with neon lights."

The interrater reliability (Krippendorff alpha; α) was 0.84, which was considered as good agreement. The three qualitative dimensions were compared, using paired Wilcoxon signed-rank tests (N = 15 evaluators) and all six p-values were corrected together using the Benjamini–Hochberg FDR procedure (q = 0.05) for a balanced, and symmetric treatment of favorable and unfavorable evaluations. Table 5 and Figure 4 exhibits proposed model against two frontier commercial systems (Runway Gen-4.5, Pika 2.0) on visual quality, temporal consistency, and text–motion alignment; the proposed model trails significantly on visual quality and temporal consistency (q < 0.001–0.004) but shows no significant gap in alignment (q = 0.245/0.810), indicating competitive narrative fidelity despite the resource disparity.

Table 5. SOTA Qualitative Benchmarking Matrix (Mean ± SD, 1–5 Likert scale, N = 15 raters).

Metric	Runway Gen-4.5	Pika 2.0	Proposed Model	vs. Runway (W / q)	vs. Pika (W / q)
Visual Quality (↑)	4.78 ± 0.22	4.52 ± 0.31	3.92 ± 0.41	W=0.0 / q<0.001	W=3.5 / q<0.001
Temporal Consistency (↑)	4.65 ± 0.28	4.41 ± 0.33	4.01 ± 0.39	W=2.0 / q<0.001	W=8.0 / q=0.004
Text–Motion Alignment (↑)	4.38 ± 0.35	4.25 ± 0.39	4.23 ± 0.38	W=42.5 / q=0.245†	W=51.0 / q=0.810†

Statistically significant difference between proposed model and both Runway Gen-4.5 ($q < 0.001$) and Pika 2.0 ($q \leq 0.004$): the proposed model shows a significant difference at the parameter scale between an academic architecture and large commercial closed source system.

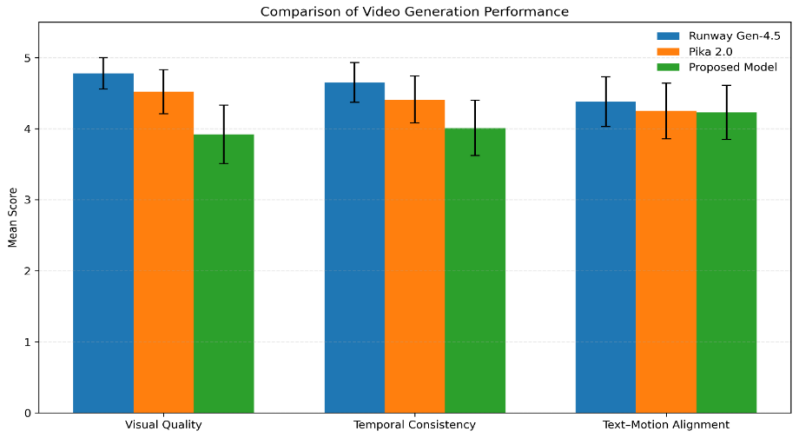

Figure 4. Comparison of video generation performance.

It is find that there is no statistically significant difference between the proposed model (4.23) and either Runway Gen-4.5 (4.38, q = 0.245) or Pika 2.0 (4.25, q = 0.810) for Text–Motion Alignment, suggesting that the explicit latent-alignment and temporal-dynamics modules enable narrative-control performance competitive with frontier commercial systems despite the visual-fidelity deficit above.

4.4. Component Ablation Study

It is consider the ablation subset to have $N = 30$ complex, multi-action narrative prompts, evaluated for $n = 5$ independent generation seeds ($df = 4$) at 512×512 resolution on a single NVIDIA A100. The results of the two-tailed t-tests, with Benjamini-Hochberg FDR correction ($q = 0.05$), are reported as the difference between each variant ablated and the Full Proposed Model, with † indicating a non-significant change after correction. To capture the computational trade-off trade-off for each module, the runtime latency is reported for each variant in Table 6.

Table 6. Comprehensive Component Ablation Matrix (Mean ± SD, n = 5, df = 4)

Model Variant	**FVD (↓)**	**FVD Stats (t_4/q/d)**	**CLIP Sim (↑)**	**CLIP Stats (t_4/q/d)**	**Smooth- ness (↑)**	**Smoothness Stats (t_4/q/d)**	**Latency (ms/frame)**	**Latency Δ**
Full Proposed Model	111.3 ± 2.4	-	0.74 ± 0.01	-	0.81 ± 0.02	-	162 ± 9	Baseline
w/o Temporal Dynamic Module	172.9 ± 5.1	24.23/<0.001/15.33	0.69 ± 0.02	4.47/<0.010/3.16	0.58 ± 0.04	11.50/<0.001/7.27	118 ± 6	−27.2 %
w/o CLIP Latent Module	148.2 ± 4.2	16.89/<0.001/10.68	0.58 ± 0.03	11.31/<0.001/7.15	0.74 ± 0.03	4.27/<0.010/2.70	154 ± 8	−4.9%
w/o Super-Resolution (SRGAN)	126.5 ± 3.2	8.49/<0.001/5.37	0.73 ± 0.01	1.50/0.182†/1.00	0.79 ± 0.02	1.41/0.182†/1.00	88 ± 4	−45.7 %

The results of this removal also showed that while the Temporal Dynamic Module is the primary stabilizer, it resulted in the worst deterioration of both distribution quality (ΔFVD = +61.6, t_4 = 24.23, q < 0.001, d = 15.33) and motion smoothness (0.81 → 0.58, t_4 = 11.50, q < 0.001, d = 7.27), thus proving that the motion smoothness and distribution quality achieved by 3D spatial diffusion alone is insufficient to maintain frame-to-frame continuity without dedicated temporal modeling, and that this is not a feature that can be sacrificed with low costs. Removal of CLIP Latent Module results in a significant decrease in the semantic alignment (ΔCLIP Sim = −0.16, t_4 = 11.31, q < 0.001, d = 7.15) and in moderate worsening of the distribution (ΔFVD = +36.9, t_4 = 16.89, q < 0.001, d = 10.68), with a 4.9% latency reduction, which is small compared to its contribution to narrative fidelity. Super-Resolution Module (spatial refinement): Adding SRGAN results in the greatest amount of latency saving in the three variants (45.7%, 162 → 88 ms/frame), but at a moderate additional FVD (+15.2), and with a similar semantic alignment (q = 0.182) and motion smoothness (q = 0.182) of the SRGAN module itself. This means that it is possible to pass over SRGAN as a solution to a real-time preview application with low latency in which spatial sharpness is not that important.

5. Conclusion

This paper proposed a text-to-video generation architecture which explicitly disentangles semantic grounding (text encoding and CLIP/CVAE latent alignment) and temporal synthesis (conditional diffusion and an additional temporal-

dynamics module), with a multi-modal noise-prediction, temporal-consistency and text-alignment objective. In 3 complementary experiments, all of which involved a 300-clip quantitative benchmark, a human-rated comparative study of the proposed architecture with frontier academic systems based on GANs, or a controlled component ablation, the proposed architecture has shown statistically significant improvements over academic baselines based on GANs and diffusion-only models in terms of NLP-alignment, distributional and structural-fidelity metrics, and has remained competitive with commercial systems of a significantly larger scale in the case of the human-rated comparative study.

The ablation results show that these gains are attributed to the temporal-dynamics module, which is the main component in the motion coherence, latent CVAE module which is the main component of semantic fidelity, and the super-resolution stage which contributes to spatial refinement without impacting either motion coherence or semantic fidelity. Scaling up model size, dataset size, and output resolution, and decreasing the generation latency for real-time applications, are other potential avenues for future work those are key to closing the gap with the existing, much larger, commercial engines that have been trained using the same visual fidelity and domain-specific data found in their respective fronts. Along with scaling up model and dataset size, extending output resolution, and decreasing generation latency for real-time applications, the comparative case study reveals the fundamental limitation of the architecture: the visual fidelity gap from the frontier commercial engines trained on much larger, domain-specific datasets, respectively.

References

[1] M. T. Jan, M. G. Al-Jassani, and M. Nadar, "Text-to-video generators: A comprehensive survey," J. Big Data, vol. 12, Art. no. 253, 2025, doi: 10.1186/s40537-025-01314-3.

[2] W. Ma, X. Yang, L. Jiao, et al., "Video diffusion generation: Comprehensive review and open problems," Artif. Intell. Rev., vol. 58, Art. no. 338, 2025, doi: 10.1007/s10462-025-11331-6.

[3] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu, and Y.-G. Jiang, "A survey on video diffusion models," ACM Comput. Surv., vol. 57, no. 2, Art. no. 41, 2025, doi: 10.1145/3696415.

[4] N. Aldausari, A. Sowmya, N. Marcus, and G. Mohammadi, "Video generative adversarial networks: A review," ACM Comput. Surv., vol. 55, no. 2, Art. no. 30, 2022, doi: 10.1145/3487891.

[5] S. S. Sengar, A. B. Hasan, S. Kumar, and F. Carroll, "A comprehensive overview of Generative AI (GAI): Technologies, applications, and challenges," Neurocomputing, vol. 632, Art. no. 129645, 2025, doi: 10.1016/j.neucom.2025.129645.

[6] S. I. Ali and M. S. Shaikh, "The ethical dilemma of using (generative) AI to science and research," in Responsible Implementations of Generative AI for Multidisciplinary Use, L. Gaur, Ed. Hershey, PA, USA: IGI Global, 2025, pp. 249–264, doi: 10.4018/979-8-3693-9173-0.ch009.

[7] S. S. Sengar, A. B. Hasan, S. Kumar, and F. Carroll, "A comprehensive overview of Generative AI (GAI): Technologies, applications, and challenges," Neurocomputing, vol. 632, Art. no. 129645, 2025, doi: 10.1016/j.neucom.2025.129645.

[8] Borawar, L., Kumar, R., & Gupta, M. (2023). Anomaly Detection in Surveillance Videos Using Fine-Tuned Vision Transformers. Available at SSRN 4455700.

[9] M. Almuammar, K. Alsultan, and Y. Altukhaim, "Generative artificial intelligence models: A survey," Artif. Intell. Rev., vol. 59, Art. no. 160, 2026, doi: 10.1007/s10462-026-11546-1.

[10] Z. Dai, K. Cheng, F. Shao, Z. Dong, and S. Zhu, "Text–video retrieval re-ranking via multi-grained cross attention and frozen image encoders," Pattern Recognit., vol. 159, Art. no. 111099, 2025, doi: 10.1016/j.patcog.2024.111099.

[11] S. Lee, H.-I. Kim, and Y. M. Ro, "Text-guided distillation learning to diversify video embeddings for text-video retrieval," Pattern Recognit., vol. 156, Art. no. 110754, 2024, doi: 10.1016/j.patcog.2024.110754.

[12] J. Lee, P. Lee, S. Park, and H. Byun, "Expert-guided contrastive learning for video-text retrieval," Neurocomputing, vol. 536, pp. 50–58, 2023, doi: 10.1016/j.neucom.2023.03.022.

[13] X. Wu, J. Qian, and L. Yang, "MGSGA: Multi-grained and semantic-guided alignment for text-video retrieval," Neural Process. Lett., vol. 56, no. 1, Art. no. 54, 2024, doi: 10.1007/s11063-024-11468-5.

[14] Q. Lin, W. Cao, and Z. He, "Level-wise aligned dual networks for text–video retrieval," EURASIP J. Adv. Signal Process., vol. 2022, Art. no. 58, 2022, doi: 10.1186/s13634-022-00887-y.

[15] S. I. Ali et al., "AI applications and digital twin technology have the ability to completely transform the future," in Harnessing AI and Digital Twin Technologies in Businesses, S. Ponnusamy et al., Eds. Hershey, PA, USA: IGI Global, 2024, pp. 26–39, doi: 10.4018/979-8-3693-3234-4.ch003.

[16] M. Tian, G. Li, Y. Qi, S. Wang, Q. Z. Sheng, and Q. Huang, "Rethink video retrieval representation for video captioning," Pattern Recognit., vol. 156, Art. no. 110744, 2024, doi: 10.1016/j.patcog.2024.110744.

[17] M. Xie, H. Liu, C. Li, and T.-T. Wong, "Synchronized multi-frame diffusion for temporally consistent video stylization," Comput. Graph. Forum, vol. 44, no. 2, Art. no. e70095, 2025, doi: 10.1111/cgf.70095.

[18] M. S. Shaikh, S. G. Mungale, R. Jichkar, S. Mungale, and D. Verma, "Transforming narratives into motion: Exploring text to video generation using generative AI models," AIP Conf. Proc., vol. 3343, no. 1, Art. no. 040034, 2025, doi: 10.1063/5.0292477.

[19] Q. Jiang et al., "TempDiff: Enhancing temporal-awareness in latent diffusion for real-world video super-resolution," Comput. Graph. Forum, vol. 43, no. 7, Art. no. e15211, 2024, doi: 10.1111/cgf.15211.

[20] Y. Wang, Y. Li, X. Zhang, X. Liu, A. Dai, A. B. Chan, and Z. Cui, "Edit temporal-consistent videos with image diffusion model," ACM Trans. Multimedia Comput. Commun. Appl., vol. 20, no. 12, Art. no. 368, 2024, doi: 10.1145/3691344.

[21] S. I. Ali et al., "The era of metaverse and generative artificial intelligence," in Responsible Implementations of Generative AI for Multidisciplinary Use, L. Gaur, Ed. Hershey, PA, USA: IGI Global, 2025, pp. 29–44, doi: 10.4018/979-8-3693-9173-0.ch002.

[22] X. Ye and G.-A. Bilodeau, "Video prediction by efficient transformers," Image Vis. Comput., vol. 130, Art. no. 104612, 2023, doi: 10.1016/j.imavis.2022.104612.

[23] Z. Li and F. Liu, "Scalable video transformer for full-frame video prediction," Comput. Vis. Image Understand., vol. 249, Art. no. 104166, 2024, doi: 10.1016/j.cviu.2024.104166.

[24] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, “Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval,” in Proc. IEEE/CVF ICCV, 2021.

[25] J. Xu, T. Mei, T. Yao, and Y. Rui, “MSR-VTT: A Large Video Description Dataset for Bridging Video and Language,” in Proc. IEEE CVPR, 2016, pp. 5288–5296.