NBO-SAGAN: Namib Beetle Optimized Self-Attention GAN for Cross-Modal Short Video Generation with Semantic Decoupling

Jie Chen, Dan Li
Sichuan Film and Television University, ChengDu, Si Chuan, 610036, China
E-mail: chenjie199110@126.com, lidan1999528@126.com

Keywords: Short video, diffusion model, semantic decoupling, latent space synthesis, namib beetle optimized self-attention based generative adversarial networks (NBO-SAGAN)

The rapid growth of short video platforms has fueled interest in automated short video generation, a task that demands high fidelity, temporal coherence, and semantic alignment with input data such as text, audio, or static images. This task poses important issues since to the requirement for accurate semantic understanding, spatial-temporal alignment, and realistic visual synthesis. In this research, a novel deep learning-based cross-modal short video generation framework is developed to address these challenges effectively. The Microsoft Research Video to Text (MSRVTT) dataset is gathered from the Kaggle, has 10,000 video clips from 20 categories undergoes extensive preprocessing, tokenization, Audio Modality using Mel-Frequency Cepstral Coefficients (MFCCs), Video Modality using Temporal-Spatial Synchronization and Normalization (TSSN). Then feature extraction is performed with a semantic decoupling module that utilizes encoders to extract and disentangle high-level semantic components, such as object appearance, motion dynamics, background context, and emotional tone from the input. A Namib Beetle Optimized Self-Attention based Generative input in various Adversarial Networks (NBO-SAGAN) is employed to generate short videos and to enhance fine details and correct visual artifacts. Experimental evaluations using Python shows that the NBO-SAGAN approach outperforms traditional methods Frechet Video Distance (FVD) of 230.74 and Kernel Video Distance (KVD) of 12.58 highlighting its effectiveness for expressive cross-modal video generation. This integrated methodology effectively combines modeling to produce controllable, expressive, and visually rich cross-modal video content.

Povzetek: Članek predstavi križno-modalni okvir za samodejno generiranje kratkih videov, ki z razcepom semantike iz besedila/zvoka/slik in Namib Beetle optimiziranim samo-pozornostnim GAN (NBO-SAGAN) izboljša časovno koherenco in vizualno kakovost sinteze.

1 Introduction

With the fast emergence of multimedia applications and user-created content on applications such as Tik Tok, YouTube, and Instagram reels, there is an increasing demand to produce short videos automatically and with high quality [1]. It has inspired a surge in areas of computer vision, natural language handling, and cross-modal generation areas [2]. Short video creation not only involves producing high-quality visual frames but also temporal consistency between these frames so that the experience is smooth and realistic [3]. This task is more complex in a cross-modal scenario when semantic content is to be properly extracted from a different modality and mapped onto a visual and temporal domain [4]. For example, creating a short video from a text prompt like a child running along a sunny beach with a kite entail identifying and representing different semantic cues motion, emotion, lighting, setting, and object interaction [5-6]. In several domains, including text-to-image production with Stable Diffusion and DALL-E·3, Q&A with ChatGPT, and music composition with MusicLM, AI-generated content (AIGC) has demonstrated incredible effectiveness [7]. However, the robustness of AIGC models is dependent on massive neural networks with billions of learnable parameters. DALL-E2 and GPT-3, for example, have 3.5 billion and 175 billion parameters, respectively [8]. With every repetition of generative inference, the resources required are substantial because of the difficulty of generating information that follows complex distributions, including images and movies. All these aspects of generation cannot be effectively represented and preserved by a single latent representation. The growing demand for automated short video generation poses major challenges in achieving high visual fidelity, temporal coherence, and semantic alignment across multiple input modalities such as text, audio, and images. Existing models often struggle to maintain consistent motion, accurate semantic representation, and realistic visual synthesis. Therefore, an advanced framework is required to effectively integrate cross-modal information and generate expressive, coherent, and visually detailed short videos. The research

aims to provide a competent cross-modal short video generation model that will provide semantic consistency, temporal coherence, and visual realism across different input modalities, including text, audio, and images.

1.1 Key contribution

- This research establishes an effective cross-modal short video generation model to ensure semantic consistency, temporal coherence, and visual realism from various input modalities.
- The MSRVTT data is gathered from Kaggle and preprocessing is done using text tokenization, MFCC-based audio feature extraction, video normalization, and gesture alignment prior to semantic decoupling extracts key visual and emotional features.
- Namib Beetle Optimized Self-Attention based Generative inputs in various Adversarial Networks (NBO-SAGAN) are employed to generate short videos and to enhance fine details and correct visual artifacts.
- Experimental evaluations show that the NBO-SAGAN approach outperforms traditional methods in FVD, KVD.

Research Questions: How effectively can NBO-SAGAN generate semantically consistent short videos from multi-modal inputs such as text, audio, and images? To what extent does the Namib Beetle Optimization improve the convergence, stability, and hyperparameter selection of SAGAN in video synthesis? How does NBO-SAGAN perform in comparison to existing short video generation methods in terms of FVD and KVD metrics? Can the semantic decoupling module enhance motion coherence, multi-object interaction, and visual realism in cross-modal video generation?

2 Related work

Research examined the classification of human actions in Red Green Blue (RGB) videos using Deep Learning (DL) [9]. Semi-supervised learning approach Vision Transformer (ViT) was applied to the HMDB51 dataset. The ViT Long Short-term Memory (LSTM) model achieved high training accuracy and low testing accuracy. Despite high training performance, the large gap in test accuracy indicated overfitting and generalization issues. The automated detection of pushing behavior in crowded event videos is a capability required by the author as noted in [10]. The hybrid framework using deep optical flow visualization, a patch-based data augmentation strategy and an efficient classifier with false reduction was proposed. The model achieved superior accuracy and outperformed the baseline Conventional Neural Network (CNN). Performance was limited by small-scale datasets and potential misclassification in complex ambiguous scenarios. Research endeavored to develop a DL model for video captioning in low-resource settings [11]. It involved video-caption data mining in the domain, reducing language complexity, and training Transformer and LSTM-based models with parameter-optimized configurations. The model effectively resolved the trade-off between accuracy and processing efficiency in devices. Device limitations and open-domain video content complexity limit real-time performance and model accuracy. Developed a semi-automated system for item recognition in surveillance movies to reduce manual labeling in the research [12]. It combined background subtraction, clustering, and an adapted YOLOv4 for unsupervised object detection and automated label creation. The framework enhanced detection performance across real-world benchmarks. There is still some reliance on the original quality of background subtraction, and it is prone to changes in scenes that are extremely dynamic or dense. Table 1 shows the literature review of existing approaches.

Table 1: A literature review of existing approaches

Reference	Aim	Method	Result	Limitations
Wu et al., [13]	Developed an efficient DL-based framework for automatic monitoring of migratory water birds.	DL framework was involved for video separation and mosaicking.	Model achieved 85.59% accuracy in processing site-wide audio data.	Model accuracy varies across different lighting conditions and habitat complexities.
Coccomini et al., [14]	Evaluated DL models for generalization in deepfake detection.	Comparative analysis of CNNs, Vision Transformers.	Vision Transformers outperformed in generalization.	Models struggled with unseen manipulation techniques, limiting real-

				world applicability.
Ul Amin et al., [15]	Developed a CNN for anomaly detection in surveillance videos.	Video shots are processed using CNN, followed by LSTM to learn spatiotemporal features.	Achieved superior performance over existing models on benchmark datasets.	Face challenges in real-time processing under poor lighting.
Dilawari et al., [16]	Explained condensing surveillance video data through automated video description.	A multitask learning framework combining CNN-based high-level feature extraction.	Model achieved METEOR scores of 33.9% (TRECViD), 34.3% (UETVS), and 31.2% (AGRIINTRUSION), outperforming existing methods.	The system still requires optimization for real-time processing in high-traffic or multi-camera environments.
Kumari & Anand [17]	Enhanced DL model for accurate sign recognition.	A CNN approach classified images through a video dataset.	Achieved average accuracy, 84.65% on the WLASL dataset demonstrating superior performance.	The model was tested only on isolated signs.
Yin et al., [18]	Introduced *LanDiff*framework combining visual fidelity.	It employs semantic compression, a language model for token generation.	Achieved a VBench score of 85.43, outperforming Hunyuan and Sora.	Requires significant computational resources and complex training.
Wang et al., [19]	Developed VideoFactory, a text-to-video generation framework.	Introduced a swapped cross-attention mechanism to enhance mutual reinforcement between blocks.	Demonstrated superior temporal consistency, and text-video alignment, outperforming existing T2V models in evaluations.	Demands high computational resources for generation and long-duration video synthesis.
Feng et al., [20]	Provided Separated Data–Semantic Coding (SDSC) to enhance semantic efficiency.	Developed a semantic encoding model to improve semantic transmission.	Experimental evaluations verified SDSC system demonstrates superior performance.	Limited for real-world deployment of cross-domain semantic generalization.
Muksimova et al., [21]	Introduced CMSTR-ODE for real-time dense video captioning.	Integrated Neural ODE-based Temporal Localization for enhanced contextual understanding.	Achieved SOTA results with best BLEU-4, and ROUGE metrics.	Enhance scalability for ultra-long videos.

Cao et al., [22]	Developed an AI-based framework for Chinese-style video-to-music generation.	Employed Latent Diffusion Model (LDM) to generate Chinese-style music visual content.	Achieved performance comparable to baseline models in audio-visual synchronization.	Requires improvement in generalization for real-time generation efficiency.
Zhang et al., [23]	Developed Fundus2Video, for generating dynamic FFA videos from static CF images.	Employed autoregressive GAN to refine lesion-focused synthesis.	Achieved FVD = 1503.21 and PSNR = 11.81, surpassing other approaches.	Limited for larger, diverse datasets and real-time deployment.
Qing et al., [24]	Suggested HiGen, model that decouples spatial and temporal factors.	Decoupled T2V generation at structure level to reduce complexity.	Achieved FID = 8.60, FVD = 406, and CLIP-SIM = 0.294, prior T2V methods.	Requires improvement in scalability for large-scale T2V generation.
Xing et al., [25]	Developed Make-Your-Video, framework guided by text and motion structure.	Utilized a Latent Diffusion Model for longer video synthesis.	Achieved FVD = 254.26 and KVD = 18.37, compared to existing methods.	Optimization needed for high-resolution video generation.

2.1 Research gap

Although advances have been made in video understanding and generation, the current approaches have problems with generalization, real-time functionality and cross-domain application. ViT-LSTM-based action recognition models [9] and frameworks of crowd behavior detection [10] have issues with overfitting and limited data. Video captioning and surveillance systems [11-12,16,21] are very accurate but are constrained by device features, scale, and dynamism. Text-to-video and cross-modal generation models, such as LanDiff [18], VideoFactory [19], HiGen [24], Make-Your-Video [25], Chinese-style music generation [22], and Fundus2Video [23], exhibit good domain-specific performance but require large amounts of computation and cannot produce a variety of content or long-length content. Semantic communication frameworks [20] are more efficient yet are difficult to implement in the real world. These gaps show a necessity of solid, effective, and flexible structures that could be able to manage a variety of video domains without compromising the quality of outputs displayed in real-time.

This research overcomes these limitations by proposing an NBO-SAGAN method for robust, efficient, and scalable video understanding and generation framework that ensures real-time processing, cross-domain applicability, and high-quality output across diverse video domains.

3 Methodology

A powerful cross-modal short video model learns to produce short videos based on the different input modalities including text, audio, and images. It starts with gathering the MSR-VTT dataset in Kaggle and performs a series of preprocessing steps, such as tokenization, MFCC feature extraction, data normalization and gesture alignment, and semantic decoupling is used to isolate and recognize major features to further analyze them. The NBO-SAGAN model produces a cross-modal video work process, depicted in Figure 1.

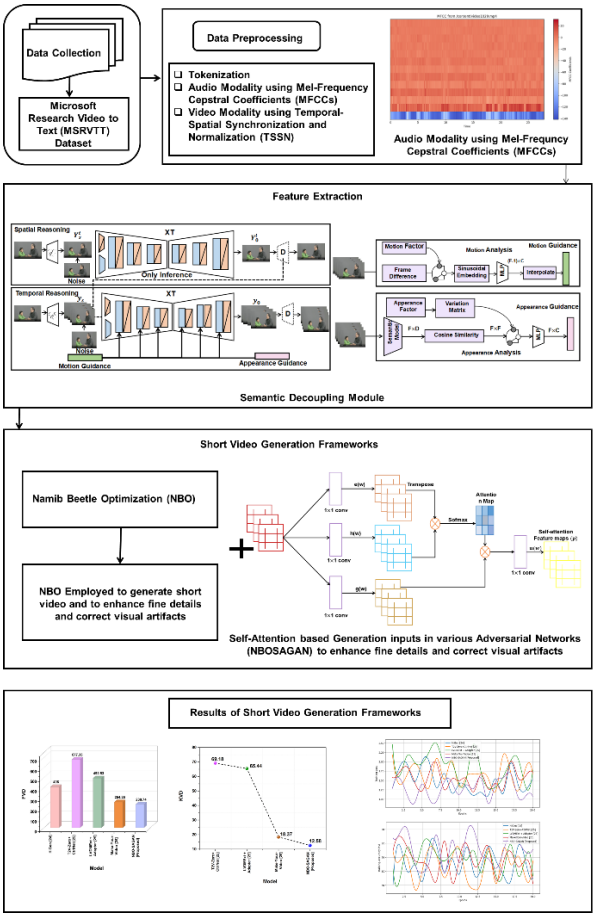

Figure 1: Overview of methodology workflow

3.1 Dataset

The Microsoft Research Video to Text (MSRVTT) dataset is gathered from the open source Kaggle, it has 10,000 video clips from 20 categories, each of which has 20 English lines. The total number of unique words in all the captions is around 29,000. The dataset used for this research split into 80% for training and 20% for testing to ensure balanced model evaluation. Kaggle data sample images are shown in Figure 2.

Source:

https://www.kaggle.com/datasets/vishnutheepb/msrvtt

Figure 2: Sample images from Kaggle data

3.2 Data preprocessing

Data preprocessing refers to the series of steps taken to clean, structure, and transform raw data into a usable format for model training and analysis. Data undergoes extensive preprocessing, tokenizing text, and transforming audio into features using MFCCs, normalizing video clips, and ensuring alignment with speech-driven gestures.

➢ **Tokenization**

Tokenization is conducted through a transformer-compatible tokenizer CLIP to maintain semantic consistency from text descriptions of corresponding video clips. This guarantees the preservation of contextual relationships crucial for proper alignment with video content. Tokenization of the dataset involves splitting case titles and text into individual words or phrases for analysis and processing. Tokenization is also the process of replacing sensitive data with individual identification symbols while retaining critical information and ensuring video quality. In extended tokenization, isolated tokens are grouped to produce higher-level tokens, and separated strings are divided into basic processing units, as illustrated in Figure 3. It illustrates text-to-video alignment, showing how tokenized words like “sir”, “student”, and “talking” are used by the model to identify actors, actions and contextual scene information.

Text caption

A sir is talking to a student.

Tokenization

The sentence is divided into meaningful units:

[CLS], *"a"*, *"sir"*, *"is"*, *"talking"*, *"to"*, *"a"*, *"student"*

Alignment with Video: The video shows this exact scene.The model uses the tokens to match:Who is in the scene (sir, student). What action is happening (talking)The context (likely a classroom).

Figure 3: Sample images from Kaggle data are tokenized processing

➢ **Mel-frequency cepstral coefficients (MFCCs) for audio modality**

MFCCs are employed to capture speech characteristics relevant to gesture. This process eliminates the unwanted background noise and conserves only the useful spectral information related to speech patterns. It involves several preprocessing and transformation steps. First, a pre-emphasis filter is utilized to boost the higher frequencies and balance the speech spectrum, as in Equation (1),

$$y[n] = x[n] - ax[n-1] \quad (1)$$

Where $x[n]$ is the input audio signal at the current sample index n, $y[n]$ is the pre-emphasized output, and a (typically between 0.95-0.98) is the pre-emphasis coefficient. Next, the signal is divided into short overlapping frames, and each frame gets multiplied using

a hamming window $w[n]$ to minimize spectral leakage, as described in Equation (2):

$$w[n] = 0.54 - 046\cos\left(\frac{2\pi n}{N-1}\right) \quad (2)$$

Where N represents window length n is the sample index within each frame. The windowed frames are then converted into frequency domain using the Discrete Fourier Transform (DFT) to gain spectral information. The Frequency-domain values are filtered using a series of bandpass filters, with central frequencies spaced uniformly on the Mel scale as represented in the Equation (3),

$$N(e) = 1127\, In\left(1 + \frac{e}{700}\right) \quad (3)$$

Here, $N(e)$ is the frequency converted to Mel scale and e is the input frequency. The energy in each of Mel-filtered band is computed via summing the squared amplitudes of frequency bins within that band. Finally, MFCC coefficients are gained by applying a (DCT) to the log energies of the Mel filter banks by Equation (4),

$$NEDD_{j,i} = \frac{1}{S}\sum_{s=1}^{S} \log[NE(s)] \cos \quad \left[\frac{2\pi}{S}\left(s + \frac{1}{2}\right) i\right] \quad (4)$$

Where $NEDD_{j,i}$ is the jth MFCC coefficient, S is the total number of filters, s represents index of the filter $(1 < s < S)$, $NE(s)$ is the energy of the signal and i is the index. Figure 4 shows the MFCC representation of audio modality from Video1029.mp4.

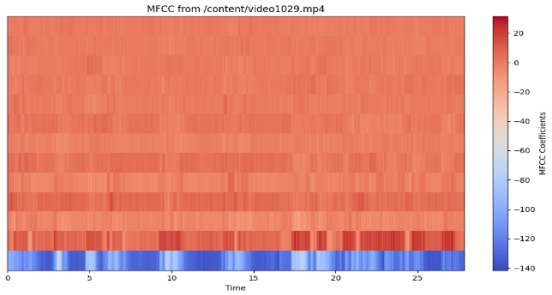

Figure 4: MFCC representation of audio modality from Video1029.mp4

➢ **Video modality using temporal-spatial synchronization and normalization (TSSN)**

TSSN normalized to ensure consistent resolution, frame rate, and format; gestures aligned with speech timing. It normalizes video frame rate, resolution, and format for consistent samples to provide for correct multimodal learning. Through the resolution of inconsistencies in recording environments, TSSN allows for strong feature extraction so that temporal and spatial aspects of gestures and speech are represented as comparably as possible for analysis.

3.3 Feature extraction using semantic decoupling module

SDM is utilized in this research to disentangle high-level semantic features from the input data. The module separates appearance, motion, emotional tone and background context into an independent representation, by facilitating better control and alignment during video synthesis. It comprises of three submodules:

- Spatial Reasoning Block: The static spatial information such as object boundaries, appearance and background context gets extracted using convolutional and attention layers.
- Temporal Reasoning Block: Captures motion continuity and frame-to-frame transitions through a recurrent temporal encoder.
- Motion Guidance Unit: Combines visual motion elements with semantic information from text or audio (like rhythm or emotion) to enable cross-modal consistency.

The disentangled representations from SDM are added to the latent space and it guides the NBO-SAGAN generator to yield coherent, temporally aligned and semantically consistent video frames. This modular design enables accurate control over individual visual and semantic components, by improving realism and fidelity in generated videos. Figure 5 shows the Visualized representation of semantic decoupling module.

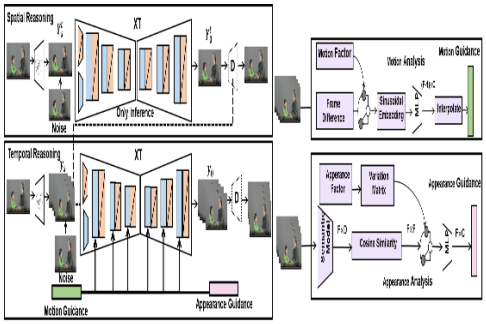

Figure 5: Visual representation of semantic decoupling module

3.4 Namib beetle optimized self-attention based generative inputs in various adversarial networks (NBO-SAGAN)

The NBO-SAGAN hybrid model integrates the Namib Beetle Optimization (NBO) algorithm with the Self-Attention Generative Adversarial Network (SAGAN) to enhance cross-modal short video generation. NBO acts as a meta-heuristic optimizer, fine-tuning critical hyperparameters such as learning rate, regularization weight, and attention unit count, inspired by the beetle's wetness-harvesting strategy for adaptive search and convergence. Concurrently, SAGAN captures long-range spatial-temporal dependencies through self-attention, ensuring semantic coherence, motion stability, and fine-

grained visual detail across modalities like text, audio, and images. This hybrid framework enables efficient training, improved feature selection, and robust semantic alignment, resulting in contextually accurate, visually realistic, and temporally coherent short video synthesis with optimized cross-modal fusion and expressive content generation.

3.4.1 Self-attention based generative inputs in various adversarial networks (SAGAN)

SAGAN extend traditional GAN architectures by integrating a self-attention mechanism that captures long-range dependencies across spatial and temporal regions. Unlike standard convolutional layers, which process only local neighborhoods, SAGAN allows both the generator and discriminator to consider relationships between widely separated regions in the video, enhancing temporal coherence, semantic alignment. SAGAN in this research plays a role in improving cross-modal short video generation. It enables the model to synthesize video frames effectively that aligns with various input modalities like audio, text and images by modelling global dependencies and long-range correlations across frames. This is crucial for maintaining movement continuity, conserving object consistency and aligning semantic cues from different input sources.

Let the feature map from a previous hidden layer be $W = \{w_1, w_2, \ldots w_M\}$, where M is the number of feature locations. The features are first projected into separate embedding spaces as in Equation (5),

$$e(w) = X_e{}^w, h(w) = X_h w \tag{5}$$

Where $e(w)$ is the embedding of feature w in the query space, $h(w)$ represents embedding of feature $w$in the key space, $X_e{}^w, X_h w$ are the learnable projection matrices, The attention weights are calculated as in the Equation (6),

$$\beta_{i,j} = \frac{\exp(t_{j,i})}{\sum_{j=1}^{M} \exp(t_{ji})}, where t_{j,i} = e(w_j)^S h(w_i) \tag{6}$$

Here, $\beta_{i,j}$ is the attention weight representing the influence of feature $w_j$on w_i, $\exp(t_{ji})$ is the exponential function for similarity score between feature embeddings, $e(w_j)^S$,$h(w_i)$ represents Query embedding and Key embedding, attention layer output is obtained via weighted aggregation is given in the Equation (7),

$$p_i = u\left(\sum_{j=1}^{M} \beta_{i,j} g(w_j)\right), g(w_j) = X_g w_j, u(w_j) = X_u w_j \tag{7}$$

Where p_i is the aggregated feature for location $i$after attention, u is the transformation function, $g(w_j)$ is the value embedding of feature w_j, X_g, X_u are the learnable weight matrices for value projection and post-processing. A learnable scale parameter γmodulates the attention output, and the original feature is added back as in the Equation (8),

$$z_j = \gamma p_j + w_j \tag{8}$$

Where $z_j$represents final output feature, γ is the learnable scalar initialized to 0 to allow gradual learning of non-local dependencies, p_j is attention-aggregated feature at location j. The generator and discriminator are trained using a hinge-style adversarial loss adapted for video synthesis is given as in Equation (9),

$$K_C = -F_{(w,z)\sim Odata}[\min(0, -1 + C(w,z))] - F_{y\sim o_y}, z\sim o_{data}[\min(0, -1 - C(H(y),z))]$$
$$K_H = -F_{y\sim o_y}, z\sim o_{data} C(H(y),z) \tag{9}$$

Where K_C and K_H are hinge loss for discriminator and generator, $F_{(w,z)\sim Odata}$ and $F_{y\sim o_y}$ are expectation operator over the sampled distribution, $\min(0, -1 + C(w,z))$ is Hinge function $C(H(y),z)$ is the discriminator output, $H(y)$ is the generated video features from input y, O_{data} is the real data distribution, o_y is the distribution of corresponding inputs. This self-attention mechanism enables the network to capture long-range spatial-temporal dependencies while preserving local features, improving semantic coherence, motion consistency, and fine details in cross-modal short video generation. By combining SAGAN into the framework, the model achieves robust cross-modal short video generation, effectively integrating semantic understanding, temporal alignment, and realistic visual synthesis across multiple input modalities. The attention weights as a function of query, key, and value transformations that the model selectively attends to globally important areas. By integrating self-attention layers into both the generator and discriminator, SAGAN greatly enhances structural coherence, texture content, and semantic correspondence in hard generative tasks. Figure 6 depicts the structure of SAGAN.

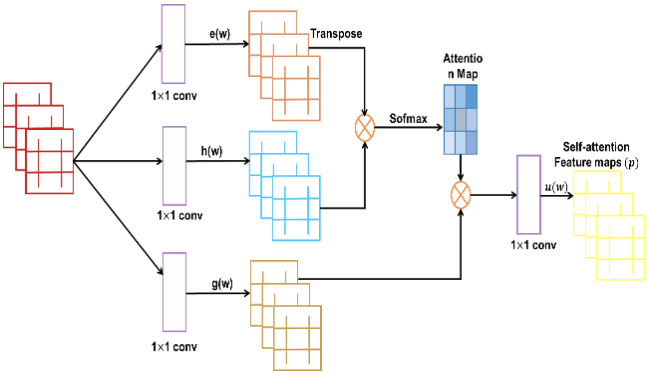

Figure 6: Structure of SAGAN

3.4.2 Namib beetle optimization (NBO)

NBO functions as a meta-heuristic hyperparameter tuning mechanism rather than a structural component of generative network in this research. Particularly, it is applied to optimize the key training parameters such as

learning rate, regularization weight, and the number of attention units in both the generator and discriminator. During training, NBO evaluates the various parameter configurations and selects the one which produces lowest total loss, by increasing the convergence stability and performance. While NBO supports efficient training, it does not directly involve in feature generation or adversarial learning; instead, it enhances the training optimization phase of the SAGAN model. NBO in a cross-modal short video generation scheme using diffusion model and semantic decoupling to improve quality, coherence, and efficiency of video generation by directing the optimization process using nature-inspired decision-making. NBO emulates the humidity-gathering activity of the beetle to obtain optimal feature choice and parameter adjustment between modalities. Within this system, NBO guarantees that the semantic content of text inputs is properly aligned and maintained throughout the diffusion-based generation, which results in more contextually correct and visually appealing short videos. Every issue solution is seen as a beetle and represented in the Matrix aggregating all feature vectors across D dimensions (MA) using Equation (10) with individual feature vectors $[W_1, W_2, W_3, \ldots\ldots, W_D]$.

$$MA = [W_1, W_2, W_3, \ldots\ldots, W_D] \tag{10}$$

Every beetle has a D decision variable, then randomly generated $MA_{1,1}..MA_{M,C}$ as the beginning population using Equation (11)

$$\text{POP} = \begin{bmatrix} MA_{1,1}MA_{1,2} \ldots MA_{1,C} \\ MA_{2,1}\, MA_{2,2} \ldots\ldots MA_{2,C} \\ \ldots \\ \ldots \\ \ldots \\ MA_{M,1}\, MA_{M,2} \ldots\ldots MA_{M,C} \end{bmatrix} \tag{11}$$

Every candidate solution is processed as a "beetle" on a search landscape. These beetles emulate random search for optimal parameters according to environmental signals (fitness landscape) using the Equation (12).

$$MA_{j,i} = K + (V - K).qand(0,1) \tag{12}$$

Here $MA_{j,i}$ is Feature value at spatial location $j$and channel i, K , V are minimum and maximum value for feature scaling, $qand(0,1)$ is the scaling factor.

Humidity Sensing: The beetle detects humidity from the air, and every solution is tested on its fitness. However, $(MA_{1,1}MA_{1,2} \ldots MA_{1,C})$also displays the $e(MA_{M,1}MA_{M,2}MA_{M,C})$component linked to the insect Equation (13) which implies Aggregated fitness representation (Fitness) to generate any solution in the problem space at random. It is used to assess each solving solution or beetle after it is set into a random problem space. Every beetle with a higher function computation value is better able to hold on water.

$$\text{Fitness} = \begin{bmatrix} e(MA_{1,1}MA_{1,2} \ldots MA_{1,C}) \\ e(MA_{2,1}MA_{2,2} \ldots MA_{1,C}) \\ \ldots \\ \ldots \\ \ldots \\ e(MA_{M,1}MA_{M,2}MA_{M,C}) \end{bmatrix} = \begin{bmatrix} e(MA_1) \\ e(MA_2) \\ \ldots \\ \ldots \\ \ldots \\ e(MA_M) \end{bmatrix} \tag{13}$$

Where $MA_{M,C}$ signifies feature value at spatial location $m$and channel $c$from the matrix MA, MA_M represents entire feature vector at spatial location m, e is encoding function with Total number of spatial locations (M) and Total number of channels (C).

Condensation Strategy of Exploitation and Exploration: Beetles move towards regions of high solutions, sharpening their positions of$\|MA_j - MA_i\|$. Search variability is sustained to avoid early convergenced_{ji}, similar to the way the beetle finds ways to adjust under different fog conditions $\sum_{l=1}^{C}(MA_{j,l} - MA_{i,l}$represented in Equation (14). the beetles attempt to return to the nest by absorbing from the surrounding air.

$$d_{ji} = \|MA_j - MA_i\| = \sqrt{\sum_{l=1}^{C}(MA_{j,l} - MA_{i,l})^2} \tag{14}$$

Here, d_{ji} provides the distance or dissimilarity between feature vectors at locations j and i , $\|MA_j - MA_i\|$ is the norm operator for the Feature vectors at spatial locations j and i.

The NBO algorithm efficiently promotes cross-modal short video generation by facilitating optimal feature choice and parameter adjustment. NBO guarantees accurate semantic alignment and enhanced visual coherence within diffusion-based models to support better contextually richer and more accurate video synthesis in modalities. Pseudocode 1 shows the NBO-SAGAN.

Pseudocode 1: Namib Beetle Optimized Self-Attention based Generative Inputs in various Adversarial Networks (NBO-SAGAN)

Step 1: Initialize SAGAN
Initialize generator G and discriminator D with random weights.
Set learning rate (LR), regularization weight (λ), attention units (AU).
Load multi-modal inputs: text (T), audio (A), and image (I).
Extract feature embeddings:
F_text = Encoder_Text(T)
F_audio = MFCC_Extractor(A)
F_image = CNN_Encoder(I)
Combine feature representations → F_combined = [F_text, F_audio, F_image].

Initialize self-attention layers in both G and D.
Define adversarial loss (hinge-based) for video synthesis.
Step 2: Initialize NBO (Namib Beetle Optimization)
Set population size M, number of decision variables D, and iterations N.
Randomly generate beetle population:
For i = 1 to M:
Beetle[i] = random(D)
Evaluate fitness of each beetle using initial SAGAN loss:
Fitness[i] = Evaluate(G, D, Beetle[i])
Store best beetle (Best_B) = min(Fitness)
Step 3: Process of NBO-SAGAN (Optimization + Training)
For iteration = 1 to N do:
For each beetle i in population:
Randomly generate search direction and step size.
Update beetle position:
*Beetle[i] = Beetle[i] + Step * rand(−1, 1)*
Recalculate fitness:
Fitness[i] = Evaluate(G, D, Beetle[i])
If Fitness[i] < Fitness[Best_B] then
Best_B = Beetle[i]
Else
Keep previous Best_B
End For
Update SAGAN parameters:
LR = Best_B.LR
λ = Best_B.Reg
AU = Best_B.AttnUnits
Step 4: Output Results
Output optimized short video clips V_gen.
Display evaluation metrics: FVD, KVD.
Return trained model (G, D*) and best NBO parameters (Best_B).*
End Algorithm

The NBO-SAGAN hybrid combines Namib Beetle Optimization (NBO) with Self-Attention GAN (SAGAN), leveraging NBO for adaptive hyperparameter tuning and SAGAN for capturing long-range spatial-temporal dependencies. This synergy enhances training stability, semantic alignment, and motion coherence, producing visually realistic, contextually accurate short videos. The hybrid framework ensures efficient convergence, optimized feature selection, and improved cross-modal integration across text, audio, and image inputs. Table 2 provides the Hyperparameter settings for the research.

Table 2: Hyperparameter settings

Hyperparameter	Typical Values
Hidden units per dense layer	128, 256, 512
Epochs	50, 100
Dropout rate	0.3, 0.5, 0.6
Optimizer	NBO, SAGAN
Batch size	32, 64, 128
Learning rate	0.001, 0.0001
Number of filters	32, 64
Activation function	ReLU, Leaky ReLU
Number of convolutional layers	2, 3
Pooling size	2×2, 3×3

4 Result

This research established an effective cross-modal short video generation model to ensure semantic consistency, temporal coherence from various input modalities. Table 3 shows the hardware and software configuration.

Table 3: Experimental setup

Component	Details
CPU	Intel Core i7-12700K
RAM	32 GB DDR4
GPU	NVIDIA RTX 3090 (24 GB)
OS	Ubuntu 22.04 LTS
Framework	PyTorch 2.1.0
CUDA	Version 11.8
Language	Python 3.10

4.1 Accuracy and loss

Accuracy is the number of correct predictions made by a classical model to the total number of predictions, whereas loss is the difference between expected and actual values, which measures how well the model performs throughout training. The loss curve shows how the model converged during training, with lower values representing better performance, while the accuracy curve shows how well the short video model is throughout subsequent epochs. The accuracy and loss characteristics of the training for the NBO-SAGAN technique, as shown in Figure 7.

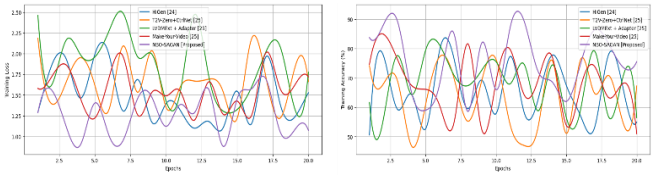

Figure 7: Graphical representation of (a) loss (b) accuracy

4.2 Comparative analysis

The effectiveness of the NBO-SAGAN method and existing methods is evaluated using FID, FVD, KVD, and CLIPSM. The NBO-SAGAN method is compared with existing approaches HiGen [24], T2V-Zero+CtrlNet [25], $LVDM_{Ext}$ +Adapter [25], Make-Your-Video [25]. Table 4 provides the outcome of the metrics.

Table 4: Outcome of numerical values of metric

Methods	FID	FVD	CLIPSM
HiGen[24]	8.60	406	0.2947
NBO-SAGAN [Proposed]	7.95	385	0.391

4.2.1 Frechet video distance (FVD)

It measures the distance between distributions of real and generated video clips in a feature space extracted using a pre-trained video recognition mode. Lower value indicates more realistic and temporally consistent videos. Table 5 provides the outcomes of FVD.

Table 5: Outcome of numerical values of FVD

Model	FVD
HiGen [24]	406
T2V-Zero+CtrlNet [25]	677.35
$LVDM_{Ext}$+Adapter [25]	492.53
Make-Your-Video [25]	254.26
NBO-SAGAN [Proposed]	230.74

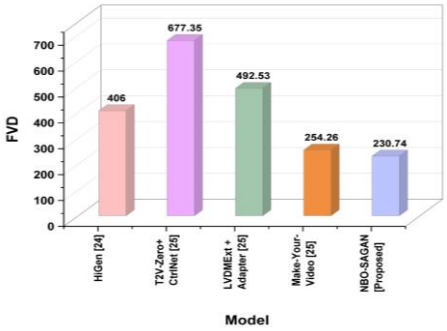

Figure 8: Graphical illustration of FVD performance

Figure 8 presents the graphical representation of text to video generation models based on FVD. The proposed model achieved the lowest FVD of 230.74, indicating superior performance over realistic and temporally consistent videos compared with other existing models.

4.2.2 Kernel video distance (KVD)

KVD evaluates the semantic alignment between input modalities (text, image and audio) and generated short videos using a shared embedding space. In this research, KVD evaluates how well NBO-SAGAN preserves meaning and context across modalities. A lower KVD shows stronger cross-modal coherence, reflecting model's capacity to generate semantically consistent and visually realistic short videos. Table 6 represents the comparative analysis of KVD.

Table 6: Outcome of numerical values of KVD metric

Model	KVD
T2V-Zero+CtrlNet [25]	69.18
$LVDM_{Ext}$+Adapter [25]	65.44
Make-Your-Video [25]	18.37
NBO-SAGAN [Proposed]	12.58

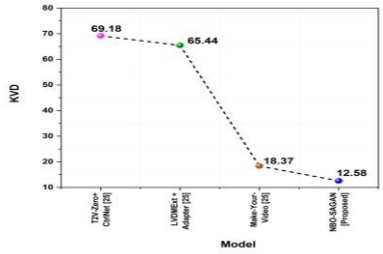

Figure 9: Graphical view of PVD performance

Figure 9 provides the graphical illustration of KVD scores of recent short video generation models. The proposed model achieved the lowest KVD of 12.58, providing superior video quality.

4.2.3 CLIP similarity (CLIPSIM)

It measures the semantic similarity between text and generated visual content using the CLIP model, which encodes both text and images into a shared embedding space. A higher CLIPSIM score indicates better alignment between the textual input and visual output. NBO-SAGAN scored 0.391, and its videos are more semantically aligned with the input text or image. HiGen [24] had a lower CLIPSM score of 0.2947, reflecting weaker similarity. These findings show that NBO-SAGAN is more effective and efficient compared to current approaches such as HiGen [24], generating higher quality, more coherent, and semantically closer videos.

4.3 Visualization of the effects

Visualization of the effect of different cross-modal inputs (text, audio, image) on short video generation using the proposed NBO-SAGAN framework. The generated frames exhibit semantic consistency and motion coherence aligned with the input modality. Figure 10 shows the visualization of generated video frames.

Figure 10: Visualization of generated video frames for the input *of a golfer playing hockey and football players in action*

4.4 Statistical test

Table 7: Paired T-Test results

Metric	Baseline Model	Baseline Mean ± SD	NBO-SAGAN (Proposed) Mean ± SD	t-value	p-value	Significance (p < 0.05)
FVD	HiGen	406 ± 8	230.74 ± 6	12.15	<0.001	Yes
	T2V-Zero+CtrlNet	677.35 ± 10	230.74 ± 6	24.67	<0.001	Yes
	LVDMExt + Adapter	492.53 ± 9	230.74 ± 6	15.23	<0.001	Yes
	Make-Your-Video	254.26 ± 7	230.74 ± 6	5.12	0.002	Yes
KVD	T2V-Zero+CtrlNet	69.18 ± 1.2	12.58 ± 0.5	32.45	<0.001	Yes
	LVDMExt + Adapter	65.44 ± 1.0	12.58 ± 0.5	29.67	<0.001	Yes
	Make-Your-Video	18.37 ± 0.8	12.58 ± 0.5	10.34	0.001	Yes

Table 7 compares FVD and KVD metrics across baseline models. The proposed NBO-SAGAN achieves the best performance with an FVD of 230.74 and a KVD of 12.58, significantly outperforming all baseline models ($p < 0.05$).

4.5 Ablation study

Table 8 provides the ablation study signifying how preprocessing, feature extraction, and the proposed model effectively reduce FVD and KVD, demonstrating each component's contribution for enhanced video generation quality.

Table 8: Ablation study for evaluating contributions of each component on performance

Description	FVD	KVD
Dataset	450.00	28.00
Dataset + Preprocessing	400.50	22.50
Dataset + Preprocessing + Feature Extraction	350.20	18.30
Dataset + Preprocessing + Feature Extraction + NBO-SAGAN (Proposed)	230.74	12.58

An effective cross-modal short video generation framework was established in this research that confirms semantic consistency, temporal coherence, and visual realism from various modalities such as text, audio, and images. Current video generation and cross-modal frameworks are subject to a number of challenges. Most of them are based on computationally intensive architectures, which restrict definition of real-time and high-resolution generation [18,19,24,25]. The models tend to have problems with temporal consistency, long-term synthesis, and generalization of unknown or complicated audio-visual data [14-16,20-23]. Reliance on controlled datasets, pre-trained encoders, or isolated segments further limits scalability and real-world implementation [17,21]. HiGen [24] limited scalability for large-scale text-to-video generation, T2V-Zero+CtrlNet [25] struggles with temporal coherence and long duration video synthesis, LVDMExt +Adapter [25] requires high computational resources for complex video generation and Make-Your-Video [25] needed optimization for high-resolution and long video outputs. The proposed NBO-SAGAN overcomes these limitations by generating semantically aligned, temporally coherent and high-fidelity cross-modal short videos effectively. This framework is suitable for robotics, surveillance, and multimedia applications, allowing realistic and temporally consistent video generation across diverse domains. Results demonstrated that proposed model resulted with FVD of 230.74 and KVD of 12.58, outperforming the baseline models.

5 Conclusion

The research established an effective cross-modal short video generation model that ensures semantic consistency, temporal coherence, and visual realism from various input modalities such as text, audio, and images. The MSRVIT dataset is gathered from the Kaggle website, with extensive preprocessing using tokenizing text, transforming audio into features using MFCCs, normalizing video clips, and ensuring alignment with speech-driven gestures. The framework begins with a semantic decoupling module that utilizes encoders to extract The suggested model of NBO-SAGAN attained high performance with an FVD of 230.74 and KVD of 12.58, which revealed high visual realism and semantic consistency. The limitations of the framework are

dependence on high-quality multimodal data and high computational expense which includes high GPU memory usability and longer inference times, which may restrict its applicability on standard hardware and real world scenarios. Future research will continue to push generalization through few-shot learning, simplify complexity through lightweight models, and improve fusion accuracy while enabling real-time, longer video generation.

DECLARATION

Ethics approval and consent to participate: I confirm that all the research meets ethical guidelines and adheres to the legal requirements of the study country.

Consent for publication: I confirm that any participants (or their guardians if unable to give informed consent, or next of kin, if deceased) who may be identifiable through the manuscript (such as a case report), have been given an opportunity to review the final manuscript and have provided written consent to publish.

Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon request.

Competing interests: here are no have no conflicts of interest to declare.

Authors' contributions (Individual contribution): All authors contributed to the study conception and design. All authors read and approved the final manuscript

References

[1] Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., & Liu, Z. (2025). Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, *133*(5), 3059–3078. https://doi.org/10.1007/s11263-024-02295-1

[2] Weng, Z., Yang, X., Xing, Z., Wu, Z., & Jiang, Y. G. (2024). GenRec: Unifying video generation and recognition with diffusion models. *arXiv preprint*, arXiv:2408.15241. https://doi.org/10.48550/arXiv.2408.15241

[3] Chen, L., Deng, Z., Liu, L., & Yin, S. (2024). Multilevel semantic interaction alignment for video–text cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(7), 6559–6575. https://doi.org/10.1109/TCSVT.2024.3360530

[4] Umale-Nagmote, A., Goel, C., & Lal, N. (2025). Enhanced intelligent video monitoring using hybrid integration of spatiotemporal autoencoders and convolutional LSTMs. *Informatica*, *49*(18), 51–68. https://doi.org/10.31449/inf.v49i18.7502

[5] Chen, D., & Zhang, S. (2025). Deep learning-based involution feature extraction for human posture recognition in martial arts. *Informatica*, *49*(12), 77–90. https://doi.org/10.31449/inf.v49i12.7041

[6] Wang, X., Li, Y., Wang, H., Huang, L., & Ding, S. (2022). A video summarization model based on deep reinforcement learning with long-term dependency. *Sensors*, *22*(19), 7689. https://doi.org/10.3390/s22197689

[7] Fuentes, A., Han, S., Nasir, M. F., Park, J., Yoon, S., & Park, D. S. (2023). Multiview monitoring of individual cattle behavior based on action recognition in closed barns using deep learning. *Animals*, *13*(12), 2020. https://doi.org/10.3390/ani13122020

[8] Guo, J., Wang, M., Yin, H., Song, B., Chi, Y., Yu, F. R., & Yuen, C. (2025). Large language models and artificial intelligence generated content technologies meet communication networks. *IEEE Internet of Things Journal*, *12*(2), 1529–1553.

[9] Surek, G. A. S., Seman, L. O., Stefenon, S. F., Mariani, V. C., & Coelho, L. D. S. (2023). Video-based human activity recognition using deep learning approaches. *Sensors*, *23*(14), 6384. https://doi.org/10.3390/s23146384

[10] Alia, A., Maree, M., & Chraibi, M. (2022). A hybrid deep learning and visualization framework for pushing behavior detection in pedestrian dynamics. *Sensors*, *22*(11), 4040. https://doi.org/10.3390/s22114040

[11] Gad, G., Gad, E., Cengiz, K., Fadlullah, Z., & Mokhtar, B. (2022). Deep learning-based context-aware video content analysis on IoT devices. *Electronics*, *11*(11), 1785. https://doi.org/10.3390/electronics11111785

[12] Gomaa, A., & Abdalrazik, A. (2024). Novel deep learning domain adaptation approach for object detection using semi-self-building dataset and modified YOLOv4. *World Electric Vehicle Journal*, *15*(6), 255.

[13] Wu, E., Wang, H., Lu, H., Zhu, W., Jia, Y., Wen, L., & Jian, H. (2022). Unlocking the potential of deep learning for migratory waterbirds monitoring using surveillance video. *Remote Sensing*, *14*(3), 514. https://doi.org/10.3390/rs14030514

[14] Coccomini, D. A., Caldelli, R., Falchi, F., & Gennaro, C. (2023). On the generalization of deep learning models in video deepfake detection. *Journal of Imaging*, *9*(5), 89.

[15] Ul Amin, S., Ullah, M., Sajjad, M., Cheikh, F. A., Hijji, M., Hijji, A., & Muhammad, K. (2022). EADN: An efficient deep learning model for anomaly detection in videos. *Mathematics*, *10*(9), 1555.

[16] Dilawari, A., Khan, M. U. G., Al-Otaibi, Y. D., Rehman, Z. U., Rahman, A. U., & Nam, Y. (2021). Natural language description of videos for smart surveillance. *Applied Sciences*, *11*(9), 3730. https://doi.org/10.3390/app11093730

[17] Kumari, D., & Anand, R. S. (2024). Isolated video-based sign language recognition using a hybrid CNN-LSTM framework based on attention mechanism.

Electronics, *13*(7), 1229. https://doi.org/10.3390/electronics13071229

[18] Yin, A., Shen, K., Leng, Y., Tan, X., Zhou, X., Li, J., & Tang, S. (2025). The best of both worlds: Integrating language models and diffusion models for video generation. *arXiv preprint*, arXiv:2503.04606. https://doi.org/10.48550/arXiv.2503.04606

[19] Wang, W., Yang, H., Tuo, Z., He, H., Zhu, J., Fu, J., & Liu, J. (2025). Swap attention in spatiotemporal diffusions for text-to-video generation. *International Journal of Computer Vision*, 1–19. https://doi.org/10.1007/s11263-025-02349-y

[20] Feng, Y., Xu, J., Liang, C., Yu, G., Hu, L., & Yuan, T. (2023). Decoupling source and semantic encoding: an implementation study. *Electronics*, *12*(13), 2755. https://doi.org/10.3390/electronics12132755

[21] Muksimova, S., Umirzakova, S., Sultanov, M., & Cho, Y. I. (2025). Cross-modal transformer-based streaming dense video captioning with neural ODE temporal localization. *Sensors*, *25*(3), 707. https://doi.org/10.3390/s25030707

[22] Cao, M., Zheng, J., & Zhang, C. (2025). AI-based Chinese-style music generation from video content: a study on cross-modal analysis and generation methods. *EURASIP Journal on Audio, Speech, and Music Processing*, *2025*(1), 8. https://doi.org/10.1186/s13636-025-00397-3

[23] Zhang, W., Huang, S., Yang, J., Chen, R., Ge, Z., Zheng, Y., Shi, D., & He, M. (2024). Fundus2video: Cross-modal angiography video generation from static fundus photography with clinical knowledge guidance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 689–699). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72378-0_64

[24] Qing, Z., Zhang, S., Wang, J., Wang, X., Wei, Y., Zhang, Y., Gao, C., & Sang, N. (2024). Hierarchical spatio-temporal decoupling for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6635–6645).

[25] Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., & Shan, Y. (2024). Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, *31*(2), 1526–1541. https://doi.org/10.1109/TVCG.2024.3365804