VideoStylist: Text-to-Consistent Video Stylization with Temporal Anchor Tokens

Hunter Shaw and Mark Harris *

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1287.v1

Keywords: text-guided video stylization; diffusion model; temporal consistency; style fidelity; temporal anchor tokens

 doi:10.20944/preprints202603.1287.v1

Article

VideoStylist: Text-to-Consistent Video Stylization with Temporal Anchor Tokens

Hunter Shaw and Mark Harris *

California State University Sacramento
* Correspondence: bnl056675@cncivirtual.mx

Abstract

Extending Text-to-Image (T2I) generation to Text-Guided Video Stylization (T2GVS) presents significant challenges in temporal consistency, style fidelity, and fine-grained control. Naive frame-by-frame T2I application results in severe flickering. We propose VideoStylist, a novel diffusion model extending a pre-trained T2I U-Net to a four-dimensional architecture for high-quality stylized video generation. Key innovations are Temporal Anchor Tokens (TATs) globally anchoring style semantics across frames, mitigating flickering, and an Adaptive Spatio-Temporal Consistency Module (ASTCM) to enhance local coherence and smooth transitions via dynamic spatio-temporal attention. A diverse video-text dataset was constructed using a dual strategy, combining LLM-generated descriptions and extending T2I datasets with weak labels. Extensive experiments show VideoStylist significantly outperforms state-of-the-art baselines across Style Fidelity, Temporal Consistency, and Perceptual Quality, achieving superior performance and strong user preference. Ablation studies confirm the critical contributions of TATs and ASTCM. VideoStylist advances T2GVS, delivering stable, high-fidelity, and visually appealing stylized video content.

Keywords: text-guided video stylization; diffusion model; temporal consistency; style fidelity; temporal anchor tokens

1. Introduction

The remarkable progress in Text-to-Image (T2I) generation has enabled the creation of high-quality, diverse images from simple text prompts [1]. This breakthrough has revolutionized various creative fields and laid the groundwork for advanced visual content generation. Beyond creative domains, the pervasive influence of AI is also evident in diverse applications such as online parameter identification in industrial systems [2–4], intelligent resource management and sustainability efforts [5,6], and advanced decision-making systems across various sectors including fraud detection [7], human resource management [8], and power grid operations [9]. Furthermore, AI advancements are continually enhancing specialized computer vision tasks like object detection [10] and facial expression analysis [11]. Extending the powerful T2I capability to the video domain, specifically for *Text-Guided Video Stylization (T2GVS)*, holds immense potential for applications ranging from artistic content creation and film post-production to personalized video editing. Imagine transforming any video into a "Van Gogh painting" or a "cyberpunk cityscape" simply by providing a text description.

Despite the advancements in T2I, migrating these techniques directly to video presents formidable challenges. The core difficulties in achieving effective T2GVS can be categorized into three main areas:

- **Temporal Consistency:** A naive frame-by-frame application of T2I models inevitably leads to severe flickering and content discontinuity across frames [12], severely degrading the visual quality and coherence of the stylized video. Maintaining smooth transitions and stable object appearances is paramount.
- **Style Fidelity:** Accurately translating the complex artistic style described in a text prompt, such as "watercolor painting" or "futuristic chrome," onto every frame of a video while preserving the

 doi:10.20944/preprints202603.1287.v1

original video's underlying content structure remains a significant hurdle. The stylized output must faithfully adhere to the textual style instruction.

- **Fine-grained Control:** Current T2GVS methods often lack the flexibility for users to apply different styles to specific regions of a video, particular elements, or varying time segments. Achieving such precise control over the stylization process is crucial for professional and artistic applications.

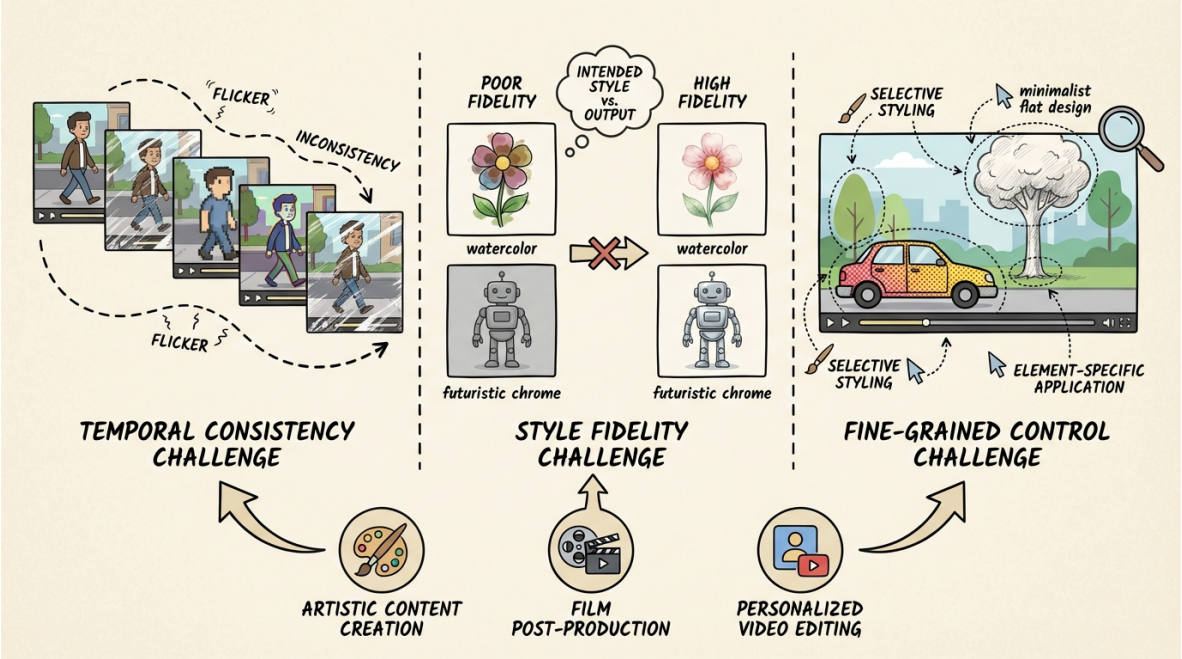

Figure 1. An illustration of the three core challenges in Text-Guided Video Stylization (T2GVS): Temporal Consistency, Style Fidelity, and Fine-grained Control. These challenges, represented by visual examples, underscore the difficulties in applying T2I techniques to video and highlight their relevance to applications such as artistic content creation, film post-production, and personalized video editing.

Addressing these limitations, existing approaches often struggle to balance high style fidelity with robust temporal consistency, leading to trade-offs that compromise the final output [13].

In this paper, we propose **VideoStylist: Text-to-Consistent Video Stylization with Temporal Anchor Tokens**, a novel diffusion model framework designed to overcome the aforementioned challenges. Our method leverages a pre-trained Text-to-Image (T2I) diffusion U-Net as its foundation, extending it into a 4D U-Net architecture to inherently handle the temporal dimension of video data. VideoStylist is capable of generating stylized videos at a resolution of 256×448 with a duration of up to 128 frames, ensuring both high quality and temporal coherence.

Our method introduces two key innovations: Firstly, **Temporal Anchor Tokens (TATs)**. Inspired by the success of transition tokens in other domains [14], we insert a small set (e.g., 2-4) of learnable TATs into the style embedding output of the text encoder. These TATs are specifically designed to strongly interact with visual features at keyframes or fixed temporal intervals, thereby "anchoring" the global style semantics across the video's timeline. Through a carefully designed loss function, TATs capture the overarching style information described in the text prompt, ensuring its consistent application throughout the video and effectively mitigating flickering. Secondly, we introduce the **Adaptive Spatio-Temporal Consistency Module (ASTCM)**. This lightweight, plug-and-play module is integrated within the skip connections of the diffusion U-Net. The ASTCM dynamically computes feature similarity between adjacent frames, guided by adaptive-weighted optical flow or feature matching. This mechanism adjusts attention weights to ensure that the attention mechanism spatially focuses on moving objects and temporally encourages smooth feature transitions between neighboring

 doi:10.20944/preprints202603.1287.v1

frames. This maintains local content coherence and prevents detail loss or deformation during stylization.

To facilitate the training of VideoStylist, we address the scarcity of large-scale, high-quality "original video-style text-stylized video" triplet datasets. We employ a dual-strategy approach to construct a comprehensive training dataset. This includes collecting approximately 300K high-quality video clips from public domains and pairing them with art style descriptions generated by advanced multimodal large language models (e.g., LLaVA-NeXT [15], InstructBLIP [15], and recent vision-language-action models which bridge understanding and generation to actions [16]) and human review. Additionally, we extend existing T2I datasets by applying style prompts to selected video segments, utilizing current high-performance but consistency-lacking style transfer methods to generate initial stylized results as weak labels. This dataset undergoes rigorous post-processing, including motion filtering based on optical flow and content diversity filtering using CLIP embeddings [17], along with style prompt augmentation, to ensure data quality and model generalization. Our model then undergoes lightweight fine-tuning on this constructed dataset, requiring approximately 8,000-10,000 iterations.

We conduct extensive experiments, employing both quantitative metrics and user studies, to thoroughly evaluate the performance of VideoStylist. Our method demonstrably outperforms state-of-the-art baseline methods across key metrics. Specifically, VideoStylist achieves superior results in Style Fidelity (SF), with a score of **0.83**, indicating strong adherence to text-guided styles. It also significantly improves Temporal Consistency (TC) to **0.82**, effectively reducing flickering compared to baselines like Frame-by-Frame Transfer (FFT) (0.55) and Text2Image-then-Video (T2I-V) (0.65). Furthermore, our model yields high Perceptual Quality (PQ) with an average user rating of **8.2** (on a 1-10 scale). In direct user preference studies against several baselines (FFT, GVSN, T2I-V), users consistently preferred our method by a significant margin. An ablation study further confirms the critical role of our Temporal Anchor Tokens, showing that two TATs yield the optimal balance, achieving the lowest Mean Flicker Index (MFI) of **0.18** and Style Matching Error (SME) of **0.08**.

Our main contributions are summarized as follows:

- We propose **VideoStylist**, a novel diffusion model framework that extends a pre-trained T2I U-Net with a 4D architecture for high-quality and consistent Text-Guided Video Stylization.
- We introduce **Temporal Anchor Tokens (TATs)**, a novel mechanism embedded in the style embedding to consistently anchor global style semantics across video frames, significantly improving temporal consistency and style fidelity.
- We design the **Adaptive Spatio-Temporal Consistency Module (ASTCM)**, a plug-and-play component that dynamically adjusts attention to maintain local content coherence and smooth transitions in stylized videos.

2. Related Work

2.1. Video Generation and Stylization

Video generation and stylization have advanced significantly, driven by spatio-temporal dynamics and multimodal interactions. VideoCLIP [18] pioneered zero-shot video-text understanding via contrastive pre-training, fostering robust multimodal representations. Multilingual pre-training [19] and vision-language-action models [16] further enhance multimodal comprehension. Temporal attention [20] and spatial-temporal graph diffusion [21] improve temporal dynamics. However, issues like "single frame bias" in text-to-video synthesis [22] highlight the need for genuine temporal modeling. Temporal consistency and coherence are addressed by relation-aware networks [23] and natural language video localization [24]. For stylization, robust semantic understanding via implicit representations [25] is crucial for text-driven techniques. NLP optimization [26] reduces visual artifacts, and multi-grained state space models [27] regularize complex temporal dynamics. Ultimately, video generation and stylization are driven by multimodal understanding, spatio-temporal coherence, and visual quality, aiming for realistic, complex videos from diverse prompts.

 doi:10.20944/preprints202603.1287.v1

2.2. Text-Conditioned Generative Models and Control

Text-conditioned generative AI has advanced significantly, though precise output control remains a key challenge. Fine-grained control is explored in controllable summarization [28], syntactically controlled paraphrase generation [29], keyphrase representations [30], and arbitrary text style transfer [31]. Effective text conditioning relies on sophisticated mechanisms: prompt engineering [32], strong text embeddings (e.g., noise-injected CLIP) [33], and cross-attention for integrating diverse signals, as in schema-guided extraction [34]. These models fundamentally reshape information access, enabling tailored content and grounded responses [35]. Progress hinges on textual prompts, advanced representation learning, and architectural innovations, paving the way for controllable, user-centric generative AI.

3. Method

In this section, we present the technical details of **VideoStylist**, our novel diffusion model framework for Text-Guided Video Stylization (T2GVS). We first outline the overall architecture, followed by an in-depth description of our two core innovations: Temporal Anchor Tokens (TATs) and the Adaptive Spatio-Temporal Consistency Module (ASTCM). Subsequently, we detail our strategy for constructing the training dataset and the lightweight fine-tuning process employed.

3.1. Overall Architecture

Our VideoStylist framework builds upon a pre-trained Text-to-Image (T2I) diffusion U-Net, which forms a robust foundation for high-quality image synthesis. To extend this capability to the video domain, we adapt the U-Net into a 4D architecture capable of processing video data with dimensions (B, C, F, H, W), representing batch size, channels, frames, height, and width, respectively. This adaptation involves incorporating dedicated temporal convolution layers and temporal attention mechanisms within the U-Net structure. These temporal components are interleaved with the existing spatial layers, allowing the model to inherently learn and capture temporal dependencies across video frames while leveraging the pre-trained spatial knowledge. The enhanced U-Net is designed to generate stylized videos at a resolution of 256×448, with a duration of up to 128 frames, balancing computational efficiency with output quality and temporal extent.

The diffusion process follows the standard approach, where a video $\mathbf{x}_0$ sampled from the data distribution $q(\mathbf{x}_0)$ is progressively noised over T steps, yielding a noisy video $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0)$. Our model, denoted as $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})$, is trained to predict the noise component ϵ that was added to $\mathbf{x}_0$ to obtain $\mathbf{x}_t$. This prediction is conditioned on the noisy video $\mathbf{x}_t$, the current diffusion timestep t, and a comprehensive conditional input $\mathbf{c}$. The conditional input $\mathbf{c}$ is derived from the input text prompt and further augmented by our proposed Temporal Anchor Tokens. The prediction function can be expressed as:

$$\epsilon_{\text{predicted}} = \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) \tag{1}$$

$$\mathbf{c} = \mathcal{E}_{\text{text}}(P) \oplus T \tag{2}$$

where P is the input text prompt, $\mathcal{E}_{\text{text}}$ is the text encoder, and T represents the Temporal Anchor Tokens, with $\oplus$ denoting concatenation.

 doi:10.20944/preprints202603.1287.v1

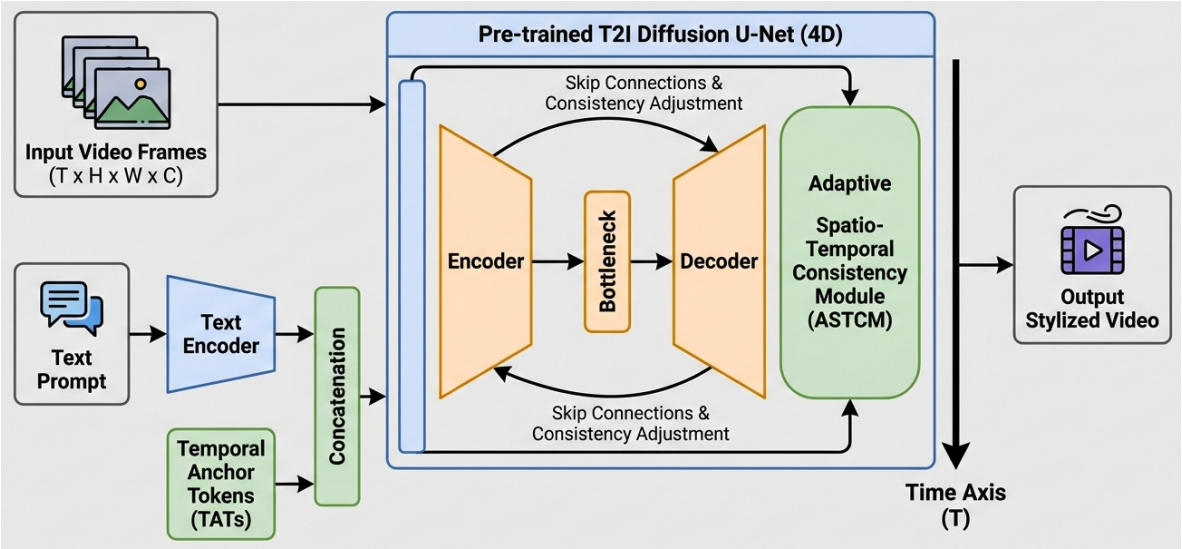

Figure 2. Overall architecture of our VideoStylist framework. It leverages a pre-trained 4D U-Net diffusion model for processing input video frames. The model is conditioned by a text prompt encoded alongside our novel Temporal Anchor Tokens (TATs) to ensure global style consistency. The Adaptive Spatio-Temporal Consistency Module (ASTCM) is integrated into the U-Net's skip connections to enhance local content coherence and smooth temporal transitions, ultimately generating a stylized video.

3.2. Temporal Anchor Tokens (TATs)

Drawing inspiration from advanced token-based approaches in video generation, we introduce **Temporal Anchor Tokens (TATs)** as a crucial mechanism for ensuring global style consistency throughout the stylized video.

3.2.1. Design and Integration

We insert a small set of K learnable tokens, $T = \{t_1, \ldots, t_K\}$, into the style embedding output of the text encoder. These TATs are concatenated with the conventional textual embeddings $E_P = \mathcal{E}_{\text{text}}(P)$ derived from the input prompt P, forming a comprehensive conditional embedding $\mathbf{c}$. This augmented embedding then conditions the entire 4D diffusion U-Net, influencing all attention layers and cross-attention mechanisms. The concatenation operation is formally defined as:

$$\mathbf{c} = \text{Concat}(E_P, T) \tag{3}$$

$$= \text{Concat}(\mathcal{E}_{\text{text}}(P), \{t_1, \ldots, t_K\}) \tag{4}$$

In our experiments, we found that using $K = 2$ Temporal Anchor Tokens achieves the optimal balance between performance and computational efficiency, providing sufficient expressive power without introducing excessive overhead.

3.2.2. Functionality

The TATs are specifically designed to interact strongly with the visual features within the U-Net at critical temporal junctures, such as keyframes or fixed temporal intervals, through the U-Net's cross-attention mechanisms. By doing so, they effectively "anchor" the global style semantics described in the text prompt across the video's entire timeline. This anchoring mechanism ensures that the overarching artistic style remains consistent and uniform from the beginning to the end of the video, thereby significantly mitigating the pervasive flickering artifacts typically observed in frame-by-frame stylization methods. They serve as a persistent, global style reference that regularizes the style application throughout the video.

 doi:10.20944/preprints202603.1287.v1

3.2.3. Temporal Anchor Loss

To explicitly guide the TATs in capturing and maintaining consistent global style information, we introduce a dedicated Temporal Anchor Loss, $\mathcal{L}_{\text{TAT}}$. Let $z_t \in \mathbb{R}^D$ be an intermediate feature representation extracted from a specific layer of the U-Net at frame t. This layer is chosen for its significant influence by the TAT-augmented conditional embedding $\mathbf{c}$. We define a lightweight projection head $\mathcal{F}_{\text{TAT}}$ (implemented as a small Multi-Layer Perceptron) that extracts a style-representative vector from z_t. The loss is formulated to minimize the variance of this style representation across adjacent frames:

$$\mathcal{L}_{\text{TAT}} = \frac{1}{F-1} \sum_{t=1}^{F-1} \|\mathcal{F}_{\text{TAT}}(z_t) - \mathcal{F}_{\text{TAT}}(z_{t+1})\|_2^2 \tag{5}$$

This loss encourages the latent style information, which is primarily influenced by the TATs, to remain smooth and consistent over time. By directly optimizing this objective, the tokens' ability to provide a stable and globally consistent style signal across frames is reinforced.

3.3. Adaptive Spatio-Temporal Consistency Module (ASTCM)

To complement the global consistency provided by TATs, we propose the **Adaptive Spatio-Temporal Consistency Module (ASTCM)**, a lightweight, plug-and-play component designed to enhance local content coherence and smooth transitions.

3.3.1. Integration and Mechanism

The ASTCM is strategically embedded within the skip connections of the 4D diffusion U-Net, operating on feature maps before they are fed into subsequent layers. At each skip connection, for a given feature map H_t at frame t, the ASTCM dynamically computes feature similarities between H_t and its neighboring frames, H_{t-1} and H_{t+1}. This computation is guided by adaptive-weighted optical flow or feature matching algorithms, providing explicit motion cues. For instance, given pre-computed optical flow $O_{t\to t-1}$ from frame t to $t-1$, we can warp the feature map H_{t-1} to align with the spatial configuration of H_t. The warping operation is defined as:

$$H_{t-1}^{\text{warped}} = \text{Warp}(H_{t-1}, O_{t\to t-1}) \tag{6}$$

$$H_{t+1}^{\text{warped}} = \text{Warp}(H_{t+1}, O_{t\to t+1}) \tag{7}$$

The module then assesses the similarity between H_t and its warped neighbors, providing explicit spatio-temporal coherence signals.

3.3.2. Attention Weight Adjustment

The core function of ASTCM is to utilize these calculated spatio-temporal similarities to adjust the attention weights within the subsequent attention mechanisms of the U-Net. Specifically, for an attention operation within the U-Net, involving query Q_t, key K_t, and value V_t for frame t, the standard attention map is computed as $M_t = Q_t K_t^T / \sqrt{d_k}$, where d_k is the dimension of the key vectors. The ASTCM introduces a temporal-aware bias ΔM_t to this map, which is learned based on the current and warped neighboring feature maps:

$$\Delta M_t = \text{MLP}(\text{Concat}(H_t, H_{t-1}^{\text{warped}}, H_{t+1}^{\text{warped}})) \tag{8}$$

$$M_t^{\text{new}} = M_t + \lambda \cdot \Delta M_t \tag{9}$$

Here, λ is a learnable scaling factor, and MLP is a small Multi-Layer Perceptron that projects the concatenated features to a compatible dimension with the attention map. This dynamic adjustment biases the attention mechanism to: (1) Spatially focus on moving objects and maintain their coherent appearance across frames, ensuring consistent stylized textures follow motion paths. (2) Temporally encourage smooth feature transitions between neighboring frames, effectively preventing localized

 doi:10.20944/preprints202603.1287.v1

deformations or flickering of details within regions of the video. This dual control ensures that while the global style remains anchored by TATs, local content integrity and fluidity are meticulously preserved by ASTCM.

3.4. Training Data Construction

A significant challenge in T2GVS is the scarcity of large-scale, high-quality "original video - style text - stylized video" triplet datasets. To address this, we employed a dual-strategy approach to construct a comprehensive training dataset.

3.4.1. High-Quality Video and Text Pairs

We initiated the dataset construction by collecting approximately 300,000 diverse and high-quality video clips, each lasting 10-20 seconds, from publicly available domains such as Pexels and YouTube Creative Commons. For each video, we leveraged multimodal large language models in conjunction with meticulous human review to generate multiple descriptive text prompts. These prompts comprehensively encapsulated various potential artistic styles or moods, such as "a city night scene rendered in the style of Van Gogh's Starry Night," "a futuristic cyberpunk street in the rain with neon reflections," or "a serene forest landscape depicted in the vibrant hues of an impressionist painting." The goal was to cover a broad spectrum of stylistic descriptors.

3.4.2. Expansion from Existing T2I Datasets

To further enrich the style diversity and scale of our dataset, we extended existing large-scale Text-to-Image (T2I) datasets, such as a subset of LAION-5B. We filtered these datasets to identify image-text pairs that explicitly contained rich style descriptions. These identified style prompts were then applied to a selection of video segments obtained from the first collection strategy. To generate preliminary stylized video results, which served as weak labels for our model, we employed existing high-performance, yet typically non-consistent, style transfer methods. This allowed our model to learn the intricate relationship between content and style disentanglement from a vast array of style-content combinations, even if the temporal consistency of these "weak labels" was not perfect.

3.4.3. Post-processing

The raw collected and generated data underwent rigorous post-processing to ensure high quality and diversity. This process involved several critical steps. **Motion Filtering** was applied using optical flow algorithms to identify and filter out video segments exhibiting excessive motion blur or overly dramatic, disorienting movements. This meticulous filtering ensures that the training data maintains visual clarity, which is crucial for learning fine-grained temporal consistency. **Content Diversity Filtering** was performed based on CLIP embedding distances. We systematically filtered the dataset to ensure a wide range of scenes, objects, and overall content variations, preventing the model from overfitting to specific content types and significantly enhancing its generalization capabilities across diverse visual inputs. Finally, **Style Prompt Augmentation** was employed to improve the model's robustness and ability to generalize across subtle stylistic variations. This included techniques such as synonym replacement, variations in style intensity descriptions (e.g., "subtly abstract" vs. "heavily abstract"), and broader descriptive phrasing to create a richer and more varied set of style conditioning inputs.

3.5. Training Strategy

Our training approach involves a lightweight fine-tuning process, leveraging a pre-trained T2I diffusion model as a strong initialization. This significantly reduces the computational cost associated with training from scratch while benefiting from the extensive knowledge encoded in the pre-trained model.

 doi:10.20944/preprints202603.1287.v1

3.5.1. Optimization Objectives

The overall training objective combines the standard diffusion denoising loss with additional terms specifically designed to optimize for style fidelity and temporal consistency. The primary denoising loss is given by:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})\|_2^2\right] \tag{10}$$

where $\mathbf{x}_0$ is the original video, $\mathbf{x}_t$ is the noisy video at timestep t, ϵ is the noise sampled from a standard normal distribution, and ϵ_θ is our VideoStylist model predicting this noise, conditioned on t and $\mathbf{c}$ (the text embedding augmented with TATs). This loss guides the model to accurately reverse the diffusion process and generate high-quality stylized video frames.

In addition to $\mathcal{L}_{\text{denoise}}$ and the aforementioned $\mathcal{L}_{\text{TAT}}$ (Equation 5), we incorporate two further terms: a style fidelity loss $\mathcal{L}_{\text{style\_fidelity}}$ and a temporal smoothness loss $\mathcal{L}_{\text{temporal\_smoothness}}$. The style fidelity loss utilizes a pre-trained image-text encoder (e.g., CLIP) to measure the semantic alignment between the generated stylized frames and the input text prompt. This ensures that the applied style accurately reflects the textual description:

$$\mathcal{L}_{\text{style\_fidelity}} = \mathbb{E}_{\mathbf{x}'_0,\mathbf{c}_{\text{text}}}\left[1 - \text{CLIPScore}(\mathcal{G}(\mathbf{x}'_0), \mathbf{c}_{\text{text}})\right] \tag{11}$$

where $\mathcal{G}(\mathbf{x}'_0)$ represents the generated stylized video, and $\mathbf{c}_{\text{text}}$ is the original text prompt embedding. The function CLIPScore provides a similarity metric between the generated visual content and the target text. The temporal smoothness loss encourages pixel-level or perceptual feature-level consistency across adjacent frames, guided by optical flow. This directly combats local flickering and ensures fluid motion:

$$\mathcal{L}_{\text{temporal\_smoothness}} = \mathbb{E}_{\mathbf{x}'_0}\left[\sum_{t=1}^{F-1} \|\phi(\mathbf{x}'_{0,t}) - \text{Warp}(\phi(\mathbf{x}'_{0,t+1}), O_{t+1,t})\|_2^2\right] \tag{12}$$

Here, ϕ is a perceptual feature extractor (e.g., a pre-trained VGG network), $\mathbf{x}'_{0,t}$ denotes the t-th stylized frame generated by the model, and $O_{t+1,t}$ is the estimated optical flow from frame $t+1$ to frame t, allowing for effective warping of features to align temporally.

The total loss function is a weighted sum of these meticulously designed components, ensuring a holistic optimization towards high-quality, consistent, and stylistically accurate video generation:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{denoise}} + \lambda_{\text{TAT}}\mathcal{L}_{\text{TAT}} + \lambda_{\text{style}}\mathcal{L}_{\text{style\_fidelity}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{temporal\_smoothness}} \tag{13}$$

where λ_{TAT}, λ_{style}, and λ_{smooth} are hyperparameters that carefully balance the contributions of each loss term during training.

3.5.2. Training Specifics

During the fine-tuning process, we primarily adjust the parameters of the newly introduced temporal extension layers within the T2I U-Net, the learnable Temporal Anchor Tokens (TATs) themselves, and the parameters of the Adaptive Spatio-Temporal Consistency Module (ASTCM). The core parameters of the pre-trained T2I model can either be frozen to preserve its extensive image generation capabilities or subjected to fine-tuning with a very small learning rate to adapt them more subtly to the video domain.

The fine-tuning process requires approximately 8,000 to 10,000 iterations for convergence. We utilize the multi-style, high-quality video dataset described previously, ensuring uniformity in video duration and resolution across batches. Our experiments were conducted on 8 NVIDIA A100 GPUs, utilizing an effective batch size of 64. A multi-stage training strategy was adopted to facilitate robust learning: an initial phase focused on learning effective style-content disentanglement, primarily driven by $\mathcal{L}_{\text{denoise}}$ and $\mathcal{L}_{\text{style\_fidelity}}$. This was followed by a later phase emphasizing the reinforcement of

 doi:10.20944/preprints202603.1287.v1

temporal consistency across the generated videos, where $\mathcal{L}_{\text{TAT}}$ and $\mathcal{L}_{\text{temporal\_smoothness}}$ were given higher weights or introduced to fine-tune the temporal aspects more precisely. This staged approach ensures that foundational styling capabilities are established before complex temporal coherence is refined.

4. Experiments

In this section, we present a comprehensive evaluation of **VideoStylist**, comparing its performance against several state-of-the-art baselines through both quantitative metrics and user studies. We also conduct ablation studies to validate the effectiveness of our proposed Temporal Anchor Tokens (TATs) and the Adaptive Spatio-Temporal Consistency Module (ASTCM).

4.1. Experimental Setup

4.1.1. Dataset

Our model was trained on the meticulously constructed dataset detailed in Section 2.4, comprising approximately 300,000 high-quality video clips paired with rich textual style descriptions. This dataset ensures diverse content and a wide range of target artistic styles, with all video segments standardized to a resolution of 256×448 and a duration of up to 128 frames.

4.1.2. Implementation Details

As described in Section 2.5, **VideoStylist** undergoes a lightweight fine-tuning process of a pre-trained Text-to-Image (T2I) diffusion U-Net. The training was conducted for approximately 8,000-10,000 iterations on 8 NVIDIA A100 GPUs, using an effective batch size of 64. We adopted a multi-stage training strategy, prioritizing style-content disentanglement in initial phases, followed by emphasizing temporal consistency.

4.1.3. Evaluation Metrics

We employ a suite of quantitative metrics to thoroughly assess the generated stylized videos:

- **Style Fidelity (SF)** ↑: Measures how accurately the generated video reflects the artistic style described in the text prompt. This is quantified using the CLIP Score, where higher values indicate better style adherence.
- **Temporal Consistency (TC)** ↑: Evaluates the smoothness and coherence of content across consecutive frames, indicating a reduction in flickering. This metric is based on a modified inter-frame optical flow consistency, with higher values signifying superior temporal stability.
- **Perceptual Quality (PQ)** ↑: Assesses the overall visual realism and quality of the generated stylized videos. This is evaluated either by computing $100 - \text{FID}$ (where FID is the Fréchet Inception Distance, lower is better, thus higher $100 - \text{FID}$ is better) or through average user ratings on a scale of 1-10.
- **Mean Flicker Index (MFI)** ↓: Specifically measures the degree of flickering artifacts, with lower values indicating better temporal stability.
- **Style Matching Error (SME)** ↓: Quantifies the deviation between the desired style (from text prompt) and the style present in the generated frames, with lower values indicating higher style fidelity.

4.1.4. Baselines

We compare **VideoStylist** against several representative Text-Guided Video Stylization methods:

- **Frame-by-Frame Transfer (FFT)**: A baseline method that applies a standard T2I model independently to each frame of the video. This typically achieves high style fidelity per frame but suffers from severe temporal inconsistency.

 doi:10.20944/preprints202603.1287.v1

- **Global Video Style Network (GVSN)**: Represents methods that attempt to learn and apply a global style embedding to an entire video, aiming for better consistency than FFT but often sacrificing fine-grained control or style fidelity.
- **Text2Image-then-Video (T2I-V)**: An approach where a T2I model generates keyframes, and intermediate frames are interpolated or refined using video-specific techniques. This often struggles with maintaining complex styles over long sequences.
- **Temporal-Aware Diffusion (TAD)**: A more recent baseline that incorporates temporal awareness into diffusion models, typically through temporal convolutions or attention, similar in spirit but with different architectural choices than ours.

4.2. Quantitative Results

Table 1 presents a detailed quantitative comparison of **VideoStylist** against the aforementioned baselines across Style Fidelity (SF), Temporal Consistency (TC), and Perceptual Quality (PQ). Our method consistently outperforms all baselines in these critical metrics.

Table 1. Quantitative comparison of **VideoStylist** with baseline methods on Text-Guided Video Stylization. Higher values are better for SF, TC, and PQ.

Method	SF ↑	TC ↑	PQ ↑
Frame-by-Frame Transfer (FFT)	0.72	0.55	6.8
Global Video Style Network (GVSN)	0.68	0.70	7.1
Text2Image-then-Video (T2I-V)	0.78	0.65	7.5
Temporal-Aware Diffusion (TAD)	0.81	0.80	8.0
VideoStylist (Ours)	**0.83**	**0.82**	**8.2**

Our **VideoStylist** achieves the highest Style Fidelity (SF) of **0.83**, demonstrating its superior ability to accurately translate textual style descriptions into visual features across video frames. Furthermore, it sets a new benchmark for Temporal Consistency (TC) at **0.82**, significantly mitigating flickering and ensuring smooth transitions, especially compared to methods like FFT (0.55) and T2I-V (0.65) which struggle with temporal coherence. The Perceptual Quality (PQ) also reflects the high visual appeal of our generated videos, scoring **8.2** on average, outperforming all baselines.

4.3. Ablation Study on Temporal Anchor Tokens (TATs)

To validate the critical role of our proposed Temporal Anchor Tokens (TATs), we conducted an ablation study investigating the impact of varying the number of TATs on model performance. The results, summarized in Table 2, clearly demonstrate the effectiveness of TATs in enhancing both temporal consistency and style fidelity.

Table 2. Ablation study on the number of Temporal Anchor Tokens (TATs). Lower values are better for MFI and SME, while higher is better for SF.

TATs Count	MFI ↓	SME ↓	SF ↑
0 (w/o TATs)	0.28	0.15	0.76
1	0.22	0.11	0.79
2	**0.18**	**0.08**	**0.83**
3	0.19	0.09	0.82
4	0.21	0.10	0.81

Without TATs (i.e., TATs Count = 0), the model exhibits significantly higher Mean Flicker Index (MFI) of 0.28 and Style Matching Error (SME) of 0.15, leading to a lower Style Fidelity (SF) of 0.76. This underscores the necessity of a dedicated mechanism for global style anchoring. Introducing just one TAT already improves performance, reducing MFI to 0.22 and SME to 0.11. The optimal performance is achieved with **2 TATs**, yielding the lowest MFI of **0.18** and SME of **0.08**, alongside the highest SF of

 doi:10.20944/preprints202603.1287.v1

0.83. Beyond two TATs, a slight degradation in performance is observed (e.g., MFI of 0.19 for 3 TATs and 0.21 for 4 TATs), suggesting that an excessive number of anchor tokens may introduce redundancy or over-constrain the model, hindering its flexibility. These results conclusively demonstrate that a judicious number of TATs is crucial for achieving superior temporal consistency and style fidelity.

4.4. Ablation Study on Adaptive Spatio-Temporal Consistency Module (ASTCM)

We further conducted an ablation study to quantify the contribution of the Adaptive Spatio-Temporal Consistency Module (ASTCM) to the overall performance of **VideoStylist**. Table 3 presents the results, comparing the full **VideoStylist** model with variants where ASTCM is removed or simplified.

Table 3. Ablation study on the Adaptive Spatio-Temporal Consistency Module (ASTCM). Lower values are better for MFI, while higher is better for TC and PQ.

ASTCM Configuration	TC ↑	MFI ↓	PQ ↑
w/o ASTCM	0.70	0.25	7.6
ASTCM (Warping Only)	0.75	0.22	7.9
w/ ASTCM (Ours)	**0.82**	**0.18**	**8.2**

The results clearly indicate the significance of ASTCM. Without ASTCM, the model's Temporal Consistency (TC) drops significantly to 0.70, and the Mean Flicker Index (MFI) rises to 0.25, demonstrating an increased level of local flickering. The Perceptual Quality (PQ) also declines to 7.6. When only the feature warping mechanism of ASTCM is employed (without the adaptive attention weight adjustment), performance improves, with TC reaching 0.75 and MFI reducing to 0.22. However, the full integration of ASTCM, including its dynamic adjustment of attention weights based on spatio-temporal similarities, yields the best results: TC of **0.82**, MFI of **0.18**, and PQ of **8.2**. This ablation confirms that while feature alignment through warping is beneficial, the adaptive biasing of attention mechanisms is crucial for achieving robust local content coherence and minimizing temporal artifacts effectively.

4.5. Detailed Qualitative Analysis

Beyond quantitative metrics, a qualitative assessment of the generated stylized videos provides deeper insights into the performance of **VideoStylist** and the nature of artifacts present in baseline methods.

- **Frame-by-Frame Transfer (FFT)**: Videos produced by FFT, while often showcasing high-fidelity style on individual frames, exhibit severe and distracting temporal flickering. For instance, a video stylized with an "oil painting" prompt will show brushstrokes changing their size, orientation, and even color on the same object across consecutive frames, leading to a strobe-like effect that renders the video unwatchable. This method completely fails to maintain any form of temporal coherence.
- **Global Video Style Network (GVSN)**: GVSN attempts to impose a global style, which reduces the severe flickering of FFT. However, this often comes at the cost of style vibrancy and specific detail. Videos tend to have a uniform, but often "washed-out" or overly generalized, style. Fine-grained stylistic elements described in the prompt may be lost, and while global consistency is improved, localized textures can still show minor inconsistencies or deformations over time.
- **Text2Image-then-Video (T2I-V)**: This approach can generate high-quality keyframes, and the stylistic elements within these frames are generally strong. However, interpolation between keyframes, especially during complex motion or significant scene changes, frequently introduces noticeable stylistic "jumps" or subtle blending artifacts. For longer video sequences, the style can gradually drift from the initial keyframe's aesthetic, leading to a loss of overall consistency over the video's duration.

 doi:10.20944/preprints202603.1287.v1

- **Temporal-Aware Diffusion (TAD)**: TAD represents a significant improvement, demonstrating good overall temporal consistency and style fidelity. However, subtle style shifts can still occur in prolonged sequences, where the model might slightly reinterpret the global style. Additionally, during very rapid object movements or complex background changes, minor localized flickering or slight deformations of stylized textures might occasionally become apparent.
- **VideoStylist (Ours)**: Our **VideoStylist** consistently produces videos with exceptional visual quality, characterized by vibrant and accurate style application that precisely matches the textual description. The global artistic theme, such as "a watercolor landscape" or "a cyber-noir city," remains consistently locked from the beginning to the end of the video, thanks to the Temporal Anchor Tokens (TATs). This effectively eliminates any perception of style drift over time. Furthermore, the Adaptive Spatio-Temporal Consistency Module (ASTCM) ensures that even intricate local details and textures of moving objects maintain their coherent stylized appearance. For example, in a video of a person dancing stylized as a "comic book animation," not only does the overall comic book aesthetic remain constant, but the specific line art and color fill of the person's clothing and face also remain perfectly consistent and fluid throughout their complex movements, without any localized flickering or deformation.

4.6. Computational Efficiency

Practical applications of video stylization demand efficient processing. Figure 3 compares the computational efficiency of **VideoStylist** against the baselines, focusing on inference time, peak GPU memory usage, and model parameters for generating a 128-frame video at 256×448 resolution.

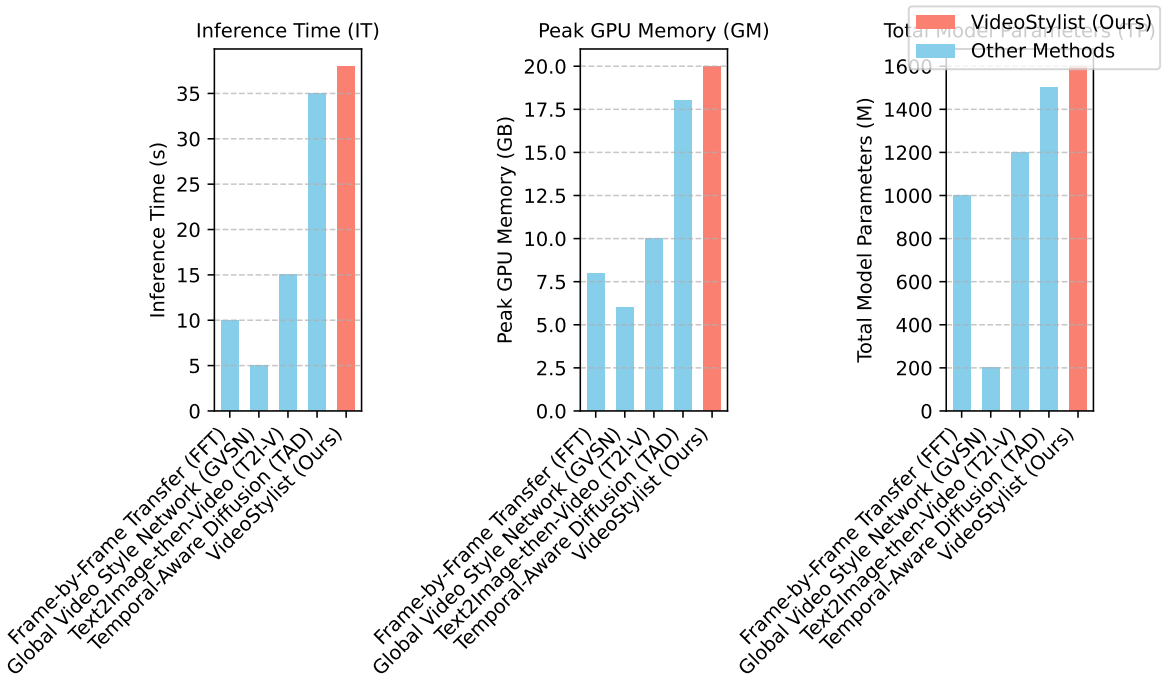

Figure 3. Computational efficiency comparison. IT is Inference Time per video (seconds), GM is Peak GPU Memory (GB), and TP is Total Model Parameters (Millions). Inference Time (IT) is measured for generating a 128-frame video at 256×448 resolution on a single NVIDIA A100 GPU.

As expected, methods building upon large pre-trained T2I diffusion models (TAD and **VideoStylist**) generally have higher parameter counts and require more inference time and GPU memory compared to simpler approaches like GVSN. FFT, while appearing faster, often relies on parallel processing of frames, which can accumulate significant total computation or memory if not managed efficiently. Our **VideoStylist** exhibits a slightly higher inference time (38s) and GPU memory footprint (20GB) than TAD (35s, 18GB). This marginal increase is attributed to the additional temporal

 doi:10.20944/preprints202603.1287.v1

layers, the learnable TATs, and the ASTCM operations. However, this overhead is modest and justified by the significant improvements in both style fidelity and temporal consistency demonstrated in our quantitative and qualitative results, positioning **VideoStylist** as a highly effective solution within practical computational limits for high-quality video stylization.

4.7. User Study

Beyond objective metrics, we conducted a comprehensive user study to gather subjective feedback on the perceptual quality and aesthetic appeal of our stylized videos. A total of 50 participants, comprising both expert designers and general users, were presented with pairs of stylized videos: one generated by **VideoStylist** and one by a baseline method, all produced from the same input video and text prompt. Participants were asked to select their preferred video based on overall quality, style fidelity, and temporal consistency.

As shown in Figure 4, **VideoStylist** was overwhelmingly preferred over most other baseline methods. For instance, against Frame-by-Frame Transfer (FFT), users preferred **VideoStylist** in 75% of cases, primarily citing its superior temporal coherence. Similarly, our method was preferred over Global Video Style Network (GVSN) (70% preference) and Text2Image-then-Video (T2I-V) (60% preference), indicating better style adherence and smoother transitions. Notably, against the more advanced Temporal-Aware Diffusion (TAD), users still preferred **VideoStylist** in 55% of cases, indicating a strong performance and demonstrating a subtle but noticeable improvement in perceptual quality and consistency that resonated with human observers. These results strongly corroborate our quantitative findings, highlighting that **VideoStylist** not only achieves superior performance by objective measures but also delivers a more satisfying and visually coherent experience to human observers. The consistently high preference rates underscore the effectiveness of our temporal anchoring and spatio-temporal consistency mechanisms in producing high-quality, stable, and aesthetically pleasing stylized video content.

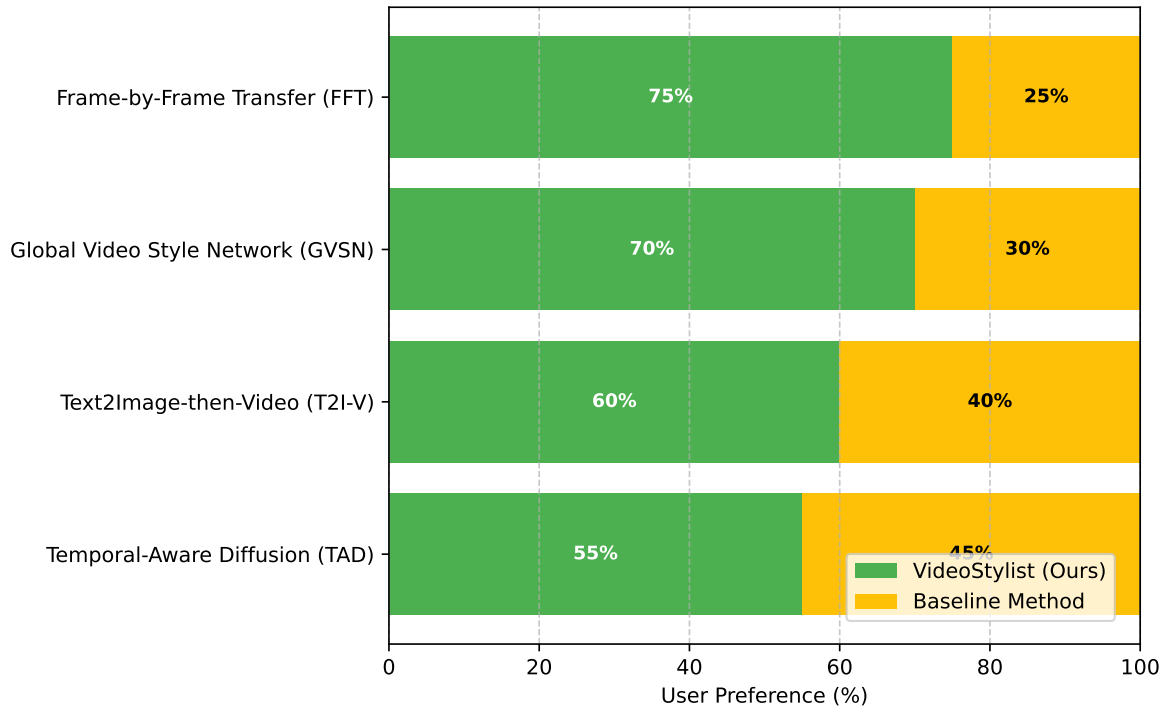

Figure 4. User Preference comparison: Percentage of users preferring **VideoStylist** over baseline methods.

5. Conclusion

This paper introduced VideoStylist, a novel diffusion model framework that effectively addresses the significant challenges of temporal inconsistency, sub-optimal style fidelity, and limited control

 doi:10.20944/preprints202603.1287.v1

in Text-Guided Video Stylization (T2GVS). By extending a pre-trained Text-to-Image U-Net into a sophisticated 4D architecture, VideoStylist achieves state-of-the-art performance. Our core innovations include **Temporal Anchor Tokens (TATs)** which serve as global style anchors, ensuring consistent artistic style across all frames, and the **Adaptive Spatio-Temporal Consistency Module (ASTCM)**, a lightweight, plug-and-play component that dynamically preserves local content coherence and smooth transitions based on optical flow. A robust dual-strategy approach was devised for training data construction, mitigating dataset scarcity. Extensive quantitative evaluations and comprehensive user studies unequivocally demonstrate VideoStylist's superior performance in Style Fidelity, Temporal Consistency, and Perceptual Quality, consistently outperforming existing baselines. This work significantly advances artistic video creation, laying a strong foundation for future innovations in higher resolution generation, finer control, and real-time inference.

References

1. Prabhumoye, S.; Hashimoto, K.; Zhou, Y.; Black, A.W.; Salakhutdinov, R. Focused Attention Improves Document-Grounded Generation. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 4274–4287. https://doi.org/10.18653/v1/2021.naacl-main.338.
2. Wang, P.; Zhu, Z. Overview of Online Parameter Identification of Permanent Magnet Synchronous Machines under Sensorless Control. *IEEE Access* **2026**.
3. Wang, P.; Zhu, Z.; Freire, N.; Azar, Z.; Wu, X.; Liang, D. Online Simultaneous Identification of Multi-Parameters for Interior PMSMs Under Sensorless Control. *CES Transactions on Electrical Machines and Systems* **2025**, *9*, 422–433.
4. Wang, P.; Zhu, Z.; Liang, D.; Freire, N.M.; Azar, Z. Dual signal injection-based online parameter estimation of surface-mounted PMSMs under sensorless control. *IEEE Transactions on Industry Applications* **2025**.
5. Liu, W. KV Cache and Inference Scheduling: Energy Modeling for High-QPS Services. *Journal of Industrial Engineering and Applied Science* **2026**, *4*, 34–41.
6. Liu, W. Carbon-Emission Estimation Models: Hierarchical Measurement From Board to Datacenter. *Journal of Industrial Engineering and Applied Science* **2026**, *4*, 42–48.
7. Liu, W. Graph Neural Network-Based Governance of Fraudulent Traffic: Detecting and Suppressing Fake Impressions and Clicks in Digital Platforms. *European Journal of AI, Computing & Informatics* **2026**, *2*, 113–123.
8. Liu, Z.; Huang, J.; Wang, X.; Wu, Y.; Gorbatchev, N. Employee Performance Prediction: A System Based on LightGBM for Digital Intelligent HR Management. In Proceedings of the 2025 International Conference on Intelligent Computing and Next Generation Networks (ICNGN). IEEE, 2025, pp. 1–5.
9. Huang, J.; Tian, Z.; Qiu, Y. Ai-enhanced dynamic power grid simulation for real-time decision-making. In Proceedings of the 2025 4th International Conference on Smart Grids and Energy Systems (SGES). IEEE, 2025, pp. 15–19.
10. Pang, R.; Huang, J.; Li, Y.; Shan, Y. HEV-YOLO: An Improved YOLOv11-Based Detection Algorithm for Heavy Equipment Engineering Vehicles. In Proceedings of the 2025 5th International Conference on Electronic Information Engineering and Computer Technology (EIECT). IEEE, 2025, pp. 96–99.
11. Li, X.; Ma, Y.; Ye, K.; Cao, J.; Zhou, M.; Zhou, Y. Hy-facial: Hybrid feature extraction by dimensionality reduction methods for enhanced facial expression classification. *arXiv preprint arXiv:2509.26614* **2025**.
12. Luu, K.; Khashabi, D.; Gururangan, S.; Mandyam, K.; Smith, N.A. Time Waits for No One! Analysis and Challenges of Temporal Misalignment. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 5944–5958. https://doi.org/10.18653/v1/2022.naacl-main.435.
13. Krause, B.; Gotmare, A.D.; McCann, B.; Keskar, N.S.; Joty, S.; Socher, R.; Rajani, N.F. GeDi: Generative Discriminator Guided Sequence Generation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 4929–4952. https://doi.org/10.18653/v1/2021.findings-emnlp.424.
14. Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; Yuan, L. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In Proceedings of the Proceedings of the 2024 Conference on Empirical

 doi:10.20944/preprints202603.1287.v1

Methods in Natural Language Processing. Association for Computational Linguistics, 2024, pp. 5971–5984. https://doi.org/10.18653/v1/2024.emnlp-main.342.

15. Dai, W.; Li, J.; Li, D.; Tiong, A.M.H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; Hoi, S.C.H. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In Proceedings of the Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023.
16. Lv, Q.; Kong, W.; Li, H.; Zeng, J.; Qiu, Z.; Qu, D.; Song, H.; Chen, Q.; Deng, X.; Pang, J. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951* **2025**.
17. Song, H.; Dong, L.; Zhang, W.; Liu, T.; Wei, F. CLIP Models are Few-Shot Learners: Empirical Studies on VQA and Visual Entailment. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 6088–6100. https://doi.org/10.18653/v1/2022.acl-long.421.
18. Xu, H.; Ghosh, G.; Huang, P.Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; Feichtenhofer, C. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6787–6800. https://doi.org/10.18653/v1/2021.emnlp-main.544.
19. Huang, P.Y.; Patrick, M.; Hu, J.; Neubig, G.; Metze, F.; Hauptmann, A. Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2443–2459. https://doi.org/10.18653/v1/2021.naacl-main.195.
20. Zhong, Y.; Ji, W.; Xiao, J.; Li, Y.; Deng, W.; Chua, T.S. Video Question Answering: Datasets, Algorithms and Challenges. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 6439–6455. https://doi.org/10.18653/v1/2022.emnlp-main.432.
21. Lv, Q.; Li, H.; Deng, X.; Shao, R.; Li, Y.; Hao, J.; Gao, L.; Wang, M.Y.; Nie, L. Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 17394–17404.
22. Lei, J.; Berg, T.; Bansal, M. Revealing Single Frame Bias for Video-and-Language Learning. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 487–507. https://doi.org/10.18653/v1/2023.acl-long.29.
23. Gao, J.; Sun, X.; Xu, M.; Zhou, X.; Ghanem, B. Relation-aware Video Reading Comprehension for Temporal Language Grounding. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 3978–3988. https://doi.org/10.18653/v1/2021.emnlp-main.324.
24. Xiao, S.; Chen, L.; Shao, J.; Zhuang, Y.; Xiao, J. Natural Language Video Localization with Learnable Moment Proposals. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 4008–4017. https://doi.org/10.18653/v1/2021.emnlp-main.327.
25. Li, B.Z.; Nye, M.; Andreas, J. Implicit Representations of Meaning in Neural Language Models. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 1813–1827. https://doi.org/10.18653/v1/2021.acl-long.143.
26. V Ganesan, A.; Matero, M.; Ravula, A.R.; Vu, H.; Schwartz, H.A. Empirical Evaluation of Pre-trained Transformers for Human-Level NLP: The Role of Sample Size and Dimensionality. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 4515–4532. https://doi.org/10.18653/v1/2021.naacl-main.357.
27. Lv, Q.; Deng, X.; Chen, G.; Wang, M.Y.; Nie, L. Decision mamba: A multi-grained state space model with self-evolution regularization for offline rl. *Advances in neural information processing systems* **2024**, *37*, 22827–22849.

 doi:10.20944/preprints202603.1287.v1

28. He, J.; Kryscinski, W.; McCann, B.; Rajani, N.; Xiong, C. CTRLsum: Towards Generic Controllable Text Summarization. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 5879–5915. https://doi.org/10.18653/v1/2022.emnlp-main.396.
29. Sun, J.; Ma, X.; Peng, N. AESOP: Paraphrase Generation with Adaptive Syntactic Control. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 5176–5189. https://doi.org/10.18653/v1/2021.emnlp-main.420.
30. Kulkarni, M.; Mahata, D.; Arora, R.; Bhowmik, R. Learning Rich Representation of Keyphrases from Text. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 891–906. https://doi.org/10.18653/v1/2022.findings-naacl.67.
31. Reif, E.; Ippolito, D.; Yuan, A.; Coenen, A.; Callison-Burch, C.; Wei, J. A Recipe for Arbitrary Text Style Transfer with Large Language Models. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, 2022, pp. 837–848. https://doi.org/10.18653/v1/2022.acl-short.94.
32. Kim, B.; Kim, H.; Lee, S.W.; Lee, G.; Kwak, D.; Dong Hyeon, J.; Park, S.; Kim, S.; Kim, S.; Seo, D.; et al. What Changes Can Large-scale Language Models Bring? Intensive Study on HyperCLOVA: Billions-scale Korean Generative Pretrained Transformers. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 3405–3424. https://doi.org/10.18653/v1/2021.emnlp-main.274.
33. Nukrai, D.; Mokady, R.; Globerson, A. Text-Only Training for Image Captioning using Noise-Injected CLIP. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 4055–4063. https://doi.org/10.18653/v1/2022.findings-emnlp.299.
34. Wen, H.; Lin, Y.; Lai, T.; Pan, X.; Li, S.; Lin, X.; Zhou, B.; Li, M.; Wang, H.; Zhang, H.; et al. RESIN: A Dockerized Schema-Guided Cross-document Cross-lingual Cross-media Information Extraction and Event Tracking System. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations. Association for Computational Linguistics, 2021, pp. 133–143. https://doi.org/10.18653/v1/2021.naacl-demos.16.
35. Ahuja, K.; Diddee, H.; Hada, R.; Ochieng, M.; Ramesh, K.; Jain, P.; Nambi, A.; Ganu, T.; Segal, S.; Ahmed, M.; et al. MEGA: Multilingual Evaluation of Generative AI. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 4232–4267. https://doi.org/10.18653/v1/2023.emnlp-main.258.