DOI: https://doi.org/10.15276/ict.03.2026.19
UDC 004.93

Context-aware video caption generation with VideoBERT and domain-aware adapters

Mariia S. Novichonok[1)]
ORCID: https://orcid.org/0009-0007-1787-0546; mariia.novichonok@nure.ua.
Sergii V. Mashtalir[1)]
ORCID: http://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
[1)] Kharkiv National University of Radio Electronic, 14, Nauky Ave. Kharkiv, 61166, Ukraine

ABSTRACT

Background. Automated video description is a relevant task for workplace reengineering and business process analysis, yet its real-world efficiency is affected by domain shift between training datasets and deployment environments. **Aim.** The study aims to develop a context-aware video caption generation architecture combining pretrained video representations with lightweight adaptation mechanisms. **Methods.** The proposed framework integrates a frozen twelve-layer VideoBERT encoder with a dual-branch Domain-Aware Adapter adaptation module and a three-layer autoregressive Transformer decoder. Visual features are extracted using an ImageNet-pretrained ResNet-50 backbone. The Domain-Invariant Adapter and Domain-Specific Adapter branches process contextualized representations in parallel and are combined through a trainable scalar fusion coefficient. A fifty percent stochastic Masked Language Modeling strategy is applied during training. **Results.** Evaluation on the Something-Something V2 benchmark shows that the complete VideoBERT, Domain-Aware Adapter and Transformer decoder configuration with fifty percent Masked Language Modeling achieves a Mean BLEU-4 of zero point one four six nine, Mean ROUGE-L of zero point three four seven four, and Mean METEOR of zero point two seven two zero. Component-wise ablation demonstrates improvements from both the dual-branch adaptation module and Masked Language Modeling regularization**. Conclusions.** The proposed architecture provides a modular, parameter-efficient framework for context-aware video caption generation, while explicit cross-domain transfer evaluation remains a direction for future research.

Keywords: context-aware video captioning; VideoBERT; Domain-Aware Adapter; domain adaptation; spatio-temporal representation; Transformer decoder; masked language modeling; parameter-efficient adaptation

For citation: **Novichonok M. S., Mashtalir S. V. "Context-aware video caption generation with VideoBERT and domain-aware adapters". *Informatics. Culture. Technology*. 2026; Vol. 3 No. 1(3): 221–235. DOI: https://doi.org/10.15276/ict.03.2026.19**

INTRODUCTION

Context-aware video understanding remains an important research task due to its potential applications across various fields and the rapidly growing volume of video content available for analysis. Existing video analysis methods are already helping to automate processes in medicine, construction, manufacturing, social network algorithms, etc. This article is a continuation of research to obtain an architecture that can analyze the video sequence of a particular production or institution and provide a detailed description of the processes performed by employees, which will subsequently become the ba-sis for reengineering by a business analyst. Previous publications have already highlighted the features [1] that must be considered when choosing video analysis methods, namely, determining the structure of the processes performed, their sequence, and the type of interaction between a person and an object or another person. Also, since the customer of reengineering can be an interested party from any field, the issue of multi-domain capability of the future solution becomes very important. In real-world applications, computer vision models often encounter a discrepancy between the data on which they were trained (source domain) and the data they work with during deployment (target domain).

This phenomenon is called domain shift, and it significantly reduces the quality of prediction [2], [3]. Therefore, the task of domain adaptation and generalization remains critically important for building systems capable of working effectively in changing domain conditions. It is this problem that receives the most attention in this work. In the context of this work, a domain refers to a specific distribution of visual environments, object categories and action contexts. The previous study investigated the conceptual integration of VideoBERT [4], Transformer decoding and a set of Domain-Aware Adapter (DA-Ada) [5] for contextual video analysis, focusing on the architectural design and theoretical justification. The present work extends that research by implementing the proposed architecture, introducing a lightweight DA-Ada integration strategy with a three-layer

Transformer decoder and providing an experimental evaluation on the Something-Something V2 dataset together with an ablation study. The proposed architecture is intended to facilitate adaptation to new domains with limited retraining. Unlike the previous publication, which focused on the architectural concept, the present paper reports the implementation details, training procedure, quantitative evaluation, and ablation experiments of the proposed framework.

1. RELATED WORKS

First, we consider the task of video caption generation – the automatic creation of descriptions of events in a video in natural language. The first methods of generating video descriptions were based on manual or automatic selection of key features (actors, actions, objects), after which templates were used to form sentences. This approach was limited in its expressiveness and scalability, as it did not allow for the combination of long-term events or the free generation of text without relying on predefined templates. In 2015, the publication of the Sequence to sequence – Video to Tex (S2VT) method [6], using a "decoder-encoder" architecture, became significant. S2VT is the first model that directly transforms video into a text description without using templates. Since the architecture is based on a bidirectional Long short-term memory (LSTM) model, S2VT learns simultaneously from visual representations and their descriptions. Sequence-to-sequence works well on videos of varying lengths, understands the sequence of actions over time, and generates descriptions similar to those that a human can create. However, the model is not adaptable to other domains and performs well on videos similar to those used for training [7]. Also, S2VT needs improvement to generate coherent descriptions of long-duration scenes.

Multimodal transformers. Transformers are multimodal models that combine images (or videos) and text in a shared spatial representation. Representative examples include Contrastive Language-Image Pre-training (CLIP) [8], Bootstrapping Language-Image Pre-training (BLIP) [9], and Frozen [10] models. These models are usually trained using self-supervised learning on [image, text] pairs, which allows them to achieve good results on classification and description generation tasks. For example, CLIP successfully transfers to new domains without retraining, making it the standard in many practical applications. However, de-spite their versatility, these models have a limited understanding of video as a sequence of related events. CLIP and BLIP mainly work with individual frames or short clips and do not consider the temporal context between frames. Frozen combines frozen CLIP-like visual encoders with a Generative Pre-trained Transformer (GPT) language model, but also lacks a distinct mechanism for handling video dynamics. All these models mainly operate on static scenes and are not adapted for in-depth analysis of actions, motion sequences, or cause-and-effect relationships in videos. This limitation has been one of the reasons why research on action description generation has gradually shifted its focus to models that specifically consider the temporal structure of video, such as VideoBERT, ClipBERT, ActionFormer, and others.

Domain adaptation (DA). One of the key challenges in generating video descriptions is domain specialization of models. Many modern architectures, including VideoBERT and CLIP-derived models, perform well on data like training data, but lose significant quality when transferred to new scenes, genres, or types of actions (e.g., from everyday videos to industrial ones). One effective solution to this problem is the use of adapter modules—small, parameterized blocks that are embedded in the transformer architecture and allow the model to adapt to new conditions without completely retraining the entire network.

Among the well-known adapters, we can highlight the Visual Language (VL) adapter [11] and Recurrent Adapter with Partial Alignment (READ) [12]. Despite architectural variations, these approaches share the prin-ciple of adapting a model by updating only a small subset of its parameters (for example, for VL, the adapter showed effectiveness when retraining only 3% of parameters).

This paper focuses on the DA-Ada adapter [5]. It consists of two parts:

1) Domain-Invariant Adapter (DIA) – for universal features.

2) Domain-Specific Adapter (DSA) – for local features specific to the domain. The authors of the approach focused specifically on the task of domain shifting, which is achieved by dividing features into those common to all domains and those specific to a particular do-main, and by combining adapters within the model by adding an attention mechanism. DA-Ada also allows for fine-tuning of the model, making it easier to scale to new domains.

VideoBERT. One of the key steps towards video-language representation was the VideoBERT model, which adapted the architecture of the bidirectional BERT transformer [13] to a multimodal context. Unlike previous approaches that relied on separate processing of video and text with subsequent merging, VideoBERT implements joint spatial learning of video and text tokens in a single model. The model is trained in self-supervised mode using token masking for both linguistic and visual sequences. This allows Video-BERT to learn generalized semantic representations without using labeled data, which can then be adapted to various tasks, such as video classification and description generation.

The authors of [4] also demonstrated the application of VideoBERT in the task of video description generation by integrating a transformer-based decoder. Within this architecture, VideoBERT serves as a source of semantic representations of videos, which are then passed to a decoder trained on a [video, description] pair. This approach enabled contextually aware generation of descriptions of actions in natural language based on knowledge gained during prior self-training. However, despite its demon-strated advantages, this architecture is still limited to the training domain: the model performs best on similar types of videos but has difficulty generalizing to new scenes.

2. PROBLEM STATEMENT

The task of contextual video understanding and generation of semantically correct action descriptions remains challenging due to a combination of several fundamental problems: the high spatio-temporal complexity of video data, the dependence of action semantics on context, and the significant domain shift between different video sources [14]. Recent video domain adaptation studies have addressed this problem through feature disentanglement, separating domain-specific information from domain-invariant representations [15]. Visually similar scenes can correspond to different actions depending on the environment and sequence of frames, while the same action can manifest itself differently in different domains (domestic scenes, industrial environments, educational videos, etc.).

Existing approaches to video classification and video description are usually either trained end-to-end on large amounts of data or require complete retraining when moving to a new domain [16]. Such methods have high computational costs, limited scalability, and poor generalization under changing data distributions. In addition, models based solely on visual features are often unable to adequately reflect the context of the action at the semantic level, which is critical for text description generation tasks.

A separate problem is the effective use of pre-trained models. Although video-language encoders such as VideoBERT can model long-term temporal dependencies and the shared space of video and text, their full retraining is resource-intensive and not always justified when adapting to new domains [17].

This creates a need for architectural solutions that allow the knowledge of previously trained models to be preserved while ensuring flexible adaptation to new conditions.

The aim of this work is to formulate and experimentally investigate a context-aware video captioning architecture designed to support subsequent adaptation under domain shift. The proposed approach is aimed at building a modular architecture that combines the frozen VideoBERT video-language encoder, the domain-oriented adaptation module DA-Ada, and an autoregressive transformer decoder for generating descriptions.

To achieve this goal, the following main tasks are formulated in the work:

1. Investigate the possibility of using frozen VideoBERT as a universal video context encoder without full retraining.

2. Implement a dual-branch adaptation mechanism based on the domain-invariant and domain-specific pathways of the DA-Ada architecture.

3. Evaluate the contribution of the adaptation module to context-aware video caption generation through component-wise ablation experiments.

4. To analyze the quality and stability of text description generation compared to baseline architectures without domain adaptation.

The approach assumes that effective separation of universal and domain-specific features allows reducing the negative impact of domain shift, and combining a pre-trained video-language encoder with lightweight adapters provides a balance between generalization ability, flexibility, and computational efficiency. As a result, the task of contextual video understanding is viewed as a problem of joint modeling of video context, action semantics, and domain invariance within a single experimental architecture.

3. PROPOSED METHOD

The architecture of the model under study is designed to solve the problem of context-aware video caption generation with the ability to support different domains. Video is input into the system, convert-ed into visual tokens, and sequentially processed by the VideoBERT encoder, DA-Ada adapter, feature fusion layer, and decoding transformer. The following subsections describe each module in detail [18].

Something-Something V2. The Something-Something V2 dataset was selected for training the model, which contains 220,847 short videos with short sequential related actions (e.g., cooking a meal). The dataset is divided into three sets containing 168,913 (training), 24,777 (validation) and 27,157 (test) videos. In total, the videos contain 174 unique action types in the format "Doing [something] with [something]," for example, "Putting [something] onto [something]" [19].

Preliminary video processing. First, the video must be converted into a fixed-length sequence of visual tokens. For each input video, 16 frames are uniformly sampled across the entire video duration. Videos containing fewer than 16 available frames are supplemented with zero frames with a corresponding temporal mask. Each frame $f_i \in R^{H\times W\times 3}$ is processed by a pre-trained visual model to obtain a feature vector. In this study, we chose the ResNet-50 convolutional network model [20], pre-trained on the ImageNet dataset. The feature vector has a dimension of 2048 and is obtained at the output of the last convolutional layer of the model.

Each feature vector $z_i \in R^{2048}$ is projected into a space with dimension $\mathrm{D} = 768$ using a linear layer:

$$\mathrm{x_i} = \mathrm{W_{proj}} \cdot \mathrm{z_i} + \mathrm{b_{proj}},\ \mathrm{x_i} \in \mathrm{R}^{768}, \quad (1)$$

where $W_{\text{proj}} \in R^{768\times 2048}$ is a weight matrix that performs a linear transformation from the ResNet output space to the transformer space, $b_{\text{proj}} \in R^{768}$ is a bias vector added after multiplication by the weights.

The resulting sequence of vectors $X = \{x_1, x_2, \dots, x_T\}$ forms the basis for further tokenization. Each token $\mathrm{x_i}$ is assigned a positional vector representation $\mathrm{p_i}$,, which considers the order of the frame in the sequence. In addition, a temporal mask $\mathrm{m} \in \{0,1\}^{\mathrm{T}}$ is introduced, which marks real frames and zero frames. This mask is used in self-attention modules to avoid processing non-informative positions.

The final representation of each frame is:

$$x_i' = x_i + p_i, \quad (2)$$

where $x_i \in R^D$ is a token obtained after projection from ResNet, $p_i \in R^D$ is a positional transformation (embedding) that considers the frame order, and x_i' is a token with positional information. The entire sequence is input to VideoBERT as $\mathrm{X}' = \{\mathrm{x}_1', \mathrm{x}_2', \dots, \mathrm{x}_\mathrm{T}'\}$ together with the temporal binary mask m.

This stage provides a structured representation of the video fragment while preserving temporal ordering information.

VideoBERT encoder. The VideoBERT module serves as the main encoder in the implemented architecture. Its function is to process a sequence of visual tokens and form contextualized representations that preserve both local and global temporal-spatial dependencies in the video. VideoBERT is based on the BERT transformer architecture, adapted for processing video clips.

The implementation uses a 12-layer transformer with a multi-head self-attention mechanism and two-layer forward propagation blocks.

Each layer includes the following components:

– a self-attention mechanism for modeling dependencies between video tokens, including long-range connections between frames;

– residual learning and normalization layers to ensure gradient stability and facilitate optimization;

– a feed-forward network for transforming each token separately.

A sequence of visual tokens $X' = \{x'_1, x'_2, \dots, x'_T\}$, and a temporal mask $m \in \{0,1\}^T$ are input to VideoBERT.

The model evaluates the contextualized representation of each token:

$$H = \text{VideoBERT}(X', m) = \{h_1, h_2, \dots, h_T\}, \quad h_i \in R^D, \tag{3}$$

where $H = \{h_1, h_2, \dots, h_T\}$ is the output sequence of feature vectors containing the spatiotemporal context of all tokens after processing in VideoBERT, and D is the dimension of the space.

These vectors are passed to the DA-Ada adaptation module for further contextual feature separation.

DA-Ada Adapter. To improve the model's ability to generalize to new domains, the DA-Ada (Domain-Aware Adapter) module is integrated into the architecture (see Fig.1). The original DA-Ada approach was developed for domain-adaptive object detection and employs separate domain-invariant and domain-specific adaptation components. In the present work, its dual-branch design is transferred to a video caption generation architecture operating on contextualized video representations.

DA-Ada is an adaptation block designed for selective modeling of both universal and domain-specific features in a transformer environment. DA-Ada is placed after the main VideoBERT encoder and modifies its output representations before they are input to the decoder. The module is built from the dual-branch design of DA-Ada and is adapted here as a parameter-efficient representation refinement stage between the video encoder and the caption decoder.

DA-Ada consists of two subcomponents: DIA and DSA. DIA is implemented as a bottle-neck MLP (Multi-Layer Perceptron) and preserves the structural role of the Domain-Invariant Adapter from the original DA-Ada formulation. In the present implementation, it constitutes an independently parameterized bottleneck transformation of the shared video representation; explicit domain-invariance supervision is not applied. This block can be represented as follows:

$$\text{DIA}(h_i) = W_2^{\text{inv}} \cdot \sigma\left(W_1^{\text{inv}} \cdot h_i + b_1^{\text{inv}}\right) + b_2^{\text{inv}}, \quad h_i \in R^D, \tag{4}$$

where $h_i \in R^D$ is the input feature vector for the i token from the VideoBERT output; $W_1^{\text{inv}} \in R^{d \times D}$ is the weight matrix of the first layer of the adapter, which compresses the vector into a space of lower dimension d; $b_1^{\text{inv}} \in R^d$ is the bias vector of the first layer; $W_2^{\text{inv}} \in R^{D \times d}$ – weight matrix of the second layer, which returns the vector to the output space D; $b_2^{\text{inv}} \in R^D$ – bias vector of the second layer and σ – ReLU activation function.

The DSA branch follows the domain-specific pathway of the original DA-Ada concept and applies a second independently parameterized transformation to the encoder representation.

In the current single-benchmark setting, domain-specific specialization is not explicitly supervised. It has the same architecture as DIA, but with separate parameters:

$$\text{DSA}(h_i) = W_2^{\text{spec}} \cdot \sigma\left(W_1^{\text{spec}} \cdot h_i + b_1^{\text{spec}}\right) + b_2^{\text{spec}}, \tag{5}$$

where $h_i \in R^D$ – is the input feature vector for the i token from the VideoBERT output; $W_1^{\text{inv}} \in R^{d \times D}$ is the weight matrix of the first layer of the adapter, which compresses the vector into a lower-dimensional space d; $b_1^{\text{inv}} \in R^d$ is the bias vector of the first layer; $W_2^{\text{inv}} \in R^{D \times d}$ is the weight matrix of the second layer, which returns the vector to the output space D; $b_2^{\text{inv}} \in R^D$ is the bias vector of the second layer, and σ is the ReLU activation function.

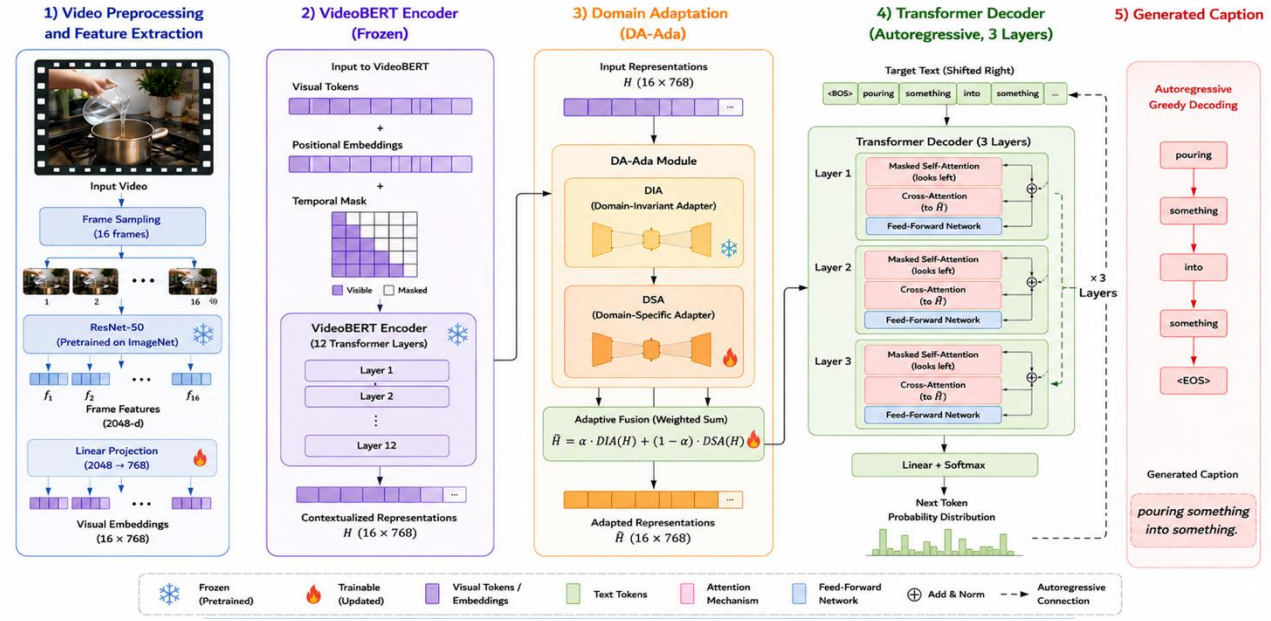


Fig. 1. **Proposed method algorithm**
Source: **compiled by the authors**

Next, the obtained features are passed to the fusion layer to form the final context for the decoder.

Feature fusion layer. Both adapters are combined at the level of each token using a trainable fusion parameter:

$$\alpha = \sigma(\beta),$$
$$\widetilde{h_i} = \alpha \cdot \text{DIA}(h_i) + (1 - \alpha) \cdot \text{DSA}(h_i), \quad (6)$$

where $\widetilde{h_i} \in R^D$ is the fused adapted feature vector; β is a trainable scalar parameter; $\alpha \in [0,1]$ is an effective fusion coefficient; $\text{DIA}(h_i) \in R^D$ is the output of the DIA adapter; $\text{DSA}(h_i) \in R^D$ is the output of the DSA adapter.

This stage is an internal fusion of adapter features and forms an adapted sequence of vectors $\widetilde{H} = \{\tilde{h}_1 \dots, \tilde{h}_T\}$, which is passed to the next stage of the model.

Transformer decoder. The final stage of the investigated architecture is the generation of a text description of the action taking place in the video. For this purpose, a transformer autoregressive decoder is used, which forms a sequence of natural language tokens based on the context obtained from adapted visual features.

Embeddings are added to the sequence $\widetilde{H} = \tilde{h}_1, \dots, \tilde{h}_T$ to preserve temporal ordering, as well as a temporal mask $m \in \{0,1\}^T$, which cuts off zero non-informative tokens. This processed sequence is passed to the cross-attention module of each decoder layer as keys and values.

The decoder consists of 3 transformer layers, each of which includes.

1. A masked self-attention mechanism for auto-completion of partially formed text;
2. A cross-attention layer to the visual context $\widetilde{H}$.
3. Two linear layers of forward propagation.
4. Normalization and dropout for stability.

The decoder operates in an autoregressive decoding mode, where each token y_t is generated sequentially, considering all previous tokens $y_{<t}$ and the context.

The task of generating a description is formulated as maximizing the probability of the text sequence $= \{y_1, y_2, \dots, y_N\}$ given the adapted visual context $\widetilde{H}$ [21]:

$$P\left(Y \mid \widetilde{H}\right) = \prod_{t=1}^{N} P\left(y_t \mid y_{<t}, \widetilde{H}\right), \quad (7)$$

where $P\left(Y \mid \widetilde{H}\right)$ is the probability of a complete sentence given the context; $P\left(y_t \mid y_{<t}, \widetilde{H}\right)$ is the conditional probability of a token at a step.

The decoder is implemented as a linguistic transformer with a trainable vocabulary, initialized randomly and optimized using a cross-entropy loss function.

At the model application stage, generation begins with the token <BOS> (begin-of-sentence) and ends when the model predicts the token <EOS> (end-of-sentence) or reaches the maximum length limit.

The decoder output is a natural language phrase that briefly describes the action in the video, for example: *"Pouring something into something"*.

The complete end-to-end processing pipeline, including feature extraction, spatiotemporal context modeling via VideoBERT, and the proposed dual-branch DA-Ada macro-architecture, is structurally mapped and illustrated in Fig. 1: the input video is sampled into 16 frames → each frame is encoded by a ResNet-50 model pretrained on ImageNet → features are projected to 768-d embeddings and processed by a frozen VideoBERT encoder (12 layers) with temporal masking and positional information → the DA-Ada module (DIA + DSA) adapts the representations and fuses them with a learnable coefficient α → the adapted representations are fed to a 3-layer transformer decoder, which generates the text description autoregressively, using generalized category labels (e.g., "something") in accordance with the Something-Something V2 dataset specification.

5. TRAINING

Each video is represented by the sequence of 16 uniformly sampled frames described in Section 3. Each frame is scaled to a size of 224×224 pixels and processed by a pre-trained ResNet-50 visual model, which extracts features from the deep level of the network. These features are projected into a 768-dimensional latent space, yielding visual embeddings that are augmented with positional information and a temporal mask to filter out padded elements.

The resulting sequence is fed into a frozen VideoBERT encoder (comprising 12 Transformer layers) to compute contextualized, spatiotemporal frame representations. These representations are subsequently mapped through the proposed DA-Ada adaptation module. Within this module, the two independently parameterized adaptation branches, DIA and DSA, process the contextualized representations in parallel following the dual-branch structure of the original DA-Ada architecture.

To mitigate linguistic mode collapse and ensure structural reliance on the visual stream, the text generation module is implemented as a lightweight three-layer Transformer decoder. During the training phase under the Teacher Forcing regime, an extreme 50% stochastic token masking (MLM) [22] is applied to the input text sequence to prevent the decoder from over-relying on linguistic n-gram statistics. The network is optimized via the cross-entropy loss function, comparing the autoregressively generated hypotheses against the ground-truth action annotations.

During training, the visual backbone and the main encoder remain fixed, whereas the visual projection layer, both adapter branches, the fusion parameter, and the Transformer decoder are optimized using the captioning cross-entropy loss. The fusion coefficient is initialized through a scalar gating parameter and optimized jointly with the remaining trainable modules.

Optimization is carried out using the AdamW optimizer [23] with a calibrated initial learning rate of 2×10^{-5}, a weight decay of 0.01, and a cosine learning rate scheduler including a warm-up period over the first 5 epochs. The training is conducted with a batch size of 4 over 10 epochs. To prevent overfitting under aggressive linguistic constraints, standard regularization is reinforced, including a dropout probability of 0.15 within the decoder layers and 0.15 within the adaptation modules. Frame-level augmentations, such as random horizontal flips and color jittering, are

applied to enhance visual robustness. Early stopping [24] is executed based on the convergence of the validation loss paired with the stabilization of the Mean ROUGE-L metric [25].

Prior to batch formation, all visual inputs undergo standard channel-wise normalization corresponding to the backbone pre-training statistics. The data pipeline utilizes dynamic padding paired with complementary temporal and causal attention masks to assemble structurally synchronized tensors within each mini-batch. This structural standardization isolates the cross-attention mechanisms from padding-induced artifacts, ensures the architecture's invariance to temporal scale variations across heterogeneous video inputs, and enforces numeric stability during gradient propagation.

6. EXPERIMENTS

Dataset and hardware setup. The empirical evaluation of the implemented architecture was conducted on the challenging Something-Something V2 (SSv2) benchmark. The dataset was selected because it emphasizes fine-grained human–object interactions rather than static object recognition. Its large diversity of manipulated objects, action categories and recording conditions makes it an appropriate benchmark for evaluating context-aware video representation learning, although it does not constitute a formal cross-domain evaluation protocol. For the computational experiments, a subset of 10,000 video sequences was extracted to reduce computational requirements while retaining a sufficiently diverse set of action samples. The sub-dataset was partitioned into mutually exclusive training set (8,000 samples), a validation (1,000 samples) and testing set (1,000 samples) to monitor generalization capacity and perform evaluation. The validation set was used for model selection and monitoring during training, whereas the test set was reserved exclusively for the final evaluation. A fixed random seed of 42 was used for dataset partitioning and other stochastic operations to ensure reproducibility. Since the current experimental validation is limited to the Something-Something V2 dataset, the presented results demonstrate the effectiveness of the proposed architecture within a single benchmark. Cross-domain transfer remains a subject of future research.

All experiments were implemented within the PyTorch framework [26] and executed in a cloud computing environment. The hardware configuration comprised an Intel Xeon CPU @ 2.20 GHz (2 vCPUs, 12 GB RAM) paired with an NVIDIA Tesla T4 GPU (16 GB VRAM). This setup accommodated the unified training of the linear projection, the DA-Ada macro-module, and the 3-layer autoregressive Transformer decoder under the specified batch constraints.

Baselines and Ablation Framework. To evaluate the contribution of individual architectural components, we conducted an ablation study to isolate the performance impact of each conceptual component.

1. Baseline Decoder Mode (Standard Teacher Forcing [27]): A baseline configuration mapping raw visual features into a standard sequence-to-sequence decoder without additional decoder-input regularization. This setup provides a reference for evaluating the effect of the proposed training-time regularization strategy.

2. Strict Text Filtering Variant (Truncation Mode): An intermediate baseline where the linguistic decoder's output distribution is conditioned post-factum via rule-based token suppression during inference, preventing adjacent word-level repetitions.

3. Ablated Spatio-temporal Context (No-VideoBERT): A structural ablation variant where ResNet-50 frame embeddings are projected directly into the adaptation module, bypassing the 12-layer frozen VideoBERT encoder. This configuration isolates the contribution of spatiotemporal contextualization provided by the VideoBERT encoder.

4. No-Adapter Variant (VideoBERT + Decoder): A structural ablation configuration in which the contextualized representations produced by the frozen VideoBERT encoder are passed directly to the Transformer decoder, bypassing both DIA and DSA branches. This configuration serves as the reference point for isolating the contribution of the dual-branch adaptation module.

5. DIA-Only Variant (VideoBERT + DIA + Decoder): A single-branch adaptation configuration in which only the DIA branch is applied to the contextualized VideoBERT

representations, while the DSA branch is removed. This variant evaluates the contribution of the independently parameterized DIA branch relative to both the no-adapter and complete dual-branch configurations.

6. DSA-Only Variant (VideoBERT + DSA + Decoder): A complementary single-branch adaptation configuration in which only the DSA branch is retained between the VideoBERT encoder and the Transformer decoder. This experiment evaluates the contribution of the independently parameterized DSA branch under otherwise identical training conditions.

7. Dual-Branch DA-Ada without MLM: The complete adaptation configuration incorporating both DIA and DSA branches and their learnable fusion, while MLM is not applied.

8. Dual-Branch DA-Ada with 25 % MLM: The complete adaptation configuration incorporating both DIA and DSA branches and their learnable fusion, adding 25 % MLM.

Proposed Method (Dual-Branch DA-Ada + 50 % MLM): The complete proposed configuration combining the frozen VideoBERT encoder, both adaptation branches with learnable fusion, and 50 % MLM.

The quantitative evaluation results summarized in Table 1 provide a component-wise assessment of the proposed architecture. The Standard Teacher Forcing baseline achieves a BLEU-4 score of 0.1520 but shows lower ROUGE-L and METEOR values than the proposed configuration, suggesting that exact n-gram matching alone does not reflect the overall quality of the generated descriptions. The Strict Text Filtering variant, which suppresses adjacent token repetitions during inference, decreases BLEU-4 to 0.1051 and ROUGE-L to 0.2609, indicating that rule-based suppression of local repetitions does not provide a sufficient substitute for training-time regularization. Removing the VideoBERT encoder also substantially reduces performance, with the No-VideoBERT configuration achieving a Mean ROUGE-L of 0.2140. This result supports the contribution of spatiotemporal contextualization to the representation of motion-centric video content.

Table 1. **Quantitative evaluation of the proposed model against ablation baselines on the Something-Something V2 dataset**

Evaluation configuration	Mean BLEU-4	Mean ROUGE-L	Mean METEOR
Mode Collapse Baseline (Standard TF)	0.1520	0.3163	0.1820
Strict Text Filtering Variant (Truncation)	0.1051	0.2609	0.0540
Ablated Spatiotemporal Context	0.0710	0.2140	0.1120
VideoBERT + Transformer Decoder	0.1321	0.3083	0.2050
VideoBERT + DIA + Transformer Decoder	0.1372	0.3222	0.2186
VideoBERT + DSA + Transformer Decoder	0.1390	0.3265	0.2263
VideoBERT + DA-Ada + Transformer Decoder (0% MLM)	0.1438	0.3370	0.2441
VideoBERT + DA-Ada + Transformer Decoder (25 % MLM)	0.1463	0.3433	0.2589
VideoBERT + Transformer Decoder	0.1321	0.3083	0.2050
Proposed Method (VideoBERT + DA-Ada + Transformer Decoder + 50% MLM)	**0.1469**	**0.3474**	**0.2720**

Source: **compiled by the authors**

The component-wise ablation further isolates the contribution of the adaptation module. The VideoBERT + Decoder configuration without adaptation and MLM achieves BLEU-4, ROUGE-L, and METEOR scores of 0.1321, 0.3083, and 0.2050, respectively. Introducing only the DIA branch increases the corresponding scores to 0.1372, 0.3222, and 0.2186, whereas the DSA-only

configuration achieves 0.1390, 0.3265, and 0.2263. When both branches are jointly employed, the scores further increase to 0.1438, 0.3370, and 0.2441. Although the current implementation does not explicitly supervise domain-invariant and domain-specific specialization of the individual branches, these results indicate that the dual-branch adaptation structure provides a greater improvement than either independently parameterized branch used alone.

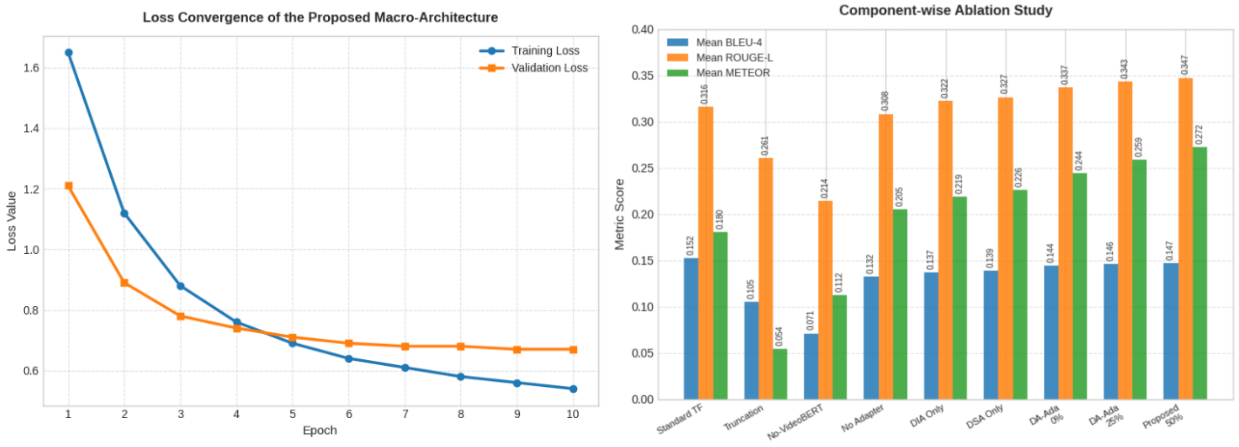
The effect of MLM was evaluated separately while keeping the VideoBERT encoder and the complete dual-branch adaptation module unchanged. Without MLM, the model achieves a Mean ROUGE-L of 0.3370 and a Mean METEOR of 0.2441. Increasing the masking ratio to 25 % improves these values to 0.3433 and 0.2589, respectively, while the 50 % configuration reaches 0.3474 and 0.2720. BLEU-4 changes more moderately, from 0.1438 without masking to 0.1463 at 25 % and 0.1469 at 50 %. These results suggest that MLM primarily improves sequence-level and semantic agreement while producing a comparatively smaller effect on exact n-gram matching.

Overall, the complete configuration combining VideoBERT, the dual-branch adaptation module, and 50 % MLM achieves the strongest overall balance across the evaluated metrics, with BLEU-4 of 0.1469, Mean ROUGE-L of 0.3474, and Mean METEOR of 0.2720. Importantly, the ablation results separate the contribution of the adaptation module from that of decoder-input masking: the former improves performance relative to the no-adapter configuration under identical masking conditions, while the latter provides an additional improvement when applied to the complete adaptation architecture.

The quantitative dynamics of the proposed framework are further illustrated in Fig. 2.

Fig. 2(a) presents the training and validation loss curves over 10 training epochs. The decrease in both losses indicates stable optimization under the selected training configuration.

Fig. 2(b) summarizes the component-wise ablation results. While the Standard Teacher Forcing baseline achieves the highest BLEU-4 score, its lower sequence-level and semantic metrics indicate that exact n-gram matching alone does not correspond to the strongest overall caption quality. The No-VideoBERT experiment demonstrates the importance of spatiotemporal contextualization, whereas the DIA-only, DSA-only, and combined dual-branch configurations progressively characterize the contribution of the adaptation module. Finally, the comparison of 0 %, 25 %, and 50 % decoder-input masking shows an improvement in ROUGE-L and METEOR as the masking ratio increases. The complete configuration with both adaptation branches and 50 % token masking therefore provides the strongest overall balance among the evaluated metrics, achieving a Mean ROUGE-L of 0.3474 and a Mean METEOR of 0.2720.



a *b*

Fig. 2. **Quantitative dynamics of the proposed framework**
Source: **compiled by the authors**

7. RESULTS AND DISCUSSIONS

Qualitative Case Studies and Visual Grounding. To evaluate the practical interpretability of the proposed DA-Ada macro-architecture, a qualitative analysis was performed on distinct

verification samples. As illustrated in Fig. 3, the generated descriptions show close agreement with the reference actions in several representative cases. For action classes involving localized deformation, such as "*Tearing something into two pieces"* (ID 179956 and ID 42068), the model shows close syntactic and semantic agreement with the reference descriptions regardless of structural background transitions or object variations.

Furthermore, the model successfully resolves abstract cognitive tasks, as demonstrated in sample ID 78483, where it correctly generates the long-tail *description "pretending to open something without actually opening it"*. This example suggests that the dual-branch adapter combined with VideoBERT temporal masks goes beyond static frame indexing, accurately extracting intention and micro-manual trajectories. Semantic variation is also observed in sample ID 130352, where the fine-grained predicate "*poking"* is functionally replaced by the broader term "*ushing"*, preserving the global physical output condition ('so that it slightly moves').

Another example of temporal interaction modeling is shown in sample ID 33756. While the ground-truth annotation frames the action as "*Rolling something on a flat surface"*, the proposed architecture generates a more detailed causal description*: "letting something roll along a flat surface"*.

Fig. 3. **Examples of successful text generation on the SSv2 validation set**
Source: **compiled by the authors**

The model captures the kinetic transition from manual acceleration to autonomous motion ("*etting something roll"*). Furthermore, this case provides an empirical validation of the calibrated evaluation pipeline: while the strict token-matching BLEU-4 metric remains conservative (0.2627), the METEOR score reaches 0.9221. This result suggests that the framework achieves high semantic similarity, effectively aligning spatial prepositions ("*on"* vs. "*along"*) based on visual context rather than rigid text-string replication.

Failure Mode and Error Propagation Analysis. Despite competitive global metrics, explicit edge-case anomalies were isolated to establish the boundaries of the architecture's generalizability. In sample ID 44850, the ground truth action "*Turning something upside down"* is incorrectly

interpreted as "*pushing something from left to right"*. This systemic failure is attributed to severe motion blur artifacts occurring within frames 3 and 4 during rapid hand acceleration. The ResNet-50 visual backbone fails to extract discrete object orientation vectors under high spatial variance, causing the temporal encoder to misinterpret the global kinetic energy as a horizontal slide.

Similarly, in sample ID 175172, the macro-module prioritizes a secondary visual cue, mistaking a lifting action for rolling motion ("*letting something roll along a flat surface"*). This error indicates that under severe scale imbalances between interacting objects (e.g., a small pen on a vast table surface), the global spatial pooling layer can suppress local motion tokens, leading to premature text generation conditioned on early frame fractions.

CONCLUSIONS AND PROSPECTS OF FURTHER RESEARCH

This paper introduces a neural network macro-architecture designed to support domain adaptation in automated video description for business process analysis. By unifying a pre-trained VideoBERT encoder with the proposed dual-branch Domain-Aware Adaptation module (DA-Ada) and a regularized Transformer decoder, the framework provides two independently parameterized adaptation pathways for refining contextualized video representations. The proposed architecture provides a parameter-efficient framework intended to support future adaptation to heterogeneous application domains. Experimental validation on the Something-Something V2 benchmark confirms its effectiveness for context-aware video caption generation, while cross-domain transfer evaluation remains future work (Fig. 4).

Fig. 4. **Failure mode analysis and typical error profiles**
Source: **compiled by the authors**

Empirical evaluation on the Something-Something V2 benchmark demonstrates that the DA-Ada modules, supported by a 50 % stochastic Masked Language Modeling (MLM) objective, reduces repetitive generation and improves sequence-level and semantic agreement. The final model achieves a Mean ROUGE-L of 0.3474 and a Mean METEOR score of 0.2720, indicating improved sequence-level and semantic agreement with the reference descriptions. Component-wise ablation further demonstrates that the combined dual-branch adaptation configuration outperforms both the no-adapter and individual DIA- and DSA-branch variants, while increasing the decoder-input masking ratio from 0% to 50% provides additional improvements in ROUGE-L and METEOR.

Limitations. The visual processing layer remains sensitive to high-speed kinetic actions that induce motion blur, which can cause the temporal encoder to misinterpret rapid axial object rotations. Furthermore, substantial spatial scale imbalances between interacting hands and small target objects can lead to local feature suppression and minor predicate errors.

Future Research Directions. Future research will focus on several core areas. First, we aim to enhance temporal feature extraction to increase the architecture's resilience against motion blur and high-frequency movement artifacts. Second, we plan to refine the adaptive token fusion by transitioning from global scalar parameters to granular, dynamic token-level gating mechanisms. Finally, explicit cross-domain evaluation using separate source and target domains and domain-supervised adaptation objectives will be conducted to assess the domain-invariant and domain-specific specialization of the DIA and DSA branches.

ACKNOWLEDGEMENTS

The results of this research were obtained under the international research project INITIATE under the grant No.101136775-HORIZON-WIDERA-2023-ACCESS-03.

Conflict of interests: Authors declare that they do not have conflict of interests, in particular financial, personal, copyright or any of a different nature, which could have influenced the research, as well as the results published in this articles

Financing: Research was carried out without financial support

Accessibility data: All data available in digital or graphic form in the main text manuscript

Using artificial intelligence tools: During preparation manuscript used tool ChatGPT, GPT-5.6 Sol, OpenAI of artificial intelligence (AI) for grammatical and stylistic text checks. All fragments processed by him were checked and edited by the author. AI was not used to generate text, formulas, and tables, formation scientific provisions, receipt results or formulation conclusions. Scientific the provisions, results and conclusions are my own, intellectual by the contribution of the authors

REFERENCES

1. Novichonok, M. S. & Normatova, T. V. “Selection of machine learning methods for contextual video classification tasks in business analysis”. *Zbirnyk Naukovykh Prats Studentiv, Mahistrantiv ta vykladachiv “Naukovyi Poshuk Molodykh Doslidnykiv”*. 2025; 18: 124–127. – Available from: https://eprints.zu.edu.ua/44884/1/1.pdf – [Accessed: May, 2026] .

2. Xu, Y., Cao, H., Chen, Z., Li, X., Xie, L. & Yang, J. “Video unsupervised domain adaptation with deep learning: A comprehensive survey”. *ACM Computing Surveys*. 2024. DOI: https://doi.org/10.1145/3679010.

3. Chen, M., Wang, Z., Zhao, S., Su, H. & Chang, S. F. “Temporal attentive alignment for video domain adaptation”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019. p. 12430–12439. DOI: https://doi.org/10.48550/arXiv.1905.10861.

4. Sun, C., Myers, A., Vondrick, C., Murphy, K. & Schmid, C. “VideoBERT: A joint model for video and language representation learning”. *arXiv*. 2019. DOI: https://doi.org/10.48550/arxiv.1904.01766.

5. Li, H., Zhang, R., Yao, H., et al. “DA-Ada: Learning domain-aware adapter for domain adaptive object detection”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arxiv.2410.09004.

6. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T. & Saenko, K. “Sequence to sequence – video to text”. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.2015. p. 4534–4542. DOI: https://doi.org/10.1109/ICCV.2015.515.

7. Yao, L., Torabi, A., Cho, K., et al. “Describing videos by exploiting temporal structure”. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015. p. 4507–4515. DOI: https://doi.org/10.48550/arXiv.1502.08029.

8. Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S. M. A., Vinyals, O. & Hill, F. “Multimodal few-shot learning with frozen language models”. In *Advances in Neural Information Processing Systems*. 2021; 34: 200–212. *NeurIPS Proceedings*. DOI: https://doi.org/10.48550/arXiv.2106.1388.

9. Li, J., Li, D., Xiong, C. & Hoi, S. “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation”. In *Proceedings of the International Conference on Machine Learning (ICML)*. 2022. p. 12888–12900. DOI: https://doi.org/10.48550/arXiv.2201.12086.

10. Bain, M., Nagrani, A., Varol, G. & Zisserman, A. “Frozen in time: A joint video and image encoder for end-to-end retrieval”. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021. p. 1728–1738. DOI: https://doi.org/10.48550/arXiv.2104.00650.

11. Sung, Y. L., Cho, J. & Bansal, M. “VL-Adapter: Parameter-efficient transfer learning for vision-and-language tasks”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022. p. 5227–5237. DOI: https://doi.org/10.48550/arXiv.2112.06825.

12. Anwar, M. U., Raza, A. & Goldberg, J. C. C. “READ: Recurrent adapter with partial alignment for parameter-efficient multimodal convergence”. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. p. 4810–4820. DOI: https://doi.org/10.48550/arXiv.2312.06950.

13. Zhang, S., Zhang, J., An, N. & Zhou, H. “A novel approach to unsupervised video domain adaptation: A disentanglement perspective”. In *International Conference on Learning Representations (ICLR). OpenReview*. 2024. – Available from: https://openreview.net/forum?id=Rp4PA0ez0m – [Accessed: May, 2026]

14. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP*). *Association for Computational Linguistics*. 2019. p. 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423.

15. Choi, J., Sharma, G., Chandraker, M. & Huang, J. B. “Unsupervised and semi-supervised domain adaptation for action recognition from drones”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021. p. 463–472. DOI: 10.1109/WACV45572.2020.9093511.

16. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M. & Schmid, C. “ViViT: A video vision transformer”. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021.p. 6816–6826. DOI: https://doi.org/10.48550/arXiv.2103.15691.

17. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A. & Wang, M. “Parameter-efficient transfer learning for NLP”. In *Proceedings of the 36th International Conference on Machine Learning. PMLR*. 2019; 97: 2790–2799. DOI: 10.48550/arXiv.1902.00751.

18. Novichonok, M. S. & Mashtalir, S. V. “Video context classification by VideoBERT and DA-Ada adapters”. *Informatics. Culture. Technology*, 2025. 2: 135–141. DOI: https://doi.org/10.15276/ict.02.2025.19.

19. Goyal, R., Ebrahimi Kahou, S., Michalski, V., et al. “The 'Something Something' video database for learning and evaluating visual common sense”. *arXiv*. 2017. p. 5842–5850. DOI: https://doi.org/10.48550/arXiv.1706.04261.

20. He, K., Zhang, X., Ren, S. & Sun, J. “Deep residual learning for image recognition”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR*) 2026. p. 770–778. DOI: https://doi.org/10.48550/arXiv.1512.03385.

21. Vaswani, A., Shazeer, N., Parmar, N., et al. “Attention is all you need”. In *Advances in Neural Information Processing Systems*. 2017. p. 5998–6008. DOI: https://doi.org/10.48550/arXiv.1706.03762.

22. Reddy, A., Paul, W., Rivera, C., Shah, K., de Melo, C. M. & Chellappa, R. “Unsupervised video domain adaptation with masked pre-training and collaborative self-training”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024. DOI: https://doi.org/10.48550/arxiv.2312.02914.

23. Loshchilov, I. & Hutter, F. “Decoupled weight decay regularization”. In *International Conference on Learning Representations (ICLR). OpenReview.* 2019. DOI: https://doi.org/10.48550/arXiv.1711.05101.

24. Prechelt, L. “Early stopping – But when?” In *G. Montavon, G. B. Orr, & K.-R. Müller (Eds.), Neural networks: Tricks of the trade (2nd ed*). 2012. p. 53–67. DOI: https://doi.org/10.1007/978-3-642-35289-8_5.

25. Lin, C. Y. “ROUGE: A package for automatic evaluation of summaries”. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*. 2004. p. 74–81. *Association for Computational Linguistics*. – Available from: https://aclanthology.org/W04-1013/ – [Accessed: May, 2026].

26. Paszke, A., Gross, S., Massa, F., et al. “PyTorch: An imperative style, high-performance deep learning library”. *Advances in Neural Information Processing Systems*, 2019; 32: 8024–8035. DOI: https://doi.org/10.48550/arXiv.1912.01703.

27. Williams, R. J., & Zipser, D. “A learning algorithm for continually running fully recurrent neural networks”. *Neural Computation*. 1989; 1 (2): 270–280. DOI: https://doi.org/10.1162/neco.1989.1.2.270.

28. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. “BLEU: A method for automatic evaluation of machine translation”. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics* 2002. p. 311–318. DOI: 10.3115/1073083.1073135.

29. Banerjee, S. & Lavie, A. “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments”. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation*. (2005). – Available from: https://aclanthology.org/W05-0909.pdf. – [Accessed: May, 2026].

30. Miller, G. A. “WordNet: A lexical database for English. Communications of the ACM”. 1995; 38 (1): 39–41. DOI: https://doi.org/10.1145/219717.219748.

Received 27.07.2026
Received after revision 25.08.2026
Accepted 04.09.2026

DOI: https://doi.org/10.15276/ict.03.2026.19
УДК 004:83

Контекстно-орієнтована генерація описів відео з використанням VideoBERT та доменно-орієнтовних адаптерів

Новічонок Марія Сергіївна[1)]
ORCID: https://orcid.org/0009-0007-1787-0546; mariia.novichonok@nure.ua
Машталір Сергій Володимирович[1)]
ORCID: http://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
[1)] Харківський Національний Університет Радіоелектроніки, проспект Науки 14. Харків, 61166, Україна

АНОТАЦІЯ

Актуальність. Автоматизований опис відео є актуальним завданням для реінжинірингу робочих процесів та бізнес-аналізу, проте його ефективність у реальних умовах залежить від доменного зсуву між навчальними наборами даних і середовищами практичного застосування. **Мета.** Метою дослідження є розроблення архітектури контекстно-орієнтованої генерації описів відео, що поєднує попередньо навчені відеопредставлення з легковаговими механізмами адаптації. **Методи.** Запропонований підхід інтегрує заморожений 12-шаровий енкодер VideoBERT із двогілковим адаптаційним модулем доменно-орієнтовних адаптерів та тришаровим авторегресійним Трансформер-декодером. Візуальні ознаки виділяються за допомогою моделі ResNet-50, попередньо навченої на наборі даних ImageNet. Гілки доменно-інваріантних адаптерів та доменно-специфічних адаптерів паралельно обробляють контекстуалізовані представлення та об'єднуються за допомогою навчуваного скалярного коефіцієнта злиття. Під час навчання застосовується стратегія стохастичного маскування 50% токенів на основі моделювання маскованої мови. **Результати.** Оцінювання на наборі даних Something-Something V2 показало, що повна конфігурація VideoBERT, доменно-орієнтовних адаптерів та Трансформер-декодера з 50 % моделюванням маскованої мови досягає середнього значення BLEU-4 0,1469, ROUGE-L 0,3474 та METEOR 0,2720. Покомпонентне абляційне дослідження демонструє покращення завдяки як двогілковому адаптаційному модулю, так і регуляризації моделювання маскованої мови. **Висновки.** Запропонована архітектура забезпечує модульний та параметрично ефективний підхід до контекстно-орієнтованої генерації описів відео, тоді як явне оцінювання міждоменного перенесення залишається напрямом подальших досліджень.

Ключові слова: контекстно-орієнтована генерація описів відео; VideoBERT; доменно-орієнтовні адаптери; доменна адаптація; просторово-часове представлення; Transformer-декодер; масковане мовне моделювання; параметрично ефективна адаптація

ABOUT THE AUTHORS

Sergii Volodymyrovych Mashtalir - Doctor of Engineering Science, Professor of Informatics Department. Kharkiv National University of Radio Electronics,14, Nauky Ave. Kharkiv, 61166, Ukraine
ORCID: https://orcid.org/0000-0002-0917-6622; sergii.mashtalir@nure.ua. Scopus Author ID: 36183980100
***Research field*:** Image and video processing; data analysis

Машталір Сергій Володимирович - доктор технічних наук, професор кафедри Інформатики Харківського національного університету радіоелектроніки, пр. Науки 14. Харків, 61166, Україна
Область наукових інтересів: Обробка зображень та відео, аналіз даних

Novichonok Mariia Sergiivna - PhD student of Informatics Department, Kharkiv National University of Radio Electronics,14, Nauky Ave. Kharkiv, 61166, Ukraine
ORCID: https://orcid.org/0009-0007-1787-0546;mariia.novichonok@nure.ua
***Research field*:** Image and video processing; data analysis

Новічонок Марія Сергіївна - аспірантка кафедри Інформатики Харківського національного університету радіоелектроніки, пр. Науки 14.Харків, 61166, Україна
Область наукових інтересів: Обробка зображень та відео, аналіз даних