Video-guided Machine Translation with Global Video Context

1st Jian Chen
Shenzhen University
2310413006@email.szu.edu.cn

2nd JinZe Lv
Shenzhen University
lvjinze777@163.com

3rd Zi Long
Shenzhen Technology University
longzi@sztu.edu.cn

4th XiangHua Fu
Shenzhen Technology University
fuxianghua@sztu.edu.cn

***Abstract*—Video-guided Multimodal Translation (VMT) has advanced significantly in recent years. However, most existing methods rely on locally aligned video segments paired one-to-one with subtitles, limiting their ability to capture global narrative context across multiple segments in long videos. To overcome this limitation, we propose a globally video-guided multimodal translation framework that leverages a pretrained semantic encoder and vector database-based subtitle retrieval to construct a context set of video segments closely related to the target subtitle semantics. An attention mechanism is employed to focus on highly relevant visual content, while preserving the remaining video features to retain broader contextual information. Furthermore, we design a region-aware cross-modal attention mechanism to enhance semantic alignment during translation. Experiments on a large-scale documentary translation dataset demonstrate that our method significantly outperforms baseline models, highlighting its effectiveness in long-video scenarios.**

***Index Terms*—Video-guided Multimodal Translation, Global Video Context, Cross-modal Attention**

I. Introduction

Neural machine translation (NMT) has achieved remarkable progress on text-only tasks [1], [2]. However, source texts often suffer from sparsity or ambiguity, hindering accurate translation. Multimodal machine translation (MMT) addresses this by incorporating auxiliary modalities like images or videos to enhance semantic understanding [3], [4].

Early MMT primarily leveraged static images for semantic alignment [5]–[7], integrating visual features during encoding [8], [9] or decoder initialization [10], [11]. Recently, video-guided multimodal translation (VMT) [12], [13] has gained attention, as videos provide richer temporal and spatial cues essential for contextual coherence.

Despite these advances, existing VMT methods often rely on local alignment [14], [15], neglecting the global narrative context crucial for long-form videos. To address this, recent works like TopicVD [16] introduced global context from documentaries. Building on this, we propose a global context-aware framework that integrates distributed visual information to improve translation accuracy beyond local alignment.

Our main contributions are:

- We retrieve subtitle-relevant video segments using semantic retrieval augmented by a context fusion strategy preserving temporal adjacency.
- We introduce a Cross-modal Attention mechanism to adaptively select critical segments, fusing unselected ones to mitigate semantic loss.
- We propose a Region-aware Cross-modal Attention module to selectively aggregate region-level visual semantics via hierarchical token-region interactions.

II. Related Work

Attention mechanisms are central to MMT for integrating visual information. Caglayan et al. [17] and Calixto et al. [18] proposed multimodal attention modules and doubly-attentive decoders, respectively, to dynamically regulate visual feature interaction. Tang et al. [6] further extended MMT to text-only datasets via retrieval-based frameworks.

VMT extends MMT by capturing temporal dynamics. Li et al. [19] and Gu et al. [14] introduced attention mechanisms to resolve subtitle ambiguities and model dynamic spatial-temporal features. Kang et al. [15] improved alignment via cross-modal contrastive learning on large-scale data. However, existing methods mostly rely on local alignment, neglecting global narrative cues essential for long-form videos.

Early datasets like How2 [20] and VATEX [12] provide large-scale pairs but often suffer from text-dependency. Smaller datasets like VISA [13] focus on disambiguation but lack robustness. While recent corpora such as BigVideo [15] and EVA [19] exceed one million samples, they primarily contain short, simple captions. To address this, TopicVD [16] introduces topic-aware, document-derived samples to facilitate global context modeling.

III. The Proposed Method

Fig. 1 illustrates the overall architecture of our proposed framework. As depicted, the model comprises five key components: Global Visual Information Retrieval, a Text-Aware Video Segment Selector, Region-Aware Cross-Modal Attention, Bi-Directional Attention with Gated Fusion and Transformer-based Text Generation [21].

A. Global Retrieval of Visual Information

Given a documentary video $D = \{(S_i, T_i), V_i\}_{i=1}^{N}$, where N denotes the number of subtitles in D, each S_i represents the i-th subtitle, T_i is its human translation, and V_i denotes the corresponding video segment. In our setting, S_i serves as the source input text, while T_i is used as the target output text for translation.

To retrieve global visual information for a given sample (S_i, V_i), we first embed the query subtitle S_i into a high-dimensional semantic space using a pretrained text encoder.

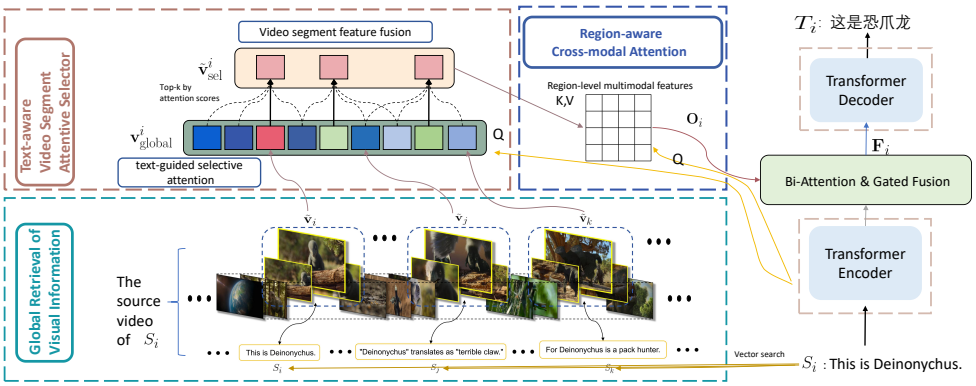

Fig. 1. The Overview of the Framework of the Proposed Method

Based on this embedding, we retrieve the top-P subtitles from the video D whose representations are most semantically similar to S_i:

$$\mathcal{I}^i_{\text{global}} = \arg\max_x \text{Top}(\text{sim}(S_x, S_i), P) \tag{1}$$

where $1 \leq x \leq N$, $\text{sim}(S_x, S_i)$ denotes the semantic similarity between the subtitle S_x and the query subtitle S_i, and $\text{Top}(l, n)$ denotes an operation that selects the n largest values from the list l. Accordingly, $\mathcal{I}^i_{\text{global}} = \{p_1, p_2, \ldots, p_P\}$ represents the indices of the P subtitles that are most semantically similar to S_i in the entire video.

We denote the video segments corresponding to the retrieved relevant subtitle indices as $\mathbf{R}^i_{\text{global}} = \{V_j \mid j \in \mathcal{I}^i_{\text{global}}\}$. To further enrich the contextual information, we exploit the temporal locality commonly observed in documentary-style videos: neighboring segments often describe related events or entities. For each selected segment V_j, we aggregate the feature vectors of its w preceding and w succeeding adjacent segments, scaled by a fixed factor γ to reduce semantic noise. The resulting enhanced segment representation is computed as:

$$\tilde{\mathbf{v}}_j = \mathbf{v}_j + \gamma \cdot \sum_{k \in \mathcal{N}_w(j)} \mathbf{v}_k, \quad j \in \mathcal{I}^i_{\text{global}}, \tag{2}$$

where $\mathcal{N}_w(j)$ denotes the indices of segments adjacent to segment j within a window of size w and $\mathbf{v}_j$ is the feature vector of V_j. The final globally contextualized video set is then constructed by collecting the enhanced features of all selected segments:

$$\mathbf{V}^i_{\text{global}} = \left\{\tilde{\mathbf{v}}_j \mid j \in \mathcal{I}^i_{\text{global}}\right\}. \tag{3}$$

This process effectively identifies subtitle-relevant video segments within potentially redundant content, leveraging a semantic retrieval strategy augmented by context fusion from neighboring segments. By incorporating adjacent segment information, the method preserves temporal and contextual cues, thereby maintaining semantic consistency along the video's temporal dimension.

B. Text-aware Video Segment Attentive Selector

To fully exploit global video context and enhance semantic richness, we propose a Text-aware Video Segment Attentive Selector. This module accepts the contextualized video set $\mathbf{V}^i_{\text{global}}$ as input, adaptively selects semantically salient segments, and integrates information from unselected segments to mitigate information loss.

Given a source subtitle S_i, encoded as $\mathbf{T}_i \in \mathbb{R}^{L \times E_t}$, we quantify the cross-modal relevance using attention-based alignment scores:

$$\boldsymbol{\alpha} = \text{Softmax}([s_1, \ldots, s_P]), \quad \text{where } s_j = \text{Attn}(\mathbf{T}_i, \tilde{\mathbf{v}}_j) \tag{4}$$

Here, $\text{Attn}(\cdot)$ denotes the attention function with $\mathbf{T}_i$ as the query and $\tilde{\mathbf{v}}_j \in \mathbf{V}^i_{\text{global}}$ as the key. Subsequently, we select the top-K relevant video segments to constitute the primary visual context $\mathbf{V}^i_{\text{sel}}$:

$$\mathcal{I}^i_{\text{sel}} = \arg\max_x \text{Top}(\alpha_x, K) \tag{5}$$

$$\mathbf{V}^i_{\text{sel}} = \left\{\tilde{\mathbf{v}}_j \mid j \in \mathcal{I}^i_{\text{sel}}\right\} \tag{6}$$

where $\mathcal{I}^i_{\text{sel}}$ denotes the indices of the K segments most relevant to S_i.

To preserve a broader scope of context, we exploit temporal locality by aggregating information from the unselected set $\mathbf{V}^i_u$. Specifically, for each selected segment $\tilde{\mathbf{v}}_j \in \mathbf{V}^i_{\text{sel}}$, we apply the following update rule:

$$\tilde{\mathbf{v}}'_j = \tilde{\mathbf{v}}_j + \frac{\lambda}{2} \sum_{K_{j-1} \leq k \leq K_{j+1}, \tilde{\mathbf{v}}_k \in \mathbf{V}^i_u} \tilde{\mathbf{v}}_k \tag{7}$$

where λ is a fixed weighting coefficient, and K_j denotes the index of $\tilde{\mathbf{v}}_j$ in $\mathbf{V}^i_{\text{global}}$, so the summation aggregates unselected segments lying between the $(j-1)$-th and $(j+1)$-th selected segments in the globally retrieved set. This enriches each selected segment with neighboring context while maintaining temporal coherence. The final updated set is denoted as:

$$\tilde{\mathbf{V}}^i_{\text{sel}} = \{\tilde{\mathbf{v}}'_j \mid j \in \mathcal{I}^i_{\text{sel}}\} \tag{8}$$

Distinct from the static construction in Section III-A, this mechanism is optimized during training, facilitating the dynamic identification of semantically relevant segments. Furthermore, by incorporating a context-aware fusion strategy, the approach mitigates local semantic loss, preserving a coherent visual context for downstream multimodal translation.

C. Region-Aware Cross-modal Attention

To facilitate fine-grained alignment between textual and visual region-level features, we propose a Region-aware Cross-modal Attention module. This module enables deep cross-modal fusion via hierarchical semantic interactions.

Let $\tilde{\mathbf{V}}^i_{\text{sel}} \in \mathbb{R}^{K\times R\times E_v}$ denote the visual features, where K, R, and E_v represent the number of video segments, spatial regions, and feature dimensions, respectively. Here, R denotes the number of spatial grid cells in the I3D feature maps, each corresponding to a local region of the video frame. To unify feature spaces, we first project the textual features $\mathbf{T}_i$ into the visual space via a learnable linear transformation:

$$\mathbf{T}'_i = \mathcal{W}_t\mathbf{T}_i + \mathbf{b}_t, \quad \mathbf{T}'_i \in \mathbb{R}^{L\times E_v} \tag{9}$$

Subsequently, $\mathbf{T}'_i$ is expanded to $\mathbb{R}^{R\times L\times E_v}$ by repeating along the region dimension, and $\tilde{\mathbf{V}}^i_{\text{sel}}$ is reshaped to $\mathbb{R}^{R\times K\times E_v}$, enabling each spatial region to independently attend to the full text sequence.

We employ multi-head attention where queries are derived from textual features, and keys/values from video features. The attended representation $\mathbf{Z}_i$ is computed via scaled dot-product attention:

$$\mathbf{Z}_i = \text{Softmax}\left(\frac{(\mathcal{W}_q\mathbf{T}'_i)(\mathcal{W}_k\tilde{\mathbf{V}}^i_{\text{sel}})^\top}{\sqrt{d_k}}\right)(\mathcal{W}_v\tilde{\mathbf{V}}^i_{\text{sel}}) \tag{10}$$

Finally, we apply average pooling over the region dimension R, followed by an output projection to yield the fused representation $\mathbf{O}_i \in \mathbb{R}^{L\times E_o}$:

$$\mathbf{O}_i = \mathcal{W}_o\left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Z}_i\right) + \mathbf{b}_o \tag{11}$$

where $\mathcal{W}_o$ and $\mathbf{b}_o$ are parameters of the output projection layer. This mechanism explicitly integrates regional visual semantics into textual representations, effectively bridging the modality gap.

D. Bi-directional Cross-modal Attention with Gated Fusion

To facilitate deep semantic interaction, we introduce a Bi-directional Cross-modal Attention mechanism [22], [23] that simultaneously captures Text-to-Video (T2V) and Video-to-Text (V2T) dependencies.

In the T2V direction, textual features $\mathbf{T}_i$ query the video features $\mathbf{O}_i$ to obtain the context-aware representation $\mathbf{H}_{\mathbf{t2v}}$:

$$\mathbf{H}_{\mathbf{t2v}} = \text{Softmax}\left(\frac{\mathbf{T}_i\cdot\mathbf{O}_i^\top}{\sqrt{d_h}}\right)\cdot\mathbf{O}_i \tag{12}$$

Conversely, in the V2T direction, video features $\mathbf{O}_i$ serve as queries to attend to textual features $\mathbf{T}_i$:

$$\mathbf{H}_{\mathbf{v2t}} = \text{Softmax}\left(\frac{\mathbf{O}_i\cdot\mathbf{T}_i^\top}{\sqrt{d_h}}\right)\cdot\mathbf{T}_i \tag{13}$$

To integrate these representations, we employ a gated fusion mechanism [24], [25] that adaptively balances visual and linguistic information. The final fused representation $\mathbf{F}_i$ is formulated as:

$$\mathbf{F}_i = (1-\mathbf{g})\odot\mathbf{H}_{\mathbf{v2t}} + \mathbf{g}\odot\mathbf{H}_{\mathbf{t2v}} + (1-\mathbf{g})\odot\mathbf{T}_i \tag{14}$$

where $\odot$ denotes element-wise multiplication. The gating matrix $\mathbf{g}$ is a learnable parameter that dynamically modulates the relative contribution of each cross-modal pathway during fusion.

E. Transformer-based Text Generation

The fused representation $\mathbf{F}_i$ encapsulates bidirectional cross-modal contextual information and serves as the multimodal input to the Transformer decoder, facilitating joint attention to both linguistic and visual semantics throughout the generation process. At each decoding step t, the decoder interacts with $\mathbf{F}_i$ via multi-head cross-attention, formulated as:

$$\mathbf{C}_t = \text{Softmax}\left(\frac{\mathbf{Q}_{\text{dec}}(t)\mathbf{F}_i^\top}{\sqrt{d_k}}\right)\mathbf{F}_i \tag{15}$$

where $\mathbf{Q}_{\text{dec}}(t)$ denotes the decoder's evolving query vector, and $\mathbf{F}_i$ functions as unified key-value pairs. The resulting contextual vector $\mathbf{C}_t$ is subsequently processed by the decoder's feed-forward layers to produce the hidden state $\mathbf{h}_t$.

The generation mechanism proceeds autoregressively through sequential token prediction. Specifically, at decoding step t, the hidden state $\mathbf{h}_t$ is projected linearly onto the target vocabulary space via $\mathbf{o}_t = \mathbf{W}_o\mathbf{h}_t + \mathbf{b}_o$, where $\mathbf{W}_o \in \mathbb{R}^{V\times d_h}$, yielding logits that are normalized by the softmax function to form the conditional probability distribution:

$$P(y_t \mid y_{1:t-1}, \mathbf{F}_i) = \text{Softmax}(\mathbf{o}_t) \tag{16}$$

The predicted token $\hat{y}_t$ is obtained by maximizing this conditional probability, formally expressed as:

$$\hat{y}_t = \arg\max_{y\in\text{vocab}} P(y_t = y) \tag{17}$$

This predicted token is then fed forward to generate the subsequent query vector $\mathbf{Q}_{\text{dec}}(t+1)$ for the next decoding iteration. This process continues iteratively until an end-of-sequence (EOS) token is produced, resulting in the complete translated sequence:

$$\hat{\mathbf{y}} = \{\hat{y}_1, \ldots, \hat{y}_T\}, \quad \hat{y}_T = \langle\text{EOS}\rangle \tag{18}$$

IV. Experiments

A. Dataset

We conduct experiments on the TopicVD dataset [16], comprising 256 documentary videos and 122,930 English–Chinese subtitle pairs. The dataset is split into 102,502 training pairs (219 videos), 10,429 validation pairs (18 videos), and 9,999 testing pairs (19 videos).

To assess generalizability, we also evaluate our method on the BigVideo dataset [15]. To ensure a fair comparison with TopicVD regarding data scale, we utilize a subset containing the first 104,790 training pairs, while keeping the original validation and test splits unchanged.[1]

[1]We also experimented on other VMT datasets [12], [13], [26]. However, due to their lack of rich contextual video information, our method yields performance comparable to existing approaches.

TABLE I
PERFORMANCE COMPARISON ON THE TOPICVD DATASET

Method	BLEU	METEOR
Text-only NMT	19.14	33.95
Image-MMT	28.93	38.59
Segment-Level Video-MMT	29.23	38.65
BigVideo (Kang et al., 2023)	25.92	36.54
Proposed Method	**30.47**	**38.96**

B. Baseline Models

We compare our proposed approach against the following baselines:

- **Text-only NMT:** A standard Transformer-based NMT model [21], implemented via the open-source `transformer-translator-pytorch` project[2].
- **Image-MMT:** An efficient visually-enhanced baseline where visual features are extracted from sampled frames. We employ FAISS [27] to retrieve the single most relevant frame for each subtitle as auxiliary input.
- **Segment-Level Video-MMT:** Utilizes only the video segment temporally aligned with the target subtitle as visual input, ignoring global context.
- **BigVideo (Kang et al.):** We reproduce the method proposed in [15], which leverages cross-modal contrastive learning to align video and text representations. This serves as a strong baseline for video-guided translation.

All models are trained on TopicVD and evaluated using 4-gram BLEU [28] and METEOR [29].

C. Experimental Settings

Model Architecture. We implement our model using Fairseq [30], following the configuration in [31]. The architecture consists of 4 encoder and decoder layers, with a hidden dimension of 128, feed-forward size of 256, 8 attention heads, and a dropout rate of 0.35. Video features are extracted via a pretrained I3D model [32] and linearly projected to align with text embeddings. The total computational complexity is approximately 1.22 GFLOPs per sample.

Hyperparameters. For global retrieval, we utilize CLIP [33] and FAISS [27] to select the top $P = 10$ subtitles. Context enrichment employs a window of $w = 2$ with a scaling factor $\gamma = 0.1$. In the selector module, we set $K = 5$ and the fusion weight $\lambda = 0.1$.[3]

Optimization. We train using RAdam [34] with a learning rate of 0.001, an inverse square root scheduler, and 4000 warm-up steps. The objective is cross-entropy with label smoothing (0.1) and a batch size of 4096 tokens. Training proceeds for up to 20,000 steps with early stopping after 10 epochs of no improvement. Mixed-precision (FP16) training is adopted for efficiency.

[2]https://github.com/devjwsong/transformer-translator-pytorch

[3]Both γ and λ are set to 0.1 to balance context enhancement against noise. Future work may explore adaptive weighting strategies.

TABLE II
PERFORMANCE COMPARISON (BLEU AND METEOR) ON THE BIGVIDEO DATASET

Method	BLEU	METEOR
BigVideo (Kang et al., 2023)	44.44	—
BigVideo (our implementation)	**47.05**	**49.63**
Proposed Method	<u>46.78</u>	<u>49.45</u>

D. Experimental Results

To ensure robustness, we report average BLEU [28] and METEOR [29] scores over three independent runs with distinct random seeds.

Results on TopicVD. As shown in Table I, our proposed method achieves the best performance on TopicVD (30.47 BLEU / 38.96 METEOR). It significantly outperforms the Text-only baseline (+11.33 BLEU) and the Image-MMT model (+1.54 BLEU). Crucially, within the video-guided domain, our approach surpasses the Segment-Level Video-MMT and BigVideo [15] baselines by 1.24 and 4.55 BLEU points, respectively, demonstrating the efficacy of global context modeling.

Results on BigVideo. Table II presents results on the BigVideo subset. Our re-implementation of the BigVideo model achieves 47.05 BLEU. Under the same setting, our method attains a comparable score of 46.78 BLEU.[4] This performance parity is expected, as BigVideo consists primarily of short segments with limited temporal context, rendering our long-range modeling less advantageous than in documentary scenarios.

Qualitative Analysis. Fig. 2 illustrates a representative case. The Segment-Level baseline fails to correctly translate "dry" due to misleading local visual cues. In contrast, our method successfully resolves this ambiguity by retrieving globally contextualized video segments, providing sufficient semantic evidence to generate an accurate translation.

V. ANALYSIS AND DISCUSSION

This section analyzes the impact of three critical hyperparameters introduced in Sections III-A and III-B: retrieval size P, context window w, and selection count K. We further conduct ablation studies to validate the contributions of core modules.

A. Hyperparameter Sensitivity

Table III summarizes the performance variations under different hyperparameter settings.

Effect of P. With $w = 2$ and $K = 5$, performance peaks at $P = 10$ (30.47 BLEU). Both smaller ($P = 5$) and larger ($P = 20$) values degrade performance, indicating that $P = 10$ strikes an optimal balance between sufficient contextual diversity and noise suppression.

[4]The inconsistency between our re-implementation and the original result is attributable to the use of a subset of the BigVideo dataset rather than hyperparameter tuning. For our proposed method, only the retrieval hyperparameters P, w, and K were adjusted to fit this subset.

Fig. 2. Example of correct translation by the proposed method. The source word "dry" is ambiguous in Chinese translation, requiring visual context to disambiguate between "干燥" (lack of moisture in air) and "干涸" (depletion of a water source).

TABLE III
ABLATION STUDIES ON HYPERPARAMETERS P, w, AND K ON THE TOPICVD DATASET

Parameter	Value	BLEU	METEOR
P ($w = 2, K = 5$)	5	29.71	38.04
	10	**30.47**	**38.96**
	20	30.23	38.68
w ($P = 10, K = 5$)	1	29.82	38.18
	2	**30.47**	**38.96**
	3	30.17	38.21
	4	30.39	38.43
	5	30.06	38.03
K ($P = 10, w = 2$)	3	30.14	38.21
	4	30.15	38.24
	5	**30.47**	**38.96**
	6	29.85	38.34
	7	29.72	38.31
	10	29.51	38.22

TABLE IV
ABLATION RESULTS ON THE TWO CORE MODULES: GR AND TVSS

Setting	BLEU	METEOR
Full Model	**30.47**	**38.96**
w/o GR	29.59	38.14
w/o TVSS	29.71	38.04
w/o Both Modules	29.48	38.06

Effect of w. Fixing $P = 10$ and $K = 5$, we find $w = 2$ yields the best results. Narrower windows ($w = 1$) limit contextual cues, whereas excessively wide windows ($w = 5$) introduce redundancy and noise, hampering translation accuracy.

Effect of K. With $P = 10$ and $w = 2$, the model achieves the highest scores at $K = 5$. Lower values ($K < 5$) risk insufficient visual grounding, while higher values ($K > 5$) tend to incorporate irrelevant or conflicting visual signals.

B. Ablation Study

To assess the efficacy of our proposed components, we conduct an ablation study by systematically removing the Global Retrieval (GR) and Text-aware Video Segment Attentive Selector (TVSS) modules. Results are detailed in Table IV.

Impact of Global Retrieval. Removing the GR module leads to a performance drop of 0.88 BLEU and 0.82 METEOR. This highlights the necessity of retrieving globally distributed segments to capture long-range dependencies, which are crucial for maintaining narrative coherence in documentaries.

Impact of Text-aware Selector. Excluding the TVSS module reduces BLEU to 29.71. This suggests that while retrieval provides candidates, our adaptive selection and fusion strategy is essential for enriching visual-semantic density and preserving subtle yet meaningful cues.

Joint Contribution. Removing both modules results in the lowest performance (29.48 BLEU). This confirms that the global retrieval and adaptive selection mechanisms are complementary, jointly enabling the model to construct comprehensive and semantically aligned visual representations.

C. Generalizability on Large Vision-Language Models

To verify the scalability and generalizability of our method within the paradigm of large vision–language models (VLMs), we further evaluate the effectiveness of the selectively incorporating global video information strategy described above on Qwen3-VL [35]. Distinct from the standard uniform sampling strategy adopted by Qwen3-VL, we first explore simple alternatives, including dot-product similarity and a jointly learned selection module, to selectively filter video frames based on their textual relevance, thereby constructing a semantically concentrated visual context.

We conduct validation experiments on Qwen3-VL-2B-Instruct [5]. The results indicate that the BLEU score improves

[5]https://huggingface.co/Qwen/Qwen3-VL-2B-Instruct

from 26.70 for the base VLM to 27.75 when using dot-product similarity, and further increases to 28.63 with the jointly learned selection module. These findings fully underscore the effectiveness of our method, demonstrating that the proposed text-aware frame selection mechanism substantially enhances the model's cross-modal alignment capabilities.

VI. Conclusions

In this work, we propose a novel global context-aware multimodal translation framework that leverages semantic relationships across video segments to improve VMT. Unlike existing methods that rely mainly on locally aligned video–text pairs, our approach constructs a coherent, context-rich video segment collection by retrieving topically related subtitles via vector-based similarity search and expanding context with adjacent segments. The experimental results confirm the effectiveness of the proposed method, while the experiments on vision–language models further demonstrate the feasibility of our approach for leveraging global video information. As future work, we plan to extend the global video context utilization strategies described in the preceding sections to VLMs, with the goal of enhancing their ability to process long video inputs.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[2] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[3] L. Specia, S. Frank, K. Sima'An, and D. Elliott, "A shared task on multimodal machine translation and crosslingual image description," in *First Conference on Machine Translation*. Association for Computational Linguistics (ACL), 2016, pp. 543–553.

[4] U. Sulubacak, O. Caglayan, S.-A. Grönroos, A. Rouhe, D. Elliott, L. Specia, and J. Tiedemann, "Multimodal machine translation through visuals and speech," *Machine Translation*, vol. 34, pp. 97–147, 2020.

[5] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual english-german image descriptions," *arXiv preprint arXiv:1605.00459*, 2016.

[6] Z. Tang, X. Zhang, Z. Long, and X. Fu, "Multimodal neural machine translation with search engine based image retrieval," *arXiv preprint arXiv:2208.00767*, 2022.

[7] Z. Long, Z. Tang, X. Fu, J. Chen, S. Hou, and J. Lyu, "Exploring the necessity of visual modality in multimodal machine translation using authentic datasets," in *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC)@ LREC-COLING 2024*, 2024, pp. 36–50.

[8] I. Calixto, Q. Liu, and N. Campbell, "Incorporating global visual features into attention-based neural machine translation," *arXiv preprint arXiv:1701.06521*, 2017.

[9] P.-Y. Huang, F. Liu, S.-R. Shiang, J. Oh, and C. Dyer, "Attention-based multimodal neural machine translation," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 2016, pp. 639–645.

[10] D. Elliott, S. Frank, and E. Hasler, "Multilingual image description with neural sequence models," *arXiv preprint arXiv:1510.04709*, 2015.

[11] P. S. Madhyastha, J. Wang, and L. Specia, "Sheffield multimt: Using object posterior predictions for multimodal machine translation," in *Proceedings of the second conference on machine translation*, 2017, pp. 470–476.

[12] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4581–4591.

[13] Y. Li, S. Shimizu, W. Gu, C. Chu, and S. Kurohashi, "Visa: An ambiguous subtitles dataset for visual scene-aware machine translation," *arXiv preprint arXiv:2201.08054*, 2022.

[14] W. Gu, H. Song, C. Chu, and S. Kurohashi, "Video-guided machine translation with spatial hierarchical attention network," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, 2021, pp. 87–92.

[15] L. Kang, L. Huang, N. Peng, P. Zhu, Z. Sun, S. Cheng, M. Wang, D. Huang, and J. Su, "Bigvideo: A large-scale video subtitle translation dataset for multimodal machine translation," *arXiv preprint arXiv:2305.18326*, 2023.

[16] J. Lv, J. Chen, Z. Long, X. Fu, and Y. Chen, "Topicvd: A topic-based dataset of video-guided multimodal machine translation for documentaries," *arXiv preprint arXiv:2505.05714*, 2025.

[17] O. Caglayan, L. Barrault, and F. Bougares, "Multimodal attention for neural machine translation," *arXiv preprint arXiv:1609.03976*, 2016.

[18] I. Calixto, Q. Liu, and N. Campbell, "Doubly-attentive decoder for multimodal neural machine translation," *arXiv preprint arXiv:1702.01287*, 2017.

[19] Y. Li, S. Shimizu, C. Chu, S. Kurohashi, and W. Li, "Video-helpful multimodal machine translation," *arXiv preprint arXiv:2310.20201*, 2023.

[20] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: a large-scale dataset for multimodal language understanding," *arXiv preprint arXiv:1811.00347*, 2018.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[22] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in neural information processing systems*, vol. 32, 2019.

[23] J. Su, J. Chen, H. Jiang, C. Zhou, H. Lin, Y. Ge, Q. Wu, and Y. Lai, "Multi-modal neural machine translation with deep semantic interactions," *Information Sciences*, vol. 554, pp. 47–60, 2021.

[24] Y. Yin, F. Meng, J. Su, C. Zhou, Z. Yang, J. Zhou, and J. Luo, "A novel graph-based multi-modal fusion encoder for neural machine translation," *arXiv preprint arXiv:2007.08742*, 2020.

[25] H. Lin, F. Meng, J. Su, Y. Yin, Z. Yang, Y. Ge, J. Zhou, and J. Luo, "Dynamic context-guided capsule network for multimodal machine translation," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1320–1329.

[26] A. Teramen, T. Ohtsuka, R. Kondo, T. Kajiwara, and T. Ninomiya, "English-to-japanese multimodal machine translation based on image-text matching of lecture videos," in *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, 2024, pp. 86–91.

[27] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[29] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*. ACL, 2005, pp. 65–72.

[30] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," *arXiv preprint arXiv:1904.01038*, 2019.

[31] B. Li, C. Lv, Z. Zhou, T. Zhou, T. Xiao, A. Ma, and J. Zhu, "On vision features in multimodal machine translation," *arXiv preprint arXiv:2203.09173*, 2022.

[32] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[34] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.

[35] Q. Team, "Qwen3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2505.09388