Multimodal Video Ordering via Task-Specific Pre-Training and Alignment-Guided Fine-Tuning

Yiping Yang✉ Yingming Li
College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310000, China
{yangyiping, yingming}@zju.edu.cn

Abstract

Multimodal sequence coherence modeling is crucial for understanding and generating temporally and logically coherent content. However, existing methods suffer from the scarcity of datasets that cover diverse temporal structures and rich logical dependencies in various scenarios. Therefore, we introduce WikiHVO, a high-quality video ordering dataset tailored to reflect the real-world temporal and logical diversity. Moreover, most coherence modeling approaches struggle to incorporate task-specific structural priors while capturing fine-grained multimodal interactions. To address these issues, we propose a unified multimodal ordering framework, VOTA (Video Ordering with Task-specific pre-training and Alignment-guided fine-tuning), that employs a two-stage training strategy to facilitate multimodal coherence reasoning. It first undergoes task-oriented pre-training on large-scale text corpora to learn ordering-required semantic logics, then is fine-tuned on our curated dataset with explicit multimodal alignment to strengthen cross-modal synergy. This design promotes VOTA to effectively bridge textual and visual semantics while preserving temporal and logical coherence. Extensive experiments and analyses demonstrate the effectiveness and superiority of VOTA. Our code and dataset are publicly available at: https://github.com/gxytsyt/MM_Video_Ordering.

Keywords: ***Multimodal Ordering, Cross-Modal Alignment, Task-Specific Pre-Training, Video Ordering Dataset.***

1. Introduction

Coherence is a fundamental property of information organization, which reflects the logical and temporal relationships between individual units and directly influences the effectiveness and comprehensibility of information [28, 29]. Sequence ordering, as a typical task in coherence modeling, aims to reorganize discrete elements while maintaining temporal and logical consistency [4, 8, 36]. In particular, multimodal sequence ordering has gradually attracted increasing attention because real-world sequence ordering scenarios inherently involve multimodal information [38]. Figure 1 illustrates an example of the video ordering task, where the model is required to comprehend both video clips and associated textual descriptions of each operational step, and reconstruct a coherent sequential order. This task requires the integration of multiple cognitive and modeling capabilities, including multimodal comprehension, temporal reasoning, and causal inference, which not only enhance sequence coherence but also provide support for general multimodal understanding and generation systems [3, 38]. It is of significant importance in diverse applications, such as multimodal content understanding, summarization, and robotic procedure planning [1, 2, 13, 35, 45].

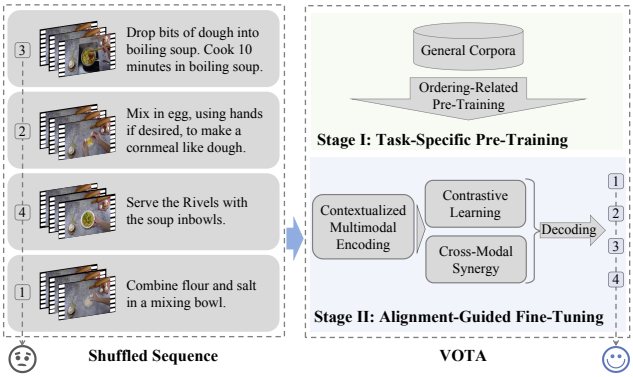

Figure 1. Illustration of the multimodal video ordering task. The left part presents an example input consisting of unordered video-text step pairs, while the right part illustrates our proposed two-stage framework. The objective is to reconstruct a coherent video sequence from the unordered input.

Early sequence ordering approaches primarily rely on statistical and linguistic features [4, 23], limited by manual feature engineering and generalization capabilities. With the rise of deep learning and large-scale pre-trained models [11, 24], neural network-based methods [8, 9, 10, 15, 31, 36, 41, 43] have been applied to sequence ordering and achieved promising results. In this work, we mainly focus

on the semantic understanding and ordering of cross-modal video-text content, aiming to enhance coherence modeling in realistic multimodal scenarios. In this regard, recent studies have introduced multimodal encoding modules [38], hierarchical modeling structures [42], cross-modal guidance mechanisms [5, 6], and diverse decoding strategies [14], showing notable improvements in multimodal understanding and reasoning performance.

Despite these advancements, existing multimodal ordering methods still face several challenges.

First, the availability of video sequence ordering datasets with diverse temporal dynamics in real-world scenarios remains limited. Most existing datasets focus on textual or image-based inputs [7, 19, 30, 38]. Although the Chinese video ordering dataset [14] provides preliminary resources, the narrow domain coverage hinders its applicability. The absence of multimodal video ordering datasets in real-world settings constrains current models' ability to handle complex and varied sequences, thereby restricting their performance and generalization in practical applications.

Second, current ordering models typically rely on simple concatenation or summation of multimodal features followed by transformer-based modeling, which does not fully consider the inherent heterogeneity of multimodal data and thus has difficulty in building semantic correspondence between modalities for effective fusion [20]. In contrast, cross-modal contrastive learning frameworks have proven effective in modeling the alignment relationships of multimodal features [12, 25, 26, 32, 37, 40], and have been widely used for learning generalized multimodal representations. However, its potential in multimodal ordering systems remains largely unexplored.

Finally, although pre-trained models are commonly adopted as backbones to capture basic feature representations for sequence ordering [10, 14, 41, 42], general pre-training architectures cannot capture ordering-related patterns and semantic structures. Instead, task-specific pre-training has demonstrated its potential to extract discriminative features from large-scale data [16, 27, 44]. However, the effectiveness of ordering-specific pre-training has not yet been fully leveraged.

Considering these limitations, we propose a novel Video Ordering framework with Task-specific pre-training and Alignment-guided fine-tuning (VOTA) to facilitate multimodal coherence reasoning. First, to tackle the scarcity of video datasets in real-world scenarios, we construct a high-quality, multi-domain video-text ordering dataset, termed WikiHVO. This dataset is derived from the widely used online knowledge platform WikiHow, which encompasses a broad spectrum of real-world operational scenarios, covering complex temporal structures and logical dependencies across diverse categories. Each sample sequence is composed of sequential steps with video clips and text descriptions, making it an ideal resource for studying multimodal coherence. In addition, to overcome the existing challenge where models struggle to incorporate task-specific structural patterns while effectively capturing fine-grained multimodal interactions, we propose a two-stage training strategy that effectively captures generalizable coherence reasoning patterns while strengthening fine-grained cross-modal interactions. In the first stage, we pre-train the model on a large-scale general text corpus to acquire ordering-related semantics, temporal dependencies, and reasoning patterns intrinsic to sequence coherence. In the second stage, we fine-tune the model on our curated multimodal ordering dataset, enabling it to transfer ordering-related patterns to multimodal settings and capture intricate dependencies between visual and textual modalities. In addition, we introduce an explicit visual-text alignment loss to enhance cross-modal consistency and semantic correspondence, enabling the model to perform more coherent reasoning and generate temporally and logically consistent multimodal sequences.

Our contributions can be concluded as follows:

- We introduce WikiHVO, a high-quality dataset for the video sequence ordering task. The dataset is publicly released to facilitate further research.
- We propose VOTA, a two-stage ordering framework specifically designed for the multimodal sequence ordering task, and introduce an explicit visual-text alignment loss to ensure consistent and effective multimodal synergy.
- We conduct extensive experiments on the video sequence ordering task. The results show that VOTA achieves state-of-the-art performance.

2. Related Work

In this section, we review sequence ordering methods, which can be broadly divided into two categories: textual sequence ordering and multimodal sequence ordering.

2.1. Textual Sequence Ordering

Text ordering aims to rearrange shuffled sentences into the original coherent sequence by leveraging semantic and logical relationships [34]. It is considered as a fundamental problem in natural language understanding. Early approaches primarily rely on statistical models and linguistic features to model coherence, including the Content Model [4], which captures topic transitions using Hidden Markov Models (HMMs), and the Entity-Grid Model [3] that learns entity transition probabilities. Although demonstrating reasonable effectiveness, these methods are heavily dependent on handcrafted features and domain-specific knowledge. With the rise of deep learning, sequence ordering

models have achieved substantial improvements. Encoder-decoder architectures based on deep neural networks, such as LSTMs [18], often combined with pointer networks [36], have been widely adopted in text ordering tasks [10, 22, 43], demonstrating superior performance in reconstructing coherent document structures compared to traditional approaches. More recently, pre-trained language models have brought substantial improvements to text ordering tasks. Models such as BERT, BART, and T5 [11, 24, 33] excel at encoding sentence-level semantics and contextual information, thereby providing powerful semantic representations that greatly enhance ordering effectiveness [8, 41].

Despite the significant progress of these models, most existing methods are confined to the textual modality, limiting their effectiveness in real-world applications that involve multimodal information. When dealing with more complex multimodal sequences, effectively understanding the content and reasoning about the sequential order remains a challenge.

2.2. Multimodal Sequence Ordering

As multimodal data is inherently present in many real-world applications, effectively modeling temporal dependencies and logical relationships within multimodal sequences has emerged as a critical research problem. Compared to traditional text ordering tasks, multimodal sequence ordering poses greater challenges in both semantic modeling and temporal reasoning.

Recent work on multimodal sequence understanding has introduced ordering-related pre-training tasks to help models capture temporal structures, such as randomly shuffling frames [27] or swapping frame pairs [44], followed by a classification task to determine if the sequence is in the correct order. While effective for learning temporal consistency, these methods treat ordering as an auxiliary classification objective, rather than directly tackling the reconstruction of coherent multimodal sequences.

To bridge this gap, researchers have begun to develop specialized approaches for multimodal sequence ordering. For instance, Wu *et al*. [38] extend the existing text-based ordering architecture, BERSON [10], by incorporating a multimodal encoding module, enabling joint modeling of image-text information for ordering. And they also incorporate pre-training to model pairwise relationships between sequence elements, which improves ordering performance. NACON [6] introduces an order-invariant context encoder combined with a non-autoregressive ordering module, effectively integrating multimodal contextual information. Additionally, SVO [14] proposes a video-text ordering framework that encodes video and text features separately and incorporates position decoding and successor prediction tasks for generation. However, these models predominantly rely on step-level or pairwise encoding schemes, which limit their ability to capture deep semantic structures and logical dependencies at the sequence level. To address this limitation, MHAONet [42] proposes a transformer-based architecture for sequence-level multimodal encoding and ordering, aiming to improve the coherence of the generated sequence through global sequence modeling. While achieving promising results, these methods still lack explicit guidance on visual-text semantic alignment and fusion, potentially limiting the effectiveness of multimodal integration and understanding.

Beyond methodological challenges, the scarcity of publicly available datasets capturing complex temporal and logical structures in real-world scenarios remains a key limitation. Existing ordering datasets primarily focus on image-text pairs or narrowly defined contexts [14, 38], offering limited coverage of the diverse structures encountered in practical tasks. Although the Chinese video ordering dataset [14] provides a valuable benchmark, it is confined to a specific domain of pre-edited media (e.g., film commentary), limiting its applicability for modeling real-world procedural coherence. The absence of high-quality video ordering datasets reflecting authentic procedural activities continues to hinder the development and deployment of effective sequence ordering models.

In summary, multimodal sequence ordering remains a challenging task. To address this, we introduce a high-quality English video ordering dataset, WikiHVO, which covers diverse real-world scenarios. Additionally, we propose a two-stage training framework that leverages task-specific pre-training to enable the model to learn ordering-related patterns. Furthermore, we incorporate an explicit visual-text alignment loss to facilitate fine-grained cross-modal interaction modeling, thereby enabling effective multimodal synergy and comprehension.

3. Methodology

In this section, we first formalize the definition of the video-text ordering task and then provide a detailed description of our proposed framework, VOTA. As illustrated in Figure 2, the framework consists of three main components: a contextualized multimodal encoding module, a cross-modal alignment and fusion module, and a pointer-based ordering decoder. Finally, we elaborate on our two-stage training strategy, where a carefully designed task-specific pre-training strategy followed by alignment-guided fine-tuning enables effective video-text understanding and coherent sequence modeling.

3.1. Task Definition

In the video sequence ordering task, each data sample consists of a set of video clips and the corresponding textual descriptions. Formally, given an unordered set of multimodal segment steps $S = [S_{o_1}, S_{o_2}, \ldots, S_{o_N}]$, where each

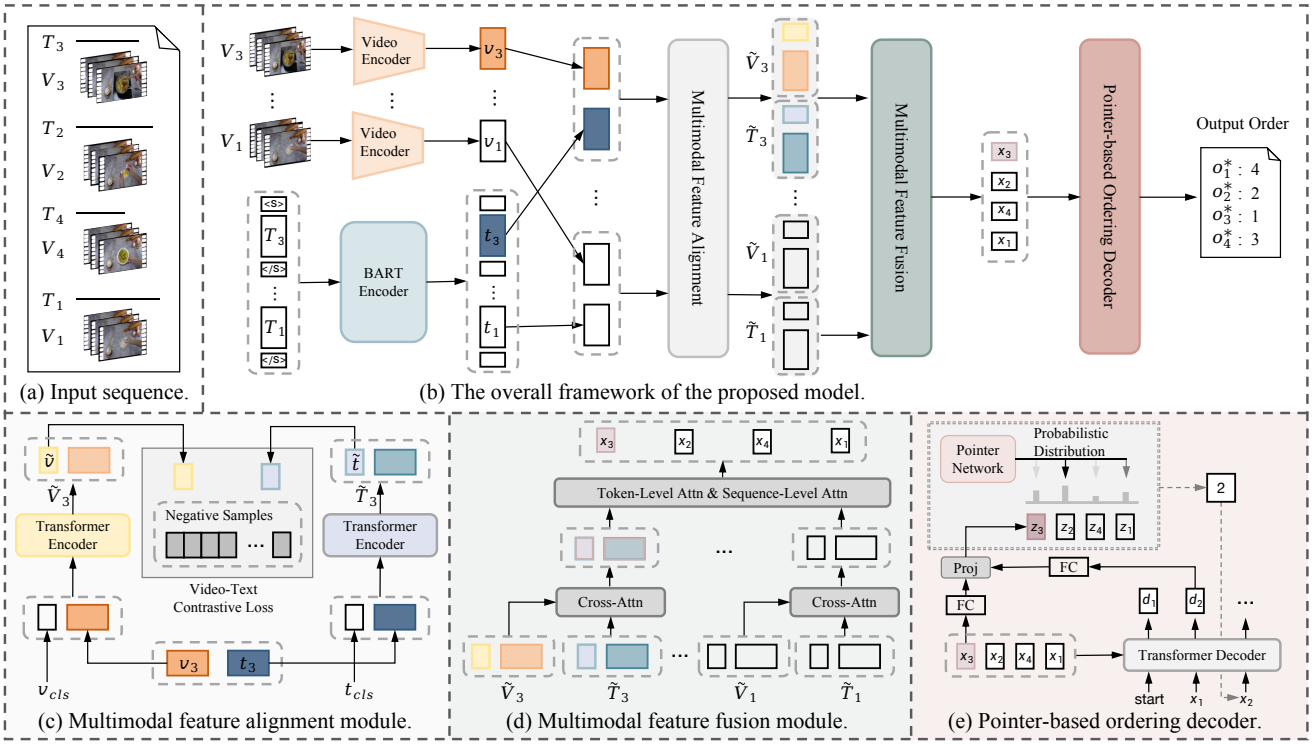

Figure 2. The architecture of VOTA. (a) Input sequence: Unordered set of N video-text pairs. (b) Overall framework: Contextualized encoding, cross-modal alignment and fusion, and pointer-based decoding. (c) Multimodal feature alignment: Cross-modal contrastive learning for video-text pairs. (d) Multimodal feature fusion: Hierarchical encoding of aligned multimodal features. (e) Pointer-based decoding: Transformer-based pointer network for autoregressive ordering.

step S_{o_i} consists of the text component T_{o_i} and the video component V_{o_i}, N is the number of steps in the sequence. Initially, the steps are presented in a randomly shuffled manner, and the goal is to reconstruct a coherent and logically consistent sequence order $o^* = [o_1^*, o_2^*, \ldots, o_N^*]$.

3.2. Contextualized Multimodal Encoding

In this section, we detail the extraction of features from both the textual descriptions and video clips associated with each operational step in the sequence. To effectively capture context-aware features, we employ a paragraph-level text encoding strategy alongside a video clip encoding module.

3.2.1 Context-Aware Paragraph-Level Text Encoder

The text encoding module is designed to capture paragraph-level semantic representations for each step. Specifically, the textual descriptions of the shuffled steps are fed into a transformer-based BART encoder [24], which leverages multi-layer self-attention mechanisms to model both intra-step semantics and inter-step dependencies, thereby obtaining contextually enriched textual representations.

In particular, we first concatenate the textual descriptions of the unordered steps using the </s> token as the delimiter. The input sequence is structured as: $[\text{<s>}, T_{o_1}, \text{</s>}, T_{o_2}, \text{</s>}, ..., T_{o_N}, \text{</s>}]$, where T_{o_i} denotes the textual description of the i-th step. For each token in the sequence, its initial embedding is computed by summing the corresponding token embedding and positional embedding. The resulting sequence representation X_{text} is then fed into a stack of K self-attention layers to capture fine-grained, paragraph-level interactions among tokens. At each transformer layer, the input features are first processed through a multi-head self-attention mechanism to enable cross-token interactions, followed by residual connections and layer normalization. The calculation process at layer l is formally defined as:

$$\tilde{\mathbf{X}}_{text}^{(l)} = LN\left(\mathbf{X}_{text}^{(l-1)} + MH\left(\mathbf{X}_{text}^{(l-1)}\right)\right), \quad (1)$$

$$\mathbf{X}_{text}^{(l)} = LN\left(\tilde{\mathbf{X}}_{text}^{(l)} + FFN\left(\tilde{\mathbf{X}}_{text}^{(l)}\right)\right), \quad (2)$$

where MH denotes the multi-head attention mechanism, FFN represents the feed-forward sub-layer, and LN indicates layer normalization.

Finally, the last layer's output from the BART encoder is taken as the context-aware representation for each token in the text sequence. The textual representation of each step is

denoted as $\{\mathbf{t}_i\}_{i=1}^{N}$, where $\mathbf{t}_i \in \mathbb{R}^{l_i \times d}$, l_i is the text length of the i-th step and d is the hidden dimension.

3.2.2 Step-Wise Video Encoding

The video encoding module extracts visual features corresponding to each step. For the video clip of each step, we uniformly sample T frames and extract frame-level visual representations using the pre-trained CLIP model [32]. The resulting visual features of the entire unordered video sequence are denoted as $\mathbf{X}_{video} \in \mathbb{R}^{N \times T \times h_{clip}}$, where h_{clip} denotes the dimensionality of the CLIP features. To align the visual features with the textual representations in a shared embedding space, we apply a projection layer to map each frame-level feature to the same hidden dimension d used by the text encoder. The final video representation for each step is denoted as $\{\mathbf{v}_i\}_{i=1}^{N}$, where $\mathbf{v}_i \in \mathbb{R}^{T \times d}$.

3.3. Interactive Cross-Modal Alignment and Fusion

To enable effective cross-modal interaction, we introduce an interactive multimodal alignment and fusion mechanism following the feature extraction stage, consisting of: (1) a visual-textual alignment module, and (2) a multimodal fusion module comprising multimodal cross-attention, token-level attention, and sequence-level attention layers to enable comprehensive multimodal integration at different granularities.

3.3.1 Visual-Textual Feature Alignment

In multimodal sequence ordering tasks, visual and textual features typically exhibit inherent heterogeneity. Direct fusion may lead to semantic inconsistencies and degraded performance when different modalities interact without proper alignment constraints. Motivated by previous works [26, 40], we introduce an explicit visual-textual alignment module before fusion to better model the alignment relationship between multimodal features, thereby enabling more effective fusion. The primary objective of this module is to bridge the semantic gap between visual and textual modalities. By explicitly aligning video and text features in a shared latent space, the module ensures that each video clip and its corresponding text description are mapped to a unified logical concept. Specifically, we employ a contrastive learning framework with a momentum encoder to align global representations across modalities. The training objective encourages semantically corresponding visual-textual pairs to be close in the shared representation space, providing a more stable and consistent multimodal representation for subsequent multimodal fusion.

We begin by prepending a modality-specific global $[CLS]$ token embedding to each individual feature representation: $\mathbf{v}_{\text{cls}}$ for visual features $\{\mathbf{v}_i\}_{i=1}^{N}$ and $\mathbf{t}_{\text{cls}}$ for textual features $\{\mathbf{t}_i\}_{i=1}^{N}$, respectively. Subsequently, these extended sequences are encoded through two modality-specific transformer encoders, each consisting of l_{mm} layers, to obtain context-enhanced representations:

$$\tilde{\mathbf{V}}_i = Transformer^{v}_{enc}([\mathbf{v}_{cls}; \mathbf{v}_i]), \tag{3}$$

$$\tilde{\mathbf{T}}_i = Transformer^{t}_{enc}([\mathbf{t}_{cls}; \mathbf{t}_i]). \tag{4}$$

We take the output representations at the $[CLS]$ token positions, $\tilde{\mathbf{V}}_i[0]$ and $\tilde{\mathbf{T}}_i[0]$, as the global representations of the visual and textual steps, respectively, and denote them as $\tilde{\mathbf{v}}$ and $\tilde{\mathbf{t}}$ for brevity.

Inspired by MoCo [17], we maintain two memory queues to store the most recent M visual-textual step features from momentum encoders for each modality. Let $\tilde{\mathbf{V}}^*$ and $\tilde{\mathbf{T}}^*$ represent the normalized $[CLS]$ features from the momentum encoders for the visual and textual modalities, respectively. To align the two modalities in the feature space, we adopt a cross-modal contrastive learning objective. Specifically, inter-modal similarity is computed using cosine similarity with temperature scaling:

$$sim(\mathbf{u}, \mathbf{r}) = \mathbf{u}^T \mathbf{r} / \sigma, \tag{5}$$

where σ is a temperature hyperparameter controlling the sharpness of the similarity distribution.

We then calculate the softmax-normalized similarities between current features and momentum encoder features in the memory queues:

$$\mathbf{S}^{V2T}_m = \frac{\exp\left(sim(\tilde{\mathbf{v}}, \tilde{\mathbf{T}}^*_m)\right)}{\sum_{m=1}^{M} \exp\left(sim(\tilde{\mathbf{v}}, \tilde{\mathbf{T}}^*_m)\right)} \tag{6}$$

$$\mathbf{S}^{T2V}_m = \frac{\exp\left(sim(\tilde{\mathbf{t}}, \tilde{\mathbf{V}}^*_m)\right)}{\sum_{m=1}^{M} \exp\left(sim(\tilde{\mathbf{t}}, \tilde{\mathbf{V}}^*_m)\right)}. \tag{7}$$

The video-text contrastive loss is defined as the average cross-entropy between the predicted similarity distributions and their corresponding ground-truth labels:

$$\mathcal{L}_{vtc} = \frac{1}{2}\mathbb{E}_{(I,T)\sim D}\left[CE(\mathbf{y}^{V2T}, \mathbf{S}^{V2T}) + CE(\mathbf{y}^{T2V}, \mathbf{S}^{T2V})\right], \tag{8}$$

where $CE(\cdot, \cdot)$ denotes the cross-entropy loss, $\mathbf{y}^{V2T}$ and $\mathbf{y}^{T2V}$ are one-hot ground-truth labels, assigning a probability of 1 to the positive pairs and 0 to all negative samples.

3.3.2 Hierarchical Multimodal Feature Fusion

To construct a unified multimodal representation for each step, we employ a transformer-based video-text fusion and hierarchical encoding module. Specifically, given the contextualized textual feature $\tilde{\mathbf{T}}_i$ and its corresponding visual

feature $\tilde{\mathbf{V}}_i$ for the i-th step, we apply a cross-attention mechanism to enable cross-modal interaction. The textual features serve as the queries, while the visual features serve as the keys and values:

$$CrossAttn(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax\left(\mathbf{Q}\mathbf{K}^T/\sqrt{d}\right)\mathbf{V}, \quad (9)$$

$$\mathbf{Q} = \tilde{\mathbf{T}}_i\mathbf{W}_{cq}, \quad \mathbf{K} = \tilde{\mathbf{V}}_i\mathbf{W}_{ck}, \quad \mathbf{V} = \tilde{\mathbf{V}}_i\mathbf{W}_{cv}, \quad (10)$$

where $\mathbf{W}_{cq}$, $\mathbf{W}_{ck}$, and $\mathbf{W}_{cv}$ are learnable parameter matrices, and $\sqrt{d}$ is the scaling factor based on the hidden dimension d. The resulting fused representation is denoted as $\mathbf{h}_i$, with the token representation denoted by $\mathbf{h}_i^k$, referring to the k-th token of the i-th step.

After cross-attention fusion, the multimodal features have incorporated information from the corresponding textual and visual modalities. To further construct a unified representation for each step, we introduce a token-level attention mechanism that enables the model to focus on the most informative components within a step. The aggregated representation for the i-th step, denoted as $\mathbf{x}_i^0$, can be computed as follows:

$$\mathbf{x}_i^0 = \sum_{k=1}^{l_i} \alpha_i^k \mathbf{h}_i^k, \quad (11)$$

where l_i denotes the number of tokens in the i-th sentence, and α_i^k is the attention weight assigned for the k-th token. The weight is calculated as follows:

$$\alpha_i^k = \frac{\exp\left(\mathbf{v}_h^\top \tanh\left(\mathbf{W}_h\mathbf{h}_i^k\right)\right)}{\sum_{k=1}^{l_i}\exp\left(\mathbf{v}_h^\top \tanh\left(\mathbf{W}_h\mathbf{h}_i^k\right)\right)}, \quad (12)$$

where $\mathbf{W}_h$ and $\mathbf{v}_h$ are learnable parameters.

Building upon the token-level attention layer, we further feed the step representations into a sequence-level transformer encoder composed of l_{seq} layers to model contextual dependencies across different steps. This produces the final step-level representations $\{\mathbf{x}_i\}_{i=1}^N$, with the entire sequence denoted as $\mathbf{X}$. This hierarchical modeling strategy not only captures local, fine-grained features but also global dependencies, thereby enhancing the model's capacity to understand and reason over multimodal sequences.

3.4. Pointer-based Ordering Decoder

After the multimodal feature alignment and fusion module, we employ a pointer network-based ordering decoder to generate the sequence order. Specifically, we utilize a transformer decoder, which consists of l_{dec} layers and generates the output representations autoregressively. At the p-th timestep, the computation of the l-th decoder layer is defined as follows:

$$\tilde{\mathbf{d}}_p^{(l)} = LN\left(\mathbf{d}_p^{(l-1)} + MMH\left(\mathbf{D}_p^{(l-1)}\right)\right), \quad (13)$$

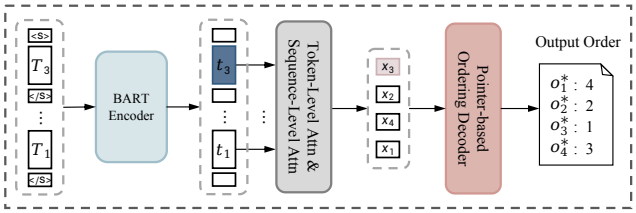

Figure 3. Model architecture for task-specific pre-training, where the paragraph-level BART encoder, token-level and sequence-level attention modules, and the pointer network-based ordering decoder are trained on large-scale text corpora using the ordering loss to learn ordering-related patterns.

$$\bar{\mathbf{d}}_p^{(l)} = LN\left(\tilde{\mathbf{d}}_p^{(l)} + MH\left(\tilde{\mathbf{d}}_p^{(l)}, \mathbf{X}, \mathbf{X}\right)\right), \quad (14)$$

$$\mathbf{d}_p^{(l)} = LN\left(\bar{\mathbf{d}}_p^{(l)} + FFN\left(\bar{\mathbf{d}}_p^{(l)}\right)\right), \quad (15)$$

where MMH denotes the masked multi-head attention mechanism, and $\mathbf{D}_p^{(l-1)}$ represents the input representations of all previously generated positions up to step $p-1$. We denote the decoder's final layer output as $\mathbf{d}_p$. For each candidate step T_{o_i}, we compute a final score representation $\mathbf{z}_{o_i}$ by combining the decoder state and the candidate's feature representation:

$$\mathbf{z}_{o_i} = \mathbf{g}^T \tanh\left(\mathbf{W}_d\mathbf{d}_p + \mathbf{W}_x\mathbf{x}_{o_i}\right), \quad (16)$$

where $\mathbf{W}_d$, $\mathbf{W}_x$, and $\mathbf{g}$ are learnable parameters. Then we select the step with the highest probability among the unordered candidate steps as the next output:

$$P\left(o'_k \mid o'_{k-1}, \cdots, o'_1, S\right) = Softmax\left(\mathbf{z}_{o_1}, \cdots, \mathbf{z}_{o_N}\right), \quad (17)$$

where previously selected steps are masked during inference to prevent duplicate predictions.

3.5. Model Training

3.5.1 Task-Specific Pre-Training

To further enhance the model's capability for sequential understanding and reasoning, we introduce a task-specific pre-training stage. In contrast to the subsequent second-stage video-text ordering fine-tuning, the pre-training stage relies solely on textual data. Considering the scarcity of large-scale multimodal pre-training resources, this design not only reduces data requirements but also enables the model to learn task-specific reasoning abilities from abundant textual corpora. In this stage, the model acquires task-specific paragraph-level modeling and sequential prediction abilities, which serve as a strong foundation for subsequent cross-modal learning. Specifically, we utilize a publicly available large-scale Wikipedia paragraph dataset [21] containing 21 million paragraphs with naturally coherent contextual structures, making it well-suited for training coherence-aware models.

Table 1. Statistical information of the WikiHVO dataset. The average video length and text length are measured in seconds and word count, respectively.

Dataset	Training	Validation	Test
Total Video Clips	14573	1818	1863
Total Tasks	2924	365	365
Average Steps	5.0	5.0	5.1
Average Video Length	10.7	10.5	10.8
Average Text Length	66.7	65.1	65.8

Figure 3 illustrates the model architecture during the pre-training process. The visual feature-related components are not trained, only textual features are incorporated as input in this setting. The paragraph-level text encoder, the token-level and sequence-level attention modules, and the pointer-based ordering decoder are trained for sequence prediction. The sequence ordering loss is adopted in this training phase:

$$\mathcal{L}_{pre} = -\frac{1}{|Q_{pre}|} \sum_{(x_{pre}, o^*_{pre}) \in Q_{pre}} \log P(o^*_{pre} | x_{pre}; \theta_{pre}), \tag{18}$$

where $Q_{pre} = \{(x_{pre}, o^*_{pre})\}$ is the general pre-training dataset, and θ_{pre} represents the model parameters activated during pre-training. We employ the negative log-likelihood between the predicted output distribution and the ground-truth labels at each step as the supervision objective. Through task-specific pre-training, the model acquires preliminary knowledge of textual sequence structure, semantic coherence, and step-by-step reasoning patterns, which are essential for downstream multimodal ordering tasks.

3.5.2 Fine-Tuning

During the multimodal fine-tuning stage, given the training set $Q = \{(x, o^*)\}$, we adopt a joint optimization objective defined as follows:

$$\mathcal{L} = \lambda \mathcal{L}_{ord} + (1 - \lambda) \mathcal{L}_{vtc}, \tag{19}$$

$$\mathcal{L}_{ord} = -\frac{1}{|Q|} \sum_{(x, o^*) \in Q} \log P(o^* | x; \theta), \tag{20}$$

where $\mathcal{L}_{ord}$ denotes the ordering loss, and $\mathcal{L}_{vtc}$ is the visual-text contrastive loss introduced in the alignment module to improve the consistency of multimodal representations. The parameter set θ includes all trainable parameters of the network, and λ is a hyperparameter controlling the trade-off between the ordering objective and the alignment objective.

By jointly optimizing the ordering and alignment losses, the model benefits from both global supervision for sequence reasoning and local alignment guidance for cross-modal feature synergy and fusion. This joint objective enhances the model's overall capacity for robust multimodal understanding and reasoning.

4. Experiments

In this section, we conduct extensive experiments to systematically evaluate the effectiveness of VOTA on multimodal sequence ordering tasks.

4.1. Datasets

We conduct experiments on two datasets: (1) WikiHVO, our newly constructed video-text ordering dataset, and (2) RecipeQA [39], a publicly available image-text ordering dataset. The results obtained across these diverse benchmarks, featuring different modality combinations demonstrate the strong generalizability and robustness of our proposed method.

WikiHVO. To support research in video sequence ordering, we develop a high-quality and well-structured dataset, WikiHVO. It is publicly available and specifically designed for the English video ordering task, providing a valuable benchmark for advancing research in related fields.

We focus on highly operational and procedural domains, such as handicrafts, cooking, and repairs, collecting task-oriented data with clearly defined multi-step structures from the widely used online knowledge-sharing platform WikiHow. The content is community-contributed and subject to editorial review, ensuring high-quality and reliable procedural information. Each data instance in the dataset represents a structured procedural task, segmented into multiple steps. Each step contains a video clip (typically 5-20 seconds) and a corresponding textual description that accurately summarizes the core action and objective of the step.

To further guarantee data consistency and usability, we implement a standardized filtering and preprocessing pipeline, including: (1) data integrity verification, where samples missing either video clips or textual descriptions are removed to ensure each instance contains complete multimodal information; and (2) task validity verification, where tasks with fewer than two steps are excluded to ensure that each sample provides a meaningful temporal structure for both training and evaluation.

The resulting dataset, named WikiHVO, comprises a collection of multi-step video-text sequences. It contains a total of 18,253 video clips across 3,654 distinct procedural tasks. Dataset statistics are summarized in Table 1. The dataset is randomly split into training, validation, and test subsets, using an 8:1:1 ratio. For evaluation purposes, the samples are categorized into two subsets: (1) WikiHVO-short, which contains short sequences with three or fewer steps; and (2) WikiHVO-long, which includes long sequences with more than three steps. In total, the dataset includes 2,638 long sequences and 1,016 short sequences.

To better illustrate the diversity and inherent procedural structure of WikiHVO, we provide sample sequences in Figure 4. The figure showcases examples from multiple categories, including cooking (e.g., making a tuna sandwich),

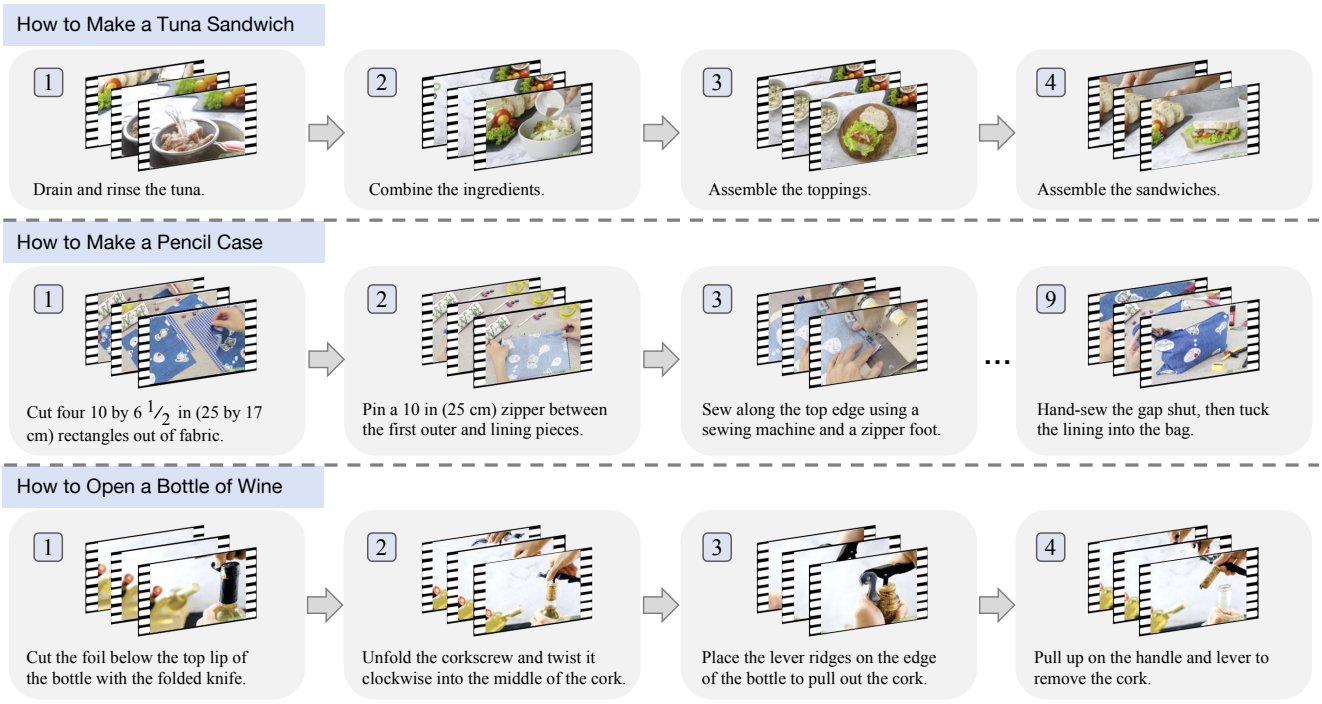

Figure 4. Illustrative examples from the WikiHVO dataset.

crafts (e.g., making a pencil case), and skills (e.g., opening a wine bottle), covering a wide range of real-world scenarios, highlighting the multi-step procedural nature of each task.

Overall, the WikiHVO dataset serves as a novel and valuable benchmark for video sequence ordering and multimodal reasoning. It fills a critical gap in current research by providing structured data specifically designed for reasoning and coherence modeling. This contribution facilitates further advancements in multimodal sequence modeling, cross-modal comprehension, and reasoning.

RecipeQA. In addition to the video sequence ordering task, we evaluate VOTA on the RecipeQA dataset. It is a multimodal dataset focused on the cooking domain, where each instance consists of an ordered set of textual descriptions paired with corresponding images. It has been widely adopted in research on multimodal reasoning and sequence ordering. Following the standard setup in prior work [38, 42], we use the original data splits: the training, validation, and test sets contain 8,032, 973, and 100 samples.

4.2. Evaluation Metrics

To comprehensively evaluate the performance of VOTA on multimodal sequence ordering tasks, we adopt three widely used evaluation metrics following prior studies [10, 41, 42]: Accuracy (Acc), Perfect Matching Ratio (PMR), and Kendall's Tau (τ) coefficient.

Accuracy measures the proportion of correctly ordered steps within each sequence. The overall accuracy is computed as the average accuracy across all sequences:

$$Accuracy = \frac{1}{N_{samples}} \sum_{i=1}^{N_{samples}} \left(\frac{1}{n_i} \sum_{k=1}^{n_i} \mathbb{I}(o_k'^i = o_k^{*i}) \right), \tag{21}$$

where $N_{samples}$ is the total number of samples, n_i is the number of steps in the i-th sample, $o_k'^i$ and o_k^{*i} denote the predicted and ground-truth indices of the k-th step, respectively.

Perfect Match Ratio (PMR) evaluates whether the entire predicted sequence exactly matches the ground truth sequence. It is defined as:

$$PMR = \frac{1}{N_{samples}} \sum_{i=1}^{N_{samples}} \mathbb{I}(o'^i = o^{*i}), \tag{22}$$

where o'^i and o^{*i} denote the predicted and ground-truth orderings of the i-th sample, respectively.

Kendall's Tau (τ) quantifies the ordinal correlation by measuring the proportion of discordant pairs:

$$\tau = 1 - \frac{2}{N_{samples}} \sum_{i=1}^{N_{samples}} \left(\frac{\#inversions(o'^i, o^{*i})}{n_i(n_i-1)/2} \right), \tag{23}$$

where $\#inversions(o'^i, o^{*i})$ denotes the number of inverted step pairs in the predicted sequence compared to the ground truth for the i-th sample, and n_i is the number of steps in that sequence.

Table 2. Experimental results on the WikiHVO dataset.

Modality	Models	WikiHVO-short			WikiHVO-long			WikiHVO-all		
		Acc	PMR	τ	Acc	PMR	τ	Acc	PMR	τ
Text-Only	B-TSort	63.39	57.35	0.50	48.90	20.69	0.50	51.92	29.00	0.51
	BERSON	73.77	63.93	0.61	61.96	42.09	0.69	64.09	46.02	0.67
	RE-BART	78.08	71.23	0.71	63.83	40.79	0.69	66.80	47.14	0.69
	BHAONet	78.54	73.97	0.73	73.22	56.83	0.77	74.33	60.40	0.76
Image-Text	NACON	73.52	65.75	0.64	56.70	32.37	0.64	60.20	39.32	0.64
	BERSON+CLIP-ViL	77.62	71.23	0.73	60.91	39.13	0.70	64.41	45.85	0.70
	MHAONet	78.54	72.60	0.69	64.92	38.46	0.74	67.91	45.95	0.73
	VOTA-image (ours)	80.37	75.34	0.76	73.80	56.12	**0.78**	75.16	60.11	0.78
Video-Text	SVO	70.78	68.97	0.68	55.49	34.89	0.60	58.67	43.01	0.62
	MHAONet-video	83.11	80.46	0.79	65.47	37.69	0.73	69.34	48.41	0.75
	VOTA (ours)	**86.76**	**82.19**	**0.85**	**74.42**	**57.55**	0.77	**76.98**	**62.68**	**0.79**

4.3. Implementation Details

The experiments are conducted using the PyTorch framework on NVIDIA RTX A6000 GPUs. For video feature extraction, we uniformly sample 16 frames from each video clip and employ the pre-trained CLIP model [32] as the visual encoder to extract frame-level features. The parameters of the CLIP encoder are frozen throughout the training process. We first perform a task-specific pre-training on the wiki_dpr dataset [21] for 60,000 steps using the BART-large encoder[1] as the backbone, with a batch size of 16. The resulting pre-trained model weights are subsequently used to initialize the model parameters for fine-tuning. Optimization is performed with the AdamW optimizer combined with a linear learning rate decay scheduler.

For the hyperparameter setting, we adopt a grid search strategy and select the best configuration based on the performance of the validation set. Specifically, the model is fine-tuned for 10 epochs on the WikiHVO dataset and 5 on the RecipeQA dataset [39]. The learning rates for different components are set as follows: for WikiHVO, the learning rates for the feature extraction module, the alignment and fusion module, and the ordering module are set to 1e-5, 5e-5, and 1e-5, respectively; On RecipeQA, the corresponding learning rates are set to 5e-6, 1e-5, and 5e-6, respectively. The hidden dimension is uniformly set to 1024, l_{mm}, l_{seq}, and l_{dec} are set to 2, 2, and 1, respectively. For visual-text contrastive learning, the queue size M is set to 65,536. To ensure fairness and reproducibility, all experiments are conducted with identical random seeds.

VOTA comprises 342.85 million trainable parameters, primarily from the BART-large encoder and transformer fusion modules, while the CLIP visual encoder remains frozen. By leveraging a lightweight pointer-based decoder, VOTA achieves a superior balance between computational economy and ordering accuracy.

[1]https://huggingface.co/facebook/bart-large

4.4. Baselines

To evaluate the effectiveness of our proposed model, we compare it with state-of-the-art sequence ordering approaches under both unimodal and multimodal settings.

Under the text-only setting, we consider four representative baselines: B-TSort [31], BERSON [10], RE-BART [8], and BHAONet [41]. Among them, B-TSort and BERSON adopt the pairwise ordering strategy, while RE-BART and BHAONet utilize sequence-level input representations. Notably, BHAONet can be viewed as a lightweight variant of our model when constrained to textual input only.

Under the multimodal setting, for the image-text configuration, we extract the middle frame from each video as the visual input. We compare VOTA with several strong baselines, including NACON [6], BERSON+CLIP-ViL [38], and MHAONet [42]. Additionally, we report the results of VOTA-image, a variant of our approach that replaces the video features with patch-level image features from the middle frame. For the video-text configuration, we compare our approach against MHAONet-video, an extended version of MHAONet that incorporates video sequence features, and SVO [14], a model specifically designed for the short video ordering task.

4.5. Main Results

We evaluate VOTA on both WikiHVO and RecipeQA to assess its effectiveness across different modality configurations and domains. WikiHVO focuses on video-text procedural ordering in diverse real-world scenarios, while RecipeQA provides an established benchmark for image-text ordering. These two benchmarks jointly cover different sequence lengths, modalities, and structural complexities, offering a comprehensive evaluation of multimodal ordering ability.

Table 3. Experimental results on the RecipeQA dataset.

Models	RecipeQA		
	Acc	PMR	τ
BERSON	74.4	52.0	0.83
RE-BART	83.0	64.0	0.88
BHAONet	83.2	66.0	0.89
NACON	76.8	64.0	0.83
BERSON+CLIP-ViL	82.6	68.0	0.88
MHAONet	85.2	72.0	0.90
VOTA-image (ours)	**86.6**	**73.0**	**0.91**
Human Performance	92.12	83.1	0.95

4.5.1 Results on the WikiHVO Dataset

For the WikiHVO dataset, we evaluate the models across different data splits: WikiHVO-short, WikiHVO-long, and WikiHVO-all, covering both unimodal and multimodal configurations. The experimental results are in Table 2. As shown, VOTA consistently outperforms all baseline models across different settings, demonstrating superior multimodal modeling and sequence ordering capabilities.

Under the text-only setting, B-TSort and BERSON focus on pairwise relationships between sentences, which results in inferior performance compared to sequence-level encoding methods. RE-BART exhibits strong performance, and BHAONet achieves further improvements. In particular, sequence-level encoding models generally outperform pairwise encoding approaches, due to their enhanced ability to capture global contextual dependencies throughout the sequence. Moreover, the hierarchical encoding design of BHAONet offers greater stability, particularly when processing longer sequences.

Under the image-text setting, VOTA shows consistent improvements over all baseline models. Compared to the step-level encoding model NACON and the pairwise encoding model BERSON+CLIP-ViL, our approach demonstrates a stronger ability to capture contextual dependencies. Compared to the current state-of-the-art image-text ordering method MHAONet, the variant VOTA-image achieves a 7.25% improvement in accuracy on the WikiHVO-all dataset. Moreover, VOTA not only maintains high accuracy but also demonstrates superior relative positional consistency, as measured by Kendall's Tau (τ).

Under the video-text setting, VOTA also achieves substantial improvements. Specifically, compared to MHAONet-video, VOTA improves accuracy by 3.65% on WikiHVO-short and 8.95% on WikiHVO-long, while yielding PMR gains of 1.73% and 19.86%, respectively. These improvements can be attributed to our task-specific pre-training and explicit cross-modal alignment supervision, which jointly enhance the model's ability in both comprehension and ordering, particularly for longer sequences containing more steps.

These results highlight the superiority of our proposed model in effectively capturing the temporal and logical relationships within multimodal sequences, thereby enabling more accurate and coherent ordering.

4.5.2 Results on the RecipeQA Dataset

We further evaluate the performance of VOTA on the RecipeQA dataset [39]. The results are presented in Table 3. As shown, VOTA-image outperforms all existing multimodal approaches across the evaluation metrics. Additionally, BHAONet surpasses RE-BART and BERSON when relying solely on textual data, validating the effectiveness in purely textual contexts. Compared to the current strongest baseline, MHAONet, VOTA-image achieves 1.4%, 1%, and 1% absolute improvements in accuracy, PMR, and τ, respectively, demonstrating capabilities that approach human-level performance. The experiments indicate that VOTA not only excels at understanding and ordering video sequences but also demonstrates robust generalization on image-text data. By leveraging task-specific pre-training and explicit multimodal alignment supervision, our approach enhances the modeling of coherence and logical relationships.

Overall, the experimental results validate the superior performance and strong generalization ability of VOTA across different modality-specific ordering tasks.

4.6. Analysis

To further evaluate the effectiveness of each component in VOTA, we conduct comprehensive experiments on the WikiHVO dataset. The validation includes ablation studies, hyperparameter analyses, and visualization of the learned multimodal representations.

4.6.1 Ablation Study

To examine the contribution of individual modules, we perform ablation studies on the WikiHVO dataset. The experimental results are presented in Table 4. The results show that the full model achieves substantial performance improvements compared to its variants.

In particular, the pre-training module plays a critical role in enhancing overall performance. When the pre-training stage is removed, the model exhibits consistent performance degradation across all evaluation metrics. This highlights the importance of the pre-training phase in providing effective model initialization and task-specific adaptation. It significantly improves the model's ability to capture semantic and structural information, thereby facilitating more accurate understanding and reasoning in sequence ordering tasks. Moreover, removing either the visual modality (w/o visual) or the textual modality (w/o text) results in performance degradation, indicating that both visual and linguistic information are essential for understanding and ordering

Table 4. Results of ablation experiment on WikiHVO dataset.

Models	WikiHVO-short			WikiHVO-long			WikiHVO-all		
	Acc	PMR	τ	Acc	PMR	τ	Acc	PMR	τ
VOTA	**86.76**	**82.19**	**0.85**	**74.42**	**57.55**	0.77	**76.98**	**62.68**	**0.79**
w/o pre-train	82.65	76.71	0.79	71.40	51.08	0.77	73.74	56.41	0.77
w/o visual	78.54	73.97	0.74	73.22	56.83	0.77	74.33	60.40	0.76
w/o text	63.47	54.79	0.43	31.17	7.19	0.28	37.89	17.09	0.31
w/o cross-modal fusion	82.65	78.08	0.78	72.49	56.47	0.76	74.60	60.97	0.77
w/o alignment	81.74	76.71	0.76	73.89	56.12	**0.78**	75.52	60.40	0.77

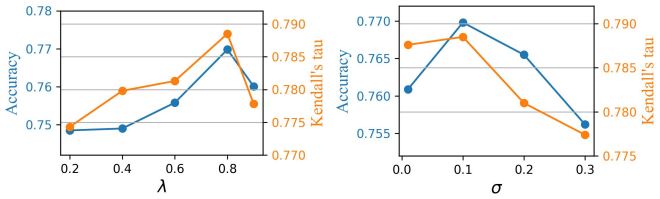

Figure 5. Hyperparameter study to analyze the effects of λ and σ.

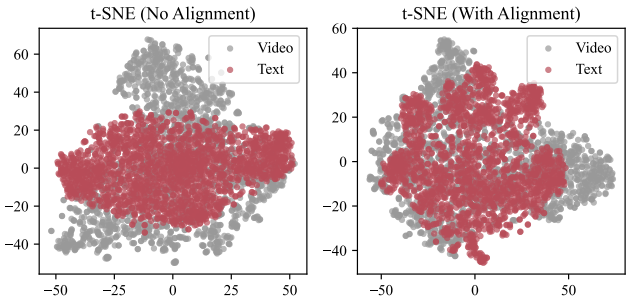

Figure 6. T-SNE visualization of learned visual-textual features.

multi-step sequences. Notably, the removal of the textual modality (w/o text) leads to a substantial performance drop, highlighting the critical role of textual information in capturing the operational objectives of each step and the logical dependencies across the sequence. In addition, removing either the cross-modal fusion module (w/o cross-modal fusion) or the alignment module (w/o alignment) also leads to performance declines, suggesting that both cross-modal fusion and alignment are essential for effective multimodal integration and reasoning.

In summary, the results of the ablation study further validate the importance of each key component in our model architecture. By jointly optimizing the ordering and cross-modal alignment objectives, VOTA demonstrates significant advantages in multimodal ordering tasks.

4.6.2 Hyperparameter Study

We conduct a hyperparameter study to evaluate the impact of the weighting factor λ in the loss function and the temperature parameter σ in cross-modal contrastive learning on model performance.

During training, the hyperparameter λ controls the trade-off between the ordering loss and the alignment loss. We evaluate model performance under varying λ values using Accuracy and Kendall's tau metrics as illustrated in Figure 5. When λ is small, the model predominantly focuses on the alignment objective while neglecting the learning of ordering. As λ increases, the ordering loss gains greater influence, and the model achieves optimal performance across the evaluation metrics at $\lambda = 0.8$, suggesting a well-balanced trade-off between the ordering and alignment objectives. However, further increasing λ introduces an imbalance where the learning of visual-text alignment is neglected, and degrades ordering accuracy.

In addition, we conduct a detailed investigation of the temperature parameter σ, which is essential for regulating the sharpness of the similarity distribution in cross-modal contrastive learning. Lower σ values enhance the model's sensitivity to distinguishing between positive and negative sample pairs, while higher values yield smoother similarity distributions. Model performance initially improves as σ increases, reaching its peak at $\sigma = 0.1$, after which it begins to decline. While lower temperatures enable finer discrimination among samples, excessively small values can lead to overfitting and reduced generalization. At $\sigma = 0.1$, the model achieves a favorable balance, demonstrating improved adaptability to diverse data samples while maintaining strong ordering performance.

In summary, selecting appropriate hyperparameters is essential for achieving an optimal balance between discriminative and generalization capability, thereby improving both the accuracy and robustness of the model.

4.6.3 Multimodal Alignment Visualization

To intuitively illustrate the impact of the contrastive alignment strategy on multimodal ordering tasks, we employ the t-SNE (t-distributed Stochastic Neighbor Embedding) algorithm to perform dimensionality reduction and visualize the visual and textual features learned by the model.

In this experiment, we extract the visual and textual features before the fusion module and project them into

a two-dimensional space for visualization. In multimodal sequence ordering, the ideal outcome is that corresponding video and text features remain close in the semantic space, thereby facilitating effective multimodal comprehension and fusion. The visualization results are presented in Figure 6. The left subfigure displays the feature distributions without the proposed alignment mechanism, while the right subfigure clearly demonstrates the distributions with the alignment strategy applied.

It can be observed that the model trained with the alignment objective exhibits a clearer alignment pattern, with greater overlap between visual and textual feature representations. This indicates that the proposed alignment strategy effectively reduces the modality gap, enabling the model to learn more semantically aligned multimodal representations. By minimizing this representational discrepancy, the alignment mechanism lays a stronger foundation for subsequent feature fusion. As the model progressively captures inter-modal correlations, the visual and textual modalities can mutually reinforce one another, thereby improving the model's ability to comprehend the sequence and reason about the logical relationships among sequential steps.

4.6.4 Limitations

Despite its effectiveness, VOTA entails several limitations that require further investigation. First, while the WikiHVO dataset spans diverse procedural domains, it primarily consists of curated instructional videos with well-defined temporal boundaries. Consequently, VOTA's efficacy in unconstrained video environments, where transitions are often fluid or ambiguous, remains to be fully established. Second, the model's scalability is bottlenecked by the quadratic complexity of the self-attention mechanism, and we observed a performance degradation as sequence lengths increased. Future research could alleviate this by integrating linear-complexity attention or hierarchical modeling frameworks. Lastly, extending the evaluation to more diverse, unstructured datasets is necessary to improve and verify the model's generalizability across broader domains.

5. Conclusion

In this paper, we propose VOTA, a two-stage modeling framework for multimodal sequence ordering. In particular, we construct a high-quality, multi-domain video ordering dataset that covers a wide range of real-world operational scenarios, providing a novel and valuable benchmark for studying multimodal sequence ordering. Building upon this, we introduce a two-stage training strategy that effectively enhances ordering performance through task-specific pre-training and alignment-guided fine-tuning. Experimental results demonstrate that VOTA achieves state-of-the-art performance in reconstructing coherent sequences from multimodal content, showcasing its superior capability in both multimodal comprehension and coherence reasoning.

Acknowledgement

This work was supported in part by National Key R&D Program of China (No.2023YFE0204200) and the Key R&D Program of Zhejiang Province (No.2023C01043).

References

[1] R. Barzilay and N. Elhadad. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55, 2002. 1

[2] R. Barzilay, N. Elhadad, and K. McKeown. Sentence ordering in multidocument summarization. In *Proceedings of the first international conference on Human language technology research*, 2001. 1

[3] R. Barzilay and M. Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008. 1, 2

[4] R. Barzilay and L. Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 113–120, 2004. 1, 2

[5] Y. Bin, J. Liao, Y. Ding, H. Li, Y. Yang, S.-K. Ng, and H. T. Shen. Leveraging weak cross-modal guidance for coherence modelling via iterative learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4630–4639, 2024. 2

[6] Y. Bin, W. Shi, J. Zhang, Y. Ding, Y. Yang, and H. T. Shen. Non-autoregressive cross-modal coherence modelling. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3253–3261, 2022. 2, 3, 9

[7] X. Chen, X. Qiu, and X. Huang. Neural sentence ordering. *arXiv preprint arXiv:1607.06952*, 2016. 2

[8] S. B. R. Chowdhury, F. Brahman, and S. Chaturvedi. Is everything in order? a simple way to order sentences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10769–10779, 2021. 1, 3, 9

[9] B. Cui, Y. Li, M. Chen, and Z. Zhang. Deep attentive sentence ordering network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4340–4349, 2018. 1

[10] B. Cui, Y. Li, and Z. Zhang. Bert-enhanced relational sentence ordering network. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6310–6320, 2020. 1, 2, 3, 8, 9

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. 1, 3

[12] N. Dvornik, I. Hadji, R. Zhang, K. G. Derpanis, R. P. Wildes, and A. D. Jepson. Stepformer: Self-supervised step discovery and localization in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18952–18961, 2023. 2

[13] L. Garattoni and M. Birattari. Autonomous task sequencing in a robot swarm. *Science Robotics*, 3(20):eaat0430, 2018. 1

[14] S. Ge, Q. Chen, Z. Jiang, Y. Yin, Z. Chen, and Q. Gu. Short video ordering via position decoding and successor prediction. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2167–2176, 2024. 2, 3, 9

[15] J. Gong, X. Chen, X. Qiu, and X. Huang. End-to-end neural sentence ordering using pointer network. *arXiv preprint arXiv:1611.04953*, 2016. 1

[16] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, and D. Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):9052–9071, 2024. 2

[17] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 5

[18] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[19] T.-H. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239, 2016. 2

[20] T. Jiao, C. Guo, X. Feng, Y. Chen, and J. Song. A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. *Computers, Materials & Continua*, 80(1), 2024. 2

[21] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020. 6, 9

[22] S. Lai, A. Wang, F. Meng, J. Zhou, Y. Ge, J. Zeng, J. Yao, D. Huang, and J. Su. Improving graph-based sentence ordering with iteratively predicted pairwise orderings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2407–2417, 2021. 3

[23] M. Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 545–552, 2003. 1

[24] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020. 1, 3, 4

[25] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[26] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 5

[27] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020. 2, 3

[28] Y. Li, B. Cui, and Z. Zhang. Efficient relational sentence ordering network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6169–6183, 2021. 1

[29] H. C. Moon, T. Mohiuddin, S. Joty, and X. Chi. A unified neural coherence model. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 2262–2272. Association for Computational Linguistics, 2019. 1

[30] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, 2016. 2

[31] S. Prabhumoye, R. Salakhutdinov, and A. W. Black. Topological sort for sentence ordering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2783–2792, 2020. 1, 9

[32] 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*, pages 8748–8763. PmLR, 2021. 2, 5, 9

[33] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3

[34] Y. Shi, H. Zhang, N. Li, and T. Yang. An overview of sentence ordering task. *International Journal of Data Science and Analytics*, 18(1):1–18, 2024. 2

[35] N. Tandon, K. Sakaguchi, B. Dalvi, D. Rajagopal, P. Clark, M. Guerquin, K. Richardson, and E. Hovy. A dataset for tracking entities in open domain procedural text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6408–6417, 2020. 1

[36] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. *Advances in neural information processing systems*, 28, 2015. 1, 3

[37] G. Wang, F. Lin, T. Wu, Z. Liu, Z. Ba, and K. Ren. Fsfm: A generalizable face security foundation model via self-supervised facial representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24364–24376, 2025. 2

[38] T.-L. Wu, A. Spangher, P. Alipoormolabashi, M. Freedman, R. Weischedel, and N. Peng. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4525–4542, 2022. 1, 2, 3, 8, 9

[39] S. Yagcioglu, A. Erdem, E. Erdem, and N. Ikizler-Cinbis. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, 2018. 7, 9, 10

[40] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. M. Chilimbi, and J. Huang. Vision-language pre-training with triple contrastive learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15650–15659, 2022. 2, 5

[41] Y. Yang, B. Cui, and Y. Li. Bart-based hierarchical attentional network for sentence ordering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4203–4207, 2024. 1, 2, 3, 8, 9

[42] Y. Yang, B. Cui, and Y. Li. A multimodal hierarchical attentional ordering network. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025. 2, 3, 8, 9

[43] Y. Yin, F. Meng, J. Su, Y. Ge, L. Song, J. Zhou, and J. Luo. Enhancing pointer network for sentence ordering with pairwise ordering predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9482–9489, 2020. 1, 3

[44] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi. Merlot: Multimodal neural script knowledge models. *Advances in neural information processing systems*, 34:23634–23651, 2021. 2, 3

[45] J. Zhu, H. Li, T. Liu, Y. Zhou, J. Zhang, and C. Zong. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164, 2018. 1