

# Order-Prompted Tag Sequence Generation for Video Tagging

Zongyang Ma<sup>1,2,4\*</sup>, Ziqi Zhang<sup>1\*</sup>, Yuxin Chen<sup>1,2,4\*</sup>, Zhongang Qi<sup>2</sup>, Yingmin Luo<sup>2</sup>, Zekun Li<sup>6</sup>,  
Chunfeng Yuan<sup>1</sup>, Bing Li<sup>†</sup>, Xiaohu Qie<sup>3</sup>, Ying Shan<sup>2</sup>, Weiming Hu<sup>1,4,5</sup>

<sup>1</sup>State Key Laboratory of Multimodal Artificial Intelligence Systems,

Institute of Automation, Chinese Academy of Sciences <sup>2</sup>ARC Lab, <sup>3</sup>Tencent PCG

<sup>4</sup>School of Artificial Intelligence, University of Chinese Academy of Sciences

<sup>5</sup>CAS Center for Excellence in Brain Science and Intelligence Technology <sup>6</sup>CNCERT/CC

{mazongyang2020, chen yuxin2019}@ia.ac.cn, {ziqi.zhang, cfyuan, bli, wmhu}@nlpr.ia.ac.cn

{zhongangqi, yingminluo, tigerqie, yingsshan}@tencent.com, lzk@cert.org.cn

## Abstract

Video Tagging intends to infer multiple tags spanning relevant content for a given video. Typically, video tags are freely defined and uploaded by a variety of users, so they have two characteristics: abundant in quantity and disordered intra-video. It is difficult for the existing multi-label classification and generation methods to adapt directly to this task. This paper proposes a novel generative model, *Order-Prompted Tag Sequence Generation (OP-TSG)*, according to the above characteristics. It regards video tagging as a tag sequence generation problem guided by sample-dependent order prompts. These prompts are semantically aligned with tags and enable to decouple tag generation order, making the model focus on modeling the tag dependencies. Moreover, the word-based generation strategy enables the model to generate novel tags. To verify the effectiveness and generalization of the proposed method, a Chinese video tagging benchmark *CREATE-tagging*, and an English image tagging benchmark *Pexel-tagging* are established. Extensive results show that *OP-TSG* is significantly superior to other methods, especially the results on rare tags improve by 3.3% and 3% over *SOTA* methods on *CREATE-tagging* and *Pexel-tagging*, and novel tags generated on *CREATE-tagging* exhibit a tag gain of 7.04%.

## 1. Introduction

Video tags are a series of discrete descriptive text in a free form, usually freely defined and uploaded by video platform users, to represent the specific content of the video. Short video platforms have a large number of videos with no tags or low-quality tags. Exploring automatic video tagging technology can effectively serve practical industrial requirements such as video recommendation, retrieval, and content review, and significantly reduce labor costs.

\* Equal contribution. † Corresponding author.

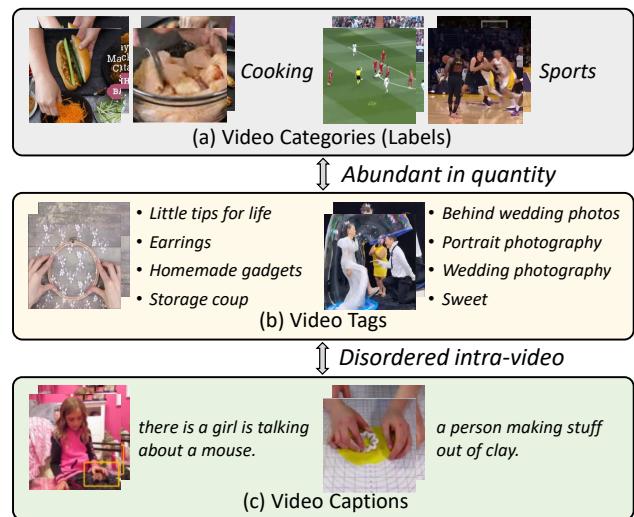


Figure 1. Comparison of video categories, tags, and captions. Video tags are more abundant in quantity than categories and are intra-video disordered compared to video captions.

Video tags have the nature of being abundant in quantity and disordered intra-video compared with video categories (labels)<sup>1</sup> and video captions, respectively, as shown in Figure 1. Compared to the fixed number of video categories strictly defined by experts, the abundance of user-defined video tags is primarily reflected in the following two aspects: (1) Multiple perspectives for the same video, such as entities, attributes, scenes, or styles; (2) Distinct granularities for the same content, such as separate words or more expressive phrases. As a result, large collections of tags can easily reach tens of thousands or even hundreds of thousands of magnitude in a large-scale scenario, presenting an extreme long-tail distribution. Compared to video captions

<sup>1</sup>Note that video categories and labels are generally represented in the form of indexes, and the number is fixed and small. This paper does not distinguish between them.

that consider grammatical correctness and fluency, multiple video tags have no fixed order within the same video, although they are correlated with each other.

The above characteristics of video tags make it difficult to directly apply current video multi-label classification models [4, 37, 6, 20] and generation models [24, 10, 34] to video tagging task. On the one hand, multi-label classification methods face a serious long-tail problem and also need to construct a classification head consistent with the pre-defined tag set, which will introduce a large number of parameters and cannot be replicated when the tag set changes. On the other hand, although the autoregressive-based generation methods can avoid bloated classification heads through word-by-word generation, the feature of tag disorder will plague the decoder of sequential generation and thus reduce the generation quality. Specifically: (1) Rule-based tag orders (e.g., frequency-first) are suboptimal; (2) Randomly ordering tags at each sampling without prompts puts the model in the dilemma of the same vision input with different tag sequences.

To this end, we propose a novel generative model, OP-TSG, which regards video tagging as a sequence generation problem guided by prompts and equips with a word-based generation strategy. OP-TSG includes a Video-Title Multimodal Hybrid Encoding module and an Order-prompted Tag Sequence Decoding module. Specifically, the encoding module integrates visual and textual multimodal information into a unified video representation. The decoding module consists of two components: the prompt encoder and the order-aware tag decoder. The prompt encoder takes a fixed number of learnable quires as input and interacts with the video representation to generate a series of sample-dependent order prompts. These order prompts are then associated with multiple annotated tags through the similarity measure function to form a similarity matrix. The Hungarian algorithm is introduced to acquire the bipartite matching between prompts and tags. In other words, the meaningful prompt is assigned to a specific tag, whereas the meaningless prompt is assigned to a pre-defined [PAD] tag. Since each prompt gets a unique assigned tag, we feed prompts into the order-aware tag decoder whose generation target is a sentence composed of corresponding tags.

Our model has several advantages: (1) Improved tag dependency modeling: we decouple the tag generation order by introducing order prompts as a guide, *i.e.* the generation order of tags depends on the input order of prompts, making the model focus on modeling the tag dependencies and thus alleviating the long tail problem; (2) Allowing the generation of novel tags: the model can infer novel tags by using the word-based generative model and pre-training; (3) Easy to extend: there is no need to pre-define a fixed number of tag sets and no need to modify the model for end-to-end training on new data.

We newly establish two benchmarks<sup>2</sup>: the Chinese video tagging benchmark CREATE-tagging and the English image tagging benchmark Pexel-tagging, to validate both the effectiveness and generalization of the proposed method. CREATE-tagging is comprised of CREATE-210K and CREATE-3M, which contain about 210K and 3M videos respectively. The larger dataset is used to validate the extensibility in pre-training mode. The Pexel-tagging contains 162k images, which is used to verify the generalization of the model in different languages and visual modalities. Videos/Images and titles are provided for each sample. The tags are separated into common high-frequency tags and rare low-frequency tags, followed by the introduction of label-based and example-based metrics to comprehensively evaluate the model’s performance at the tag and video levels, respectively. In addition, we define a novel metric, *tag gain*, to quantify the model’s ability to generate novel tags.

The contributions of this work are listed as follows:

- (1) Based on the practical application scenario, we analyze the characteristics of the video tagging task and explain the differences with the traditional multi-label classification and video captioning tasks.
- (2) We propose the order-prompted tag sequence generation approach to finish the video tagging task, which improves the modeling of tag relationships and allows the generation of novel tags.
- (3) We develop two benchmarks for evaluation, *i.e.*, CREATE-tagging and Pexel-tagging. The results of both benchmarks demonstrate that the proposed method is significantly superior to the others.

## 2. Related Work

**Multi-label Classification.** Mainstream multi-label classification methods require pre-defined categories, such as objects and actions, encoding input information through various backbone networks, such as CNN [27], GCN [4, 5, 29, 26, 30, 28] and Transformer encoder [36, 11, 37, 6], and then reasoning multiple categories simultaneously via a multi-classification head or multiple binary classification heads. In addition, a number of works [13, 20] are devoted to the investigation of novel loss functions to alleviate the problem of imbalanced positive and negative labels for each sample. These works mainly focus on the categories of video content. However, in the actual short video platform, the number of video tags far exceeds the number of video categories, making it difficult to generalize the multi-label classification methods to the task of video tagging.

**Multi-label Sequence Generation.** Some works model the multi-label classification task into a multi-label sequence generation problem and employ encoder-decoder

<sup>2</sup><https://drive.google.com/drive/folders/1gA4j9j0kWD99AIMGvpeaOK0M2xQCZHNo>

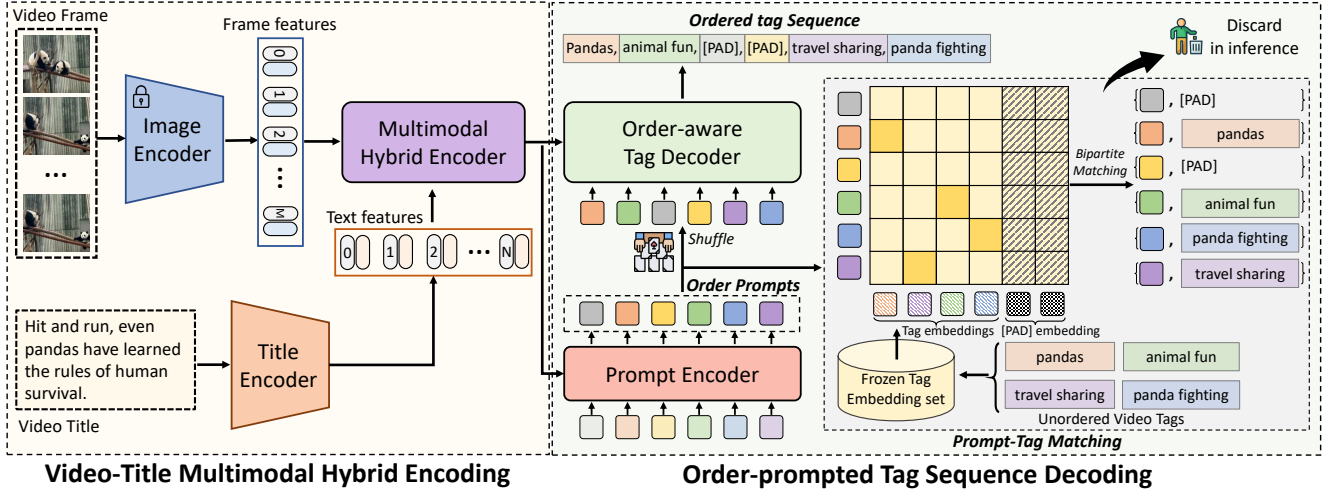


Figure 2. Overview of the proposed OP-TSG for video tagging. It consists of two major components: video-title multimodal hybrid encoding and order-prompted tag sequence decoding. Video-title multimodal hybrid encoding takes both video and title as input, which first independently encodes each modality and then performs video-title fusion. Order-prompted tag sequence decoding utilizes multimodal embeddings to produce sample-dependent order prompts, and further construct ordered tag sequence under the guidance of order prompts. OP-TSG uses a word rather than a tag as the basic generation unit.

approaches [22, 8, 31, 35, 34] to solve it. CNN-RNN [24] first explores the use of CNN and LSTM to generate a label sequence, which is sorted in descending order according to the occurrences of labels. Jin *et al.* [10] use the frequent-first and rare-first strategies to connect labels. Liu *et al.* [14] design a semantically regularised embedding layer as the interface between the CNN and RNN to enable efficient training. Order-Free [3] automatically provides a label connection order by the correlations between the visual areas and the labels. These methods still take each label as the basic unit of prediction, inevitably limiting the number of labels. Instead, our method takes the word as the base unit, flexibly predicting tags in the form of phrases and novel tags. Moreover, our order-prompt-based method is more effective than the rule-based methods at tag order decoupling.

**Learnable Queries for Decoding.** DETR [2] accomplishes object detection using encoder-decoder architecture for the first time and introduces a series of learnable queries to realize the decoding of the object sequence. The effectiveness of DETR has led to its expansion into additional fields. Moment-DETR [12] considers the queries as highlight moments in the video for language-based moment retrieval, and PDVC [25] creates connections between queries and events for dense video captioning. Inspired by prior research, we refer to learnable queries as order prompts and establish associations with tags to determine the generation order of tag sequence.

### 3. Method

In this section, we will introduce the specific workflow of the proposed OP-TSG: the Video-Title Multimodal Hybrid

Encoding in section 3.1, the Order-Prompted Tag Sequence Decoding in 3.2, and the training and inference in 3.3.

#### 3.1. Video-Title Multimodal Hybrid Encoding

We propose multimodal hybrid encoding to fully integrate the multimodal information of the video. As illustrated in Figure 2, we first sample frames from the video  $V_i$  before feeding each sampled frame into a frozen image encoder to acquire the frame features  $F_i^v$ . Since video tags usually contain a large number of specific entity concepts, such as the name of a celebrity or place, it is challenging to effectively obtain these concepts using only visual information. As a result, we choose to use the corresponding title as input, as certain concepts can be reflected directly in it. We use a pre-trained text encoder to encode the title and get text features  $F_i^t$  for each word.

The multimodal hybrid encoder  $M_E(\cdot, \cdot)$  is a multi-layer transformer encoder. Each layer consists of a multi-head self-attention head, a cross-attention head, and a feed-forward network. The frame features and text features are fused into multimodal hybrid features by  $F_i = M_E(F_i^t, F_i^v)$ , which treats the text features as queries and frame features as keys and values. The multimodal hybrid features will be used for both the encoding of order prompts related to the video and the decoding of the tag sequence.

#### 3.2. Order-Prompted Tag Sequence Decoding

Order-prompted tag sequence generation can be divided into three steps: (1) Providing learnable sample-dependent order prompts based on the interaction of initialized shared queries with multimodal hybrid features; (2) Aligning order prompts with multiple unordered tags and performing

bipartite matching to assign order prompts to unique tags; (3) Connecting the assigned tags to form the ordered tag sequence as a training target according to the order prompts.

### 3.2.1 Sample-dependent Order Prompts Encoding

The order-prompted tag sequence decoding begins with producing sample-dependent order prompts for each video. Learnable queries  $Q = \{q^{(n)}\}_{n=1}^N$  are initialized and shared between all the videos. The order prompts  $P_i = \{p_i^{(n)}\}_{n=1}^N$  for the video  $V_i$  are subsequently derived from the interaction of queries and multimodal hybrid features:

$$P_i = W_p(P_E(Q, F_i)), \quad (1)$$

where the prompt encoder  $P_E(\cdot, \cdot)$  is a cross-attention module composed of multi-layer transformers, which treats learnable queries as queries and multimodal hybrid features as keys and values;  $W_p(\cdot)$  is a linear projection. The order prompts are sample-dependent because they incorporate the specific content of the video, unlike the common visual prompts [38, 9] shared by all samples.

### 3.2.2 Alignment between Order Prompts and Tags

To make order prompts cover different semantic information, we use bipartite matching to uniquely align the order prompts to unordered video tags, as shown in the Prompt-Tag Matching part of Figure 2.

Let us denote by  $E_i = \{e_i^{(l)}\}_{l=1}^{L_i} = \{\text{LM}(t_i^{(l)})\}_{l=1}^{L_i}$  the embeddings of the ground-truth tags  $\{t_i^{(l)}\}_{l=1}^{L_i}$  extracted through a frozen pre-trained language model  $\text{LM}(\cdot)$  for the video  $V_i$ , and  $P_i = \{p_i^{(n)}\}_{n=1}^N$  the set of  $N$  order prompts. Assuming  $N$  is larger than the number of tags  $L_i$  in each video, we consider  $E_i$  as a set of size  $N$  padded with the embeddings of the pre-defined meaningless tag [PAD].  $\sigma(\cdot)$  is an index mapping function from the set  $P_i$  to the set  $E_i$ , *i.e.*,  $l = \sigma(n)$ . To find a bipartite matching between these two sets, we search for a permutation of  $N$  elements of  $\sigma$  with the lowest cost:

$$\hat{\sigma}_i = \arg \min_{\sigma} \sum_n -\mathcal{L}_{\text{match}}(p_i^{(n)}, e_i^{(\sigma(n))}), \quad (2)$$

where  $\mathcal{L}_{\text{match}}(p_i^{(n)}, e_i^{(l)})$  is a pair-wise matching cost with cosine similarity between the order prompt  $p_i^{(n)}$  and ground-truth tag embedding  $e_i^{(l)}$ . The optimal assignment is computed efficiently with the Hungarian algorithm, following prior work [2]. The final alignment result is  $\hat{l} = \hat{\sigma}_i(n)$ , and the corresponding tags are defined as:

$$\text{Align}(p_i^{(n)}) = \begin{cases} t_i^{(\hat{l})}, & \text{if } \hat{l} \in \{l\}_{l=1}^{L_i} \\ \text{[PAD]}, & \text{otherwise} \end{cases} \quad (3)$$

To further improve the accuracy of matching scores, we leverage contrastive learning by introducing InfoNCE loss

function [17] to pull the positive prompt-tag pairs and push the negative pairs, which is defined as follows:

$$\begin{aligned} \mathcal{L}_{i2p}^{i,\hat{l}} &= -\log \frac{\exp(e_i^{(\hat{l})}, p_i^{(n)}, \tau)}{\sum_{i'=1}^B \sum_{n'=1}^N \exp(e_i^{(\hat{l})}, p_{i'}^{(n')}, \tau)}, \\ \mathcal{L}_{p2t}^{i,\hat{l}} &= -\log \frac{\exp(p_i^{(n)}, e_i^{(\hat{l})}, \tau)}{\sum_{i'=1}^B \sum_{l'=1}^{L_i} \exp(p_i^{(n)}, e_{i'}^{(l')}, \tau)}, \end{aligned} \quad (4)$$

where  $\exp(x, y, \tau) = e^{x^\top y / \tau}$ ,  $\tau$  is a learnable temperature hyper-parameter, and  $B$  is the batch size. Note that only prompts associated with meaningful tags will be considered as valid positive pairs, *i.e.*  $\hat{l} = \hat{\sigma}_i(n) \in \{l\}_{l=1}^{L_i}$ . The total loss for prompt-tag contrastive learning is defined as:

$$\mathcal{L}_{cl} = \frac{\sum_{i=1}^B \sum_{\hat{l}=1}^{L_i} (\mathcal{L}_{i2p}^{i,\hat{l}} + \mathcal{L}_{p2t}^{i,\hat{l}}) / 2}{\sum_{i=1}^B \sum_{\hat{l}=1}^{L_i}}. \quad (5)$$

### 3.2.3 Ordered Tag Sequence Generation

After getting the alignment result of order prompts and tags, we rearrange the order of the ground-truth tags in the order of order prompts and combine them into a target sequence containing tags separated by commas:

$$T_s = \text{“Align}(p_i^{(1)}), \text{Align}(p_i^{(2)}), \dots, \text{Align}(p_i^{(N)})\text{”}. \quad (6)$$

Let’s take Figure 2 as an example, the target sequence after alignment is “[PAD],  $t_i^{(1)}$ , [PAD],  $t_i^{(3)}$ ,  $t_i^{(4)}$ ,  $t_i^{(2)}$ ”. To improve the modeling of tag relationships, we randomly shuffle the order of input order prompts, and tag connection order of the target tag sequence is changed accordingly, thus obtaining sequences with different tag combination patterns.

We build the order-aware tag decoder  $T_D(\cdot, \cdot, \cdot)$  based on a multi-layer transformer decoder to generate the tag sequence  $T_s$  word-by-word, conditioned on order prompts  $P_i$  and multimodal hybrid features  $F_i$  to achieve order-prompted tag sequence generation. The probability of predicting the word  $y_t$  can be expressed as follows:

$$p_{\theta}(y_t | y_{<t}, F_i, P_i) = \text{Softmax}(T_D(y_{<t}, F_i, P_i)). \quad (7)$$

The cross-entropy loss function is utilized for model training. Since there are numerous [PAD] tags in the tag sequence, using a common cross-entropy loss will cause the model to seek a shortcut, *i.e.* the model can converge quickly by simply focusing on these tags. To resolve this issue, we assign lower weights to [PAD] and higher weights to other words:

$$\Gamma(y_t) = \begin{cases} \lambda, & \text{if } y_t = \text{[PAD]} \\ 1, & \text{otherwise} \end{cases}, \quad (8)$$

where  $\lambda \in [0, 1]$ . The modified tags sequence generation loss is defined as follows:

$$\mathcal{L}_{gen} = - \sum_{i=1}^B \sum_{t=1}^{|T_s|} \Gamma(y_t) \cdot \log p_{\theta}(y_t | y_{<t}, F_i, P_i), \quad (9)$$

where  $|T_s|$  is the number of words in the sequence rather than the number of tags.

### 3.3. Training and Inference

We train the entire model end-to-end by jointly optimizing the prompt-tag contrastive loss  $\mathcal{L}_{cl}$  and tag sequence generation loss  $\mathcal{L}_{gen}$ , achieving the goal of injecting semantics into the order prompts and generating an ordered tags sequence based on the order prompts. The total loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{cl} + \mathcal{L}_{gen}. \quad (10)$$

The model outputs a tag sequence with [PAD] tags in the inference phase. We perform post-processing by splitting the tag sequence through the separator “;” and removing the meaningless [PAD] tags. The remaining multiple tags are the final inference results. Note that prompt-tag matching is only used during training to learn the semantic alignment of prompts and tags, and will be discarded during inference.

## 4. Benchmarks and Experimental Results

### 4.1. Datasets

We experiment on two newly established benchmarks, *i.e.*, the Chinese video tagging benchmark **CREATE-tagging** and the English image tagging benchmark **Pexel-tagging**, to validate the effectiveness and generalization of the proposed model on different visual and linguistic forms.

CREATE-tagging is based on the CREATE dataset [33], including the high-quality annotated CREATE-210K dataset with 210K videos for training and 5k videos for evaluation, and large-scale weakly-annotated CREATE-3M dataset with 3M videos for pre-training. The numbers of unique tags in CREATE-210K and CREATE-3M are 18464 and 57297, respectively. The 2795 tags that co-occur in the training and test set of CREATE-210K are split into 705 common high-frequency tags and 2090 rare low-frequency tags for evaluation, according to the split criterion being that the sum of common tag occurrences accounts for 90% of all tags. Pexel-tagging is based on the newly collected dataset Pexel, which includes 162k images for training and 5k images for evaluation. There are 28094 unique tags in Pexel, and 5669 tags appear in both training and test sets. Using the similar split criterion as CREATE-tagging, we obtain 1627 common tags and 4042 rare tags. Additional information regarding the two benchmarks can be found in the supplementary materials.

### 4.2. Evaluation Metrics

For tags within the tag set, we introduce the traditional label-based macro metrics [21] and example-based metrics

[32] to evaluate the Precision, Recall, and F1 score on the tag and video levels, respectively.

For tags outside the tag set, a novel metric called **tag gain** is defined to quantitatively evaluate the model’s ability to generate meaningful novel tags. Tags that are relevant to the video’s content but not appearing in the tag set contribute to tag gain. Existing cross-modal video-text matching technology [1, 16] can assess this automatically, but there are still numerous missed and false matchings. Therefore we use human evaluation, the gold standard, to determine whether the novel tags are relevant to video content. The formula is defined as follows:

$$\Delta = \frac{1}{|\mathcal{D}_V|} \sum_i \frac{|T_i^o|}{|T_i|}, \quad (11)$$

where  $|T_i^o|$  is the number of “*video matched tags*”, *i.e.* novel tags generated by the model that are relevant to the content of  $i$ -th video.  $|T_i|$  is the number of human-annotated ground-truth tags of the  $i$ -th video.  $\mathcal{D}_V$  is the set of all videos in the dataset.

### 4.3. Implementation details

OP-TSG adopts the CLIP-B/32 image encoder [18] as the frozen image encoder, and a 6-layer transformer [23] initialized with the first 6 layers of the BERT<sub>base</sub> model [7] as the title encoder. The multimodal hybrid encoder and the prompt encoder are all 6 layers, and the order-aware tag decoder is a 12-layer auto-regressive transformer decoder. Each video is uniformly sampled at 1fps, with a maximum of 60 frames. Patch features are supplied to the multimodal hybrid features instead of frame features during image tagging benchmark testing. We employ the AdamW [15] optimizer with a maximum learning rate of  $2e^{-4}$  and a weight decay of 0.002.  $\lambda$  for adjusting the loss weight is set to 0.3. The number of order prompts is set to 30 in our model. Additional implementation details are provided in the supplementary materials.

### 4.4. Comparison with Start-of-the-Art

OP-TSG is compared to advanced multi-label classification [19, 20] and generation methods [24, 34] using CREATE-tagging and Pexel-tagging benchmarks. For a fair comparison, we re-implement other methods with the same video-title input, encoder structure and initialization parameters, and training schedule as our method, and we also replace the original LSTM decoder with a 12-layer transformer decoder in the generation methods being compared.

Table 1 displays the results of CREATE-tagging. OP-TSG outperforms other methods for both label-based and example-based metrics on all tags. Moreover, OP-TSG demonstrates significant advantages in recognizing rare tags that users are interested in, achieving 3.3% and 3.5% F1 score gains over the SOTA methods Asy [20] and Open-Book [34], respectively. This verifies that our method can

| Method         | Category | Dataset                    | Label-Based |             |             |             |             |             |             |             |             | Example-Based |             |             |
|----------------|----------|----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|---------------|-------------|-------------|
|                |          |                            | Full        |             |             | Rare        |             |             | Common      |             |             |               |             |             |
|                |          |                            | P           | R           | F1          | P           | R           | F1          | P           | R           | F1          | P             | R           | F1          |
| Bin [19]       | Cls.     | CREATE-210K                | 21.1        | 28.8        | 22.4        | 14.9        | 17.1        | 14.6        | 39.4        | 63.4        | 45.5        | 38.6          | 39.6        | 39.1        |
| Asy [20]       | Cls.     |                            | 27.8        | 38.4        | 29.9        | 24.4        | 28.5        | 24.5        | 38.0        | <b>66.6</b> | <b>45.8</b> | 38.5          | 45.5        | 41.7        |
| Order-Free [3] | Gen.     |                            | 27.6        | 35.3        | 28.9        | 24.1        | 26.6        | 23.8        | 37.9        | 60.8        | 43.8        | 39.4          | 41.8        | 40.9        |
| Open-Book [34] | Gen.     |                            | 28.1        | 35.8        | 29.2        | 24.8        | 27.3        | 24.3        | 38.1        | 61.0        | 43.9        | 39.5          | 42.2        | 41.2        |
| Ours           | Gen.     |                            | <b>30.1</b> | <b>40.7</b> | <b>32.2</b> | <b>27.1</b> | <b>32.6</b> | <b>27.8</b> | <b>38.2</b> | 64.7        | 45.3        | <b>39.5</b>   | <b>46.3</b> | <b>42.7</b> |
| Bin [19]       | Cls.     | CREATE-3M +<br>CREATE-210K | 28.8        | 36.3        | 30.0        | 24.0        | 26.3        | 23.5        | 42.3        | 66.3        | 48.7        | 42.8          | 45.7        | 44.2        |
| Asy [20]       | Cls.     |                            | 30.5        | 40.1        | 32.4        | 27.1        | 30.9        | 27.2        | 40.8        | <b>67.3</b> | 47.9        | 41.1          | 46.3        | 43.6        |
| Order-Free [3] | Gen.     |                            | 31.5        | 38.2        | 32.3        | 28.0        | 29.7        | 27.2        | 41.9        | 62.7        | 47.5        | 42.4          | 44.1        | 43.2        |
| Open-Book [34] | Gen.     |                            | 31.8        | 38.6        | 32.5        | 28.3        | 29.9        | 27.5        | 42.2        | 63.2        | 47.8        | 42.6          | 44.6        | 43.5        |
| Ours           | Gen.     |                            | <b>34.1</b> | <b>42.6</b> | <b>35.5</b> | <b>31.2</b> | <b>34.6</b> | <b>30.9</b> | <b>42.6</b> | 66.3        | <b>48.9</b> | <b>43.0</b>   | <b>48.4</b> | <b>45.5</b> |

Table 1. Performance comparisons with state-of-the-art methods on CREATE-tagging. ‘‘Cls.’’ and ‘‘Gen.’’ indicate that the model belongs to multi-label classification and generation methods, respectively. ‘‘Full’’, ‘‘Rare’’, and ‘‘Common’’ denote that metrics are evaluated on all tags, rare tags, and common tags, respectively.

| Method         | Category | Dataset | Label-Based |             |             |             |             |             |             |             |             | Example-Based |             |             |
|----------------|----------|---------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|---------------|-------------|-------------|
|                |          |         | Full        |             |             | Rare        |             |             | Common      |             |             |               |             |             |
|                |          |         | P           | R           | F1          | P           | R           | F1          | P           | R           | F1          | P             | R           | F1          |
| Bin [19]       | Cls.     | Pexel   | 11.2        | 17.6        | 12.2        | 4.5         | 4.7         | 4.1         | 27.7        | 49.5        | 32.4        | 32.2          | 50.9        | 39.6        |
| Asy [20]       | Cls.     |         | 24.2        | 31.2        | 25.3        | 19.5        | 21.5        | 19.0        | 36.1        | <b>55.2</b> | 40.9        | 38.7          | <b>55.9</b> | 45.8        |
| Order-Free [3] | Gen.     |         | 37.6        | 38.6        | 36.5        | 35.7        | 35.4        | 34.0        | 42.8        | 46.7        | 42.7        | 48.2          | 49.1        | 48.4        |
| Open-Book [34] | Gen.     |         | 37.7        | 38.9        | 36.7        | 35.9        | 35.7        | 34.2        | <b>42.9</b> | 47.1        | 42.9        | <b>48.5</b>   | 49.6        | 48.8        |
| Ours           | Gen.     |         | <b>38.1</b> | <b>44.6</b> | <b>39.1</b> | <b>37.5</b> | <b>41.1</b> | <b>37.2</b> | 39.8        | 53.4        | <b>43.6</b> | 45.1          | 55.3        | <b>49.7</b> |

Table 2. Performance comparisons with state-of-the-art methods on Pexel-tagging. Evaluation settings are the same as CREATE-tagging.

| Method         | Category | # Video matched tags | $\Delta$     |
|----------------|----------|----------------------|--------------|
| Bin [19]       | Cls.     | -                    | 0.0%         |
| Asy [20]       | Cls.     | -                    | 0.0%         |
| Order-Free [3] | Gen.     | 0.11                 | 2.43%        |
| Open-Book [34] | Gen.     | 0.15                 | 3.43%        |
| Ours           | Gen.     | <b>0.30</b>          | <b>7.04%</b> |

Table 3. Comparisons of tag gain through different models. The numbers of video matched tags listed in the third column are the average of all videos.

alleviate the long-tail problem by capturing better tag dependencies. Similar conclusions are reached when introducing CREATE-3M for pre-training, demonstrating that our method scales well for pre-training.

The results of Pexel-tagging, a benchmark with more tags and complex tag distribution, are presented in Table 2. OP-TSG achieves the highest F1 scores across all settings, especially improving the F1 scores by 18.2% and 3% over Asy and Open-Book on rare tags, validating the generalization of our method.

#### 4.5. Tag gain via human evaluation.

Through the route of pre-training on CREATE-3M and then fine-tuning with CREATE-210K, the video matched tags and tag gain of different models can be evaluated on the test dataset of CREATE-210K. As shown in Table 3, clas-

sification methods, such as Bin and Asy, are unable to infer novel tags due to the classification head can only output a fixed number of tags. All the generation methods adopt a word-based generation strategy, so they all exhibit the ability to generate novel tags. Among them, OP-TSG generates an average of 0.3 video matched tags per video and obtains a tag gain of 7.04%, which significantly outperforms other generation models. This indicates that the semantics of tags injected to order prompts in pre-training can be effectively retained during fine-tuning, thus enabling the model to still generate tags that are only presented in pre-training data.

#### 4.6. Ablation study

In this section, we conduct ablation studies based on the CREATE-tagging benchmark to verify the effectiveness of each component. All experiments are performed on CREATE-210K and reported label-based F1 scores on all tags, rare tags, and common tags.

**The effect of the number of order prompts on results.** As shown in Table 4, the optimal performance is reached with 30 order prompts, and we analyze the reasons as follows: (1) When reducing the number of order prompts, it is difficult for prompts to cover all the semantics of the abundant tags. (2) When increasing the number of order prompts, the proportion of [PAD] tags relative to meaningful tags also rises, exacerbating the imbalance between the number of meaningful tags and [PAD] tags.

| Number of Prompts. | Full        | Rare        | Common      |
|--------------------|-------------|-------------|-------------|
| 5                  | 29.8        | 25.0        | 44.1        |
| 10                 | 30.9        | 26.3        | 44.7        |
| 15                 | 31.7        | 27.0        | 45.3        |
| 20                 | 31.8        | 27.3        | 45.2        |
| 30                 | <b>32.2</b> | <b>27.8</b> | <b>45.3</b> |
| 50                 | 31.6        | 26.8        | 45.1        |

Table 4. Comparisons of different numbers of order prompts.

| Weight for [PAD] Tag | Full        | Rare        | Common      |
|----------------------|-------------|-------------|-------------|
| 0.1                  | 31.4        | 27.6        | 42.8        |
| 0.3                  | <b>32.2</b> | <b>27.8</b> | 45.3        |
| 0.5                  | 31.7        | 27.0        | 45.9        |
| 0.8                  | 31.6        | 26.8        | 46.0        |
| 1.0                  | 31.2        | 26.2        | <b>46.1</b> |

Table 5. Comparisons of different [PAD] loss weights.

**The effect of [PAD] loss weight on tags prediction.** As shown in Table 5, when the [PAD] loss weight is small, *i.e.*,  $\lambda = 0.1$ , the F1 score is lower, especially for the prediction of common tags, because the model tends to produce more incorrect tags, thus reducing the Precision. Conversely, a larger  $\lambda$  also decreases the F1 score, especially for the prediction of rare tags, as the model generates more meaningless [PAD] tags, thus reducing the Recall.

**The effect of the number of prompt encoder layers on results?** According to the results in Table 6, the change in the number of layers has the greatest effect on the F1 score for rare tags. We argue that it is challenging to build the correct associations between order prompts and rare tags with fewer layers. In addition, the performance saturates when the number of layers reaches 6. We finally choose to use a 6-layer prompt encoder based on the above discussion. **Is the order-prompted tag sequence decoding mechanism effective?** Yes. Model B in Table 7 is the baseline of generation methods, *i.e.*, Video-Title Multimodal Hybrid Encoding module is directly connected to the decoder, which discards the order prompts and connects tags from high-frequency to low-frequency to construct the target tag sequence. Compared with the proposed Model A, the F1 scores of Model B on rare and common tags decreased by 4.6% and 2.1%, respectively, indicating the mechanism in OP-TSD is superior to the pre-defined tag connection rule.

**Is the shuffling of order prompts effective?** Yes. Comparing Model A with Model C in Table 7, training without shuffling the prompts yields comparable performance on common tags, but a decrease in F1 scores of 3% and 2.1% for the more important rare tags and all tags. By prompt shuffling to construct various tag combination patterns as training targets, the model can capture better tag dependencies and thus enhance the ability to generate rare tags.

**Should tag embeddings be frozen for prompt-tag association?** Yes. Model D trained by fine-tuning the language

| Depth of $P_E$ | Full        | Rare        | Common      |
|----------------|-------------|-------------|-------------|
| 1              | 31.4        | 26.6        | 45.5        |
| 2              | 31.8        | 27.2        | 45.4        |
| 4              | 31.9        | 27.4        | 45.5        |
| 6              | <b>32.2</b> | <b>27.8</b> | 45.3        |
| 8              | 32.1        | 27.5        | <b>45.6</b> |

Table 6. Comparisons of producing order prompts with different numbers of prompt encoder layers.

| Method               | Full        | Rare        | Common      |
|----------------------|-------------|-------------|-------------|
| A Ours               | <b>32.2</b> | <b>27.8</b> | 45.3        |
| B w/o OP-TSD         | 28.3        | 23.2        | 43.2        |
| C w/o shuffle        | 30.1        | 24.8        | 45.4        |
| D Learnable tag emb. | 31.4        | 26.6        | <b>45.6</b> |
| E w/o prompt-tag cl. | 28.4        | 22.9        | 44.8        |
| F Prompt label       | 29.7        | 25.0        | 43.4        |
| G Prompt cls.        | 28.7        | 23.4        | 44.5        |

Table 7. Effectiveness of the proposed components.

model to extract learnable tag embeddings shows a slight performance degradation compared to Model A. We analyze that the learnable tag embeddings will continuously change the alignment objective of the order prompts, leading to instability in prompt-tag contrastive learning during early training. Furthermore, a fixed pre-trained language model is sufficient to extract semantic embeddings of tags.

**Without prompt-tag contrastive learning.** Model E in Table 7 removes the prompt-tag contrastive learning during training, and its performance is significantly inferior to Model A, especially on rare tags. The objective of prompt-tag contrastive learning is to inject the semantics of assigned tags into order prompts, thus enabling the model to decode the correct tags based on the semantics of order prompts. Discarding the contrast learning will weaken the semantics of the order prompts and make it difficult for the model to decode the correct sequence.

**Adding [PAD] tags vs. Predicting prompt label.** Another scheme to handle the prompts aligned with [PAD] tags is to train a binary classification head to predict label 0 for these prompts and 1 for others. During training and inference, the decoder only receives prompts with label 1. Model F in Table 7 is trained in this way, and the F1 scores drop significantly compared with Model A. We attribute the results to two reasons: (1) 0, 1 labels are predicted independently on each prompt with unsatisfied accuracy. However, the generation of the [PAD] tags is performed after all prompts interact deeply in the decoder and thus enjoys a higher inference accuracy. (2) Prompts aligned with [PAD] tags may also benefit the decoding of other meaningful prompts, and thus directly removing them is not appropriate.

**Prompts for Generation vs. Prompts for Classification.** Prompts for classification mean attaching a multi-classification head on each order prompt to predict the tags

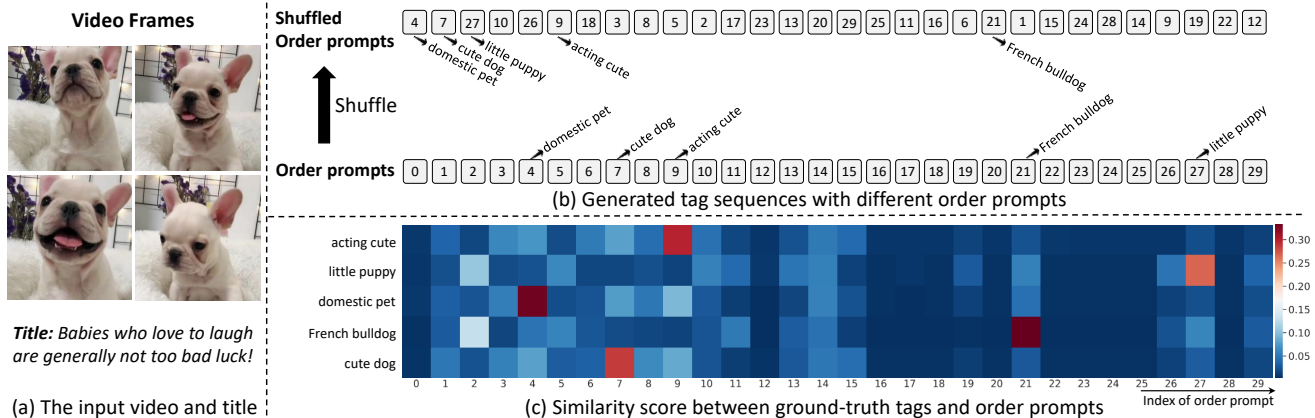


Figure 3. Visualization of tag sequence generation. We show tag sequences generated with different order of prompts in (b) and the inferred [PAD] tags are omitted for simplicity. The similarity score matrix between ground-truth tags and order prompts is shown in (c).

|   | Video | Title | Full        | Rare        | Common      |
|---|-------|-------|-------------|-------------|-------------|
| A | ✓     |       | 21.8        | 16.8        | 36.3        |
| B |       | ✓     | 23.3        | 21.6        | 28.3        |
| C | ✓     | ✓     | <b>32.2</b> | <b>27.8</b> | <b>45.3</b> |

Table 8. Ablation study on the importance of video and correspond title for tag generation.

directly after prompt-tag alignment, Model G in Table 7 implements this method. Compared to Model A, the F1 score of Model G on rare and common tags is reduced by 4.4% and 0.8%, respectively. We believe that it is caused by the independent reasoning for each order prompt lacks consideration of tag dependence. The architecture of Model F and Model G can be found in supplementary materials.

**Benefits of multimodal hybrid information.** We ablate the input multimodal hybrid information and the results are shown in Table 8. When only video frame features are available, the obtained Model A suffers a severe F1 score drop on rare tags compared to Model C. It is due to the fact that the vast entity concepts contained in the video tags are difficult to reason directly from visual information. On the other hand, Model B, which only takes titles as input, has a larger performance gap than Model C on common tags. This is because most of the useful information contained in the title are entities, which are usually beneficial for inferring rare tags rather than common tags with generic meanings.

#### 4.7. Qualitative Analysis

**Visualization of tag sequence generation.** In Figure 3, we visualize the tag sequences generated with different orders of prompts in (b) and the similarity score between ground-truth tags and order prompts in (c). We make the following qualitative observations: (1) Our model can decode the correct tag sequence according to the semantics of order prompts. *E.g.*, according to Figure 3 (c), the semantics of the prompt with index 4 match with the tag “domestic pet”



|  |             |   |
|--|-------------|---|
| (a) CREATE-Tagging<br><br><i>Title: The legendary stepping on the sword flight.</i>                 | <b>GT:</b>  | sports show off, sports master, please do not imitate, <b>extreme skydiving</b>                               |
|  | <b>Cls.</b> | sports show off, sports master, please do not imitate, <b>extreme cycling</b>                                 |
|  | <b>Gen.</b> | sports show off, please do not imitate  |
|  | <b>Ours</b> | sports show off, sports master, please do not imitate, <b>extreme skydiving</b> , <b>aerobatics</b>           |
| (b) Pexel-Tagging<br><br><i>Title: A low angle view of a female model holding a tennis racket.</i> | <b>GT:</b>  | dress, young, young woman, <b>fashion model</b> , <b>fashion tennis</b> , <b>tennis</b> , <b>tennis court</b> |
|  | <b>Cls.</b> | young woman, female, <b>tennis</b>  |
|  | <b>Gen.</b> | young, young woman, dress, <b>tennis</b> , <b>tennis court</b>  |
|  | <b>Ours</b> | young, young woman, dress, <b>fashion model</b> , <b>fashion tennis</b> , <b>tennis</b> , <b>tennis court</b> |

Figure 4. Examples of tag inference results from multiple methods. “Cls.” and “Gen.” indicates the classification method Asy and generation method Open-Book, respectively. The tags in black, green, red, and purple are common tags, correct rare tags, incorrect rare tags, and novel tags, respectively.

and our model also generates the corresponding tag from this prompt as shown at the bottom of Figure 3 (b). (2) Our model can decode the correct tag sequences regardless of the prompt order. As shown in Figure 3 (b), with shuffled order prompts, the relative order of generated tags is changed but the correspondence between prompts and generated tags remains unchanged.

**Case study of different methods.** In Figure 4, we show examples of inferred tags from different methods on two benchmarks. It can be seen that (1) On CREATE-tagging, our method achieves more accurate recognition of rare tags compared with the classification method and generation method, *e.g.*, our method identifies the tag of “extreme skydiving” while the classification method produces a wrong tag of “extreme cycling” and generation method misses this tag. In addition, our method also provides a meaningful novel tag “aerobatic”. (2) On Pexel-tagging, our method shows greater superiority in generating a more comprehensive tag set, especially for rare tags.



## 5. Conclusion

In this paper, we present a novel generative model OP-TSG for video tagging. OP-TSG introduces a novel order-prompted tag sequence decoding mechanism to deal with the disordered nature of multiple parallel tags and to improve tag dependency modeling. The word-by-word tag generation strategy is also adopted for the first time to discard the fixed tag classification head. We create two new tagging benchmarks to verify the effectiveness and generalization of our method on image/video in English/Chinese. Experimental results show that our OP-TSG is superior to previous methods, especially for the generation of rare and novel tags. In the future, we will investigate the use of retrieval enhancement technology to strengthen the ability of knowledge extraction and new word generation.

**Acknowledgments** This work is supported by National Key R&D Program of China (No. 2022ZD0118500), Beijing Natural Science Foundation (Grant No. JQ21017, L223003), the Natural Science Foundation of China (Grant No. 61972397, 62036011, 62192782, U2033210, 62172413).

## References

- [1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.
- [3] Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Order-free rnn with visual attention for multi-label classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [4] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 522–531, 2019.
- [5] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019.
- [6] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, Dong Shen, Zhongyuan Wang, Nian Shi, and Honglin Liu. Mltr: Multi-label classification with transformer. *arXiv preprint arXiv:2106.06195*, 2021.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [8] Yaosi Hu, Zhenzhong Chen, Zheng-Jun Zha, and Feng Wu. Hierarchical global-local temporal modeling for video captioning. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 774–783, 2019.
- [9] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.
- [10] Jiren Jin and Hideki Nakayama. Annotation order matters: Recurrent image annotator for arbitrary length image tagging. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2452–2457. IEEE, 2016.
- [11] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16478–16488, 2021.
- [12] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.
- [13] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [14] Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. Semantic regularisation for recurrent image annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2872–2880, 2017.
- [15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [16] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.
- [17] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [19] Jesse Read and Fernando Perez-Cruz. Deep learning for multi-label classification. *arXiv preprint arXiv:1502.05988*, 2014.
- [20] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihl Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021.
- [21] Mohammad S Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18(1):25, 2010.
- [22] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of*

- the AAAI conference on artificial intelligence, volume 31, 2017.
- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [24] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016.
- [25] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857, 2021.
- [26] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12265–12272, 2020.
- [27] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1901–1907, 2015.
- [28] Jiahao Xu, Hongda Tian, Zhiyong Wang, Yang Wang, Wenxiong Kang, and Fang Chen. Joint input and output space learning for multi-label image classification. *IEEE Transactions on Multimedia*, 23:1696–1707, 2020.
- [29] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *European Conference on Computer Vision*, pages 649–665. Springer, 2020.
- [30] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12709–12716, 2020.
- [31] Junchao Zhang and Yuxin Peng. Object-aware aggregation with bidirectional temporal graph for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8327–8336, 2019.
- [32] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.
- [33] Ziqi Zhang, Yuxin Chen, Zongyang Ma, Zhongang Qi, Chunfeng Yuan, Bing Li, Ying Shan, and Weiming Hu. Create: A benchmark for chinese short video retrieval and title generation. *arXiv preprint arXiv:2203.16763*, 2022.
- [34] Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9837–9846, 2021.
- [35] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13278–13288, 2020.
- [36] Jiawei Zhao, Ke Yan, Yifan Zhao, Xiaowei Guo, Feiyue Huang, and Jia Li. Transformer-based dual relation graph for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 163–172, 2021.
- [37] Jiawei Zhao, Yifan Zhao, and Jia Li. M3tr: Multi-modal multi-label recognition with transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 469–477, 2021.
- [38] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.