Order-Prompted Tag Sequence Generation for Video Tagging
Supplementary Materials

Data	Videos	Unique tags	Union	IS.	Common	Rare
Train Test	210k 5k	17573 3686	18464	2795	705	2090

Table 1. Statistics for the CREATE-210k dataset. "IS." is the abbreviation of intersection.

Data	Images	Unique tags	Union	IS.	Common	Rare
Train Test	162k 5k	27752 6003	28094	5669	1627	4042

Table 2. Statistics for the Pexel dataset. "IS." is the abbreviation of intersection.

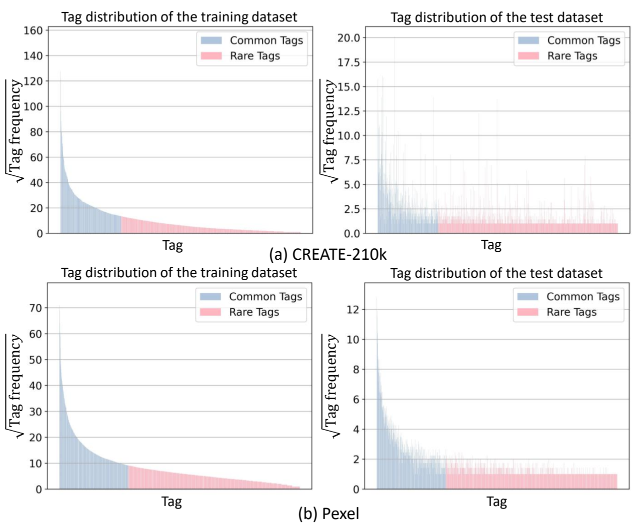


Figure 1. Tag distributions of the CREATE-210k dataset (a) and the Pexel dataset (b). The horizontal axis represents the index of the tag, and the vertical axis is the square root of the tag frequency for better visualization.

1. Benchmarks Details

1.1. CREATE-tagging Benchmark

The core part of CREATE-tagging is the CREATE-210k dataset [8]. As shown in Table 1, CREATE-210k contains 162k videos with 17573 unique tags for training and 5k videos with 3686 unique tags for test. A total of 18464 unique tags appear in CREATE-210k, of which 2795 tags appear in both the training and test data and are further divided into 705 common tags and 2090 rare tags to evaluate the performance of different models. The distributions of the 2795 tags in the training and test data are shown in Figure 1 (a), it can be seen that the tag distribution of the test data (right side) is not exactly consistent with the tag distribution of the training data (left side), *i.e.*, some rare tags in the training data appear frequently in the test data while some common tags appear infrequently in the test data.

Dataset	Data type	Videos/Images	Tags/Labels	Domain
CREATE-3M	Video Tag.	3M	57297	Open
CREATE-210k	Video Tag.	215k	18464	Open
Pexel	image Tag.	167k	28094	Open
THUMOS14 [3]	Video MLC.	413	20	Action
ActivityNet1.3 [2]	Video MLC.	15k	200	Action
MS-COCO [5]	Image MLC.	122k	80	Object
NUS-WIDE [1]	Image MLC.	210k	81	Object
WIDER Attribute [4]	Image MLC.	14k	14	Human att.
PA-100K [6]	Image MLC.	100k	26	Human att.

Table 3. Details of video/image tagging datasets and video/image multi-label classification datasets. "Video Tag.", "Image Tag.", "Video MLC." and "Image MLC." indicate that the dataset belongs to video tagging datasets, image tagging datasets, video multi-label classification datasets and image multi-label classification datasets, respectively. "Human att." is the abbreviation of human attributes.

Method	Category	Time(s) ↓	Full ↑	Rare ↑	Common ↑
ASY [7]	Cls.	**33**	25.3	19.0	40.9
Open-Book [9]	Gen.	271	36.7	34.2	42.9
Ours	Gen.	450	**39.1**	**37.2**	**43.6**

Table 4. Performance and inference time comparisons with our method, Asy and Open-Book.

1.2. Pexel-tagging Benchmark

Pexel-tagging is built on the newly collected and challenging Pexel dataset. According to the statistics in Table 2, Pexel consists of 162k images with 27752 unique tags for training and 5k images with 6003 unique tags for test. There are 28094 unique tags appearing in Pexel, of which 5669 tags appear in both the training and test data and are further divided into 1627 common tags and 4042 rare tags to evaluate the performance of different models. The distributions of the 5669 tags in the training and test data are shown in Figure 1 (b), compared with CREATE-210k, the tag distribution of the test data is more consistent with that of the training data.

2. Comparisons of Tagging Datasets and Multi-Label Classification Datasets

We list the details of video/image tagging datasets and video/image multi-label classification datasets in Table 3 to compare their differences more intuitively, and we observe two important distinctions: (1) Video/image tagging datasets are oriented towards open scenarios. In contrast, multi-label classification datasets are restricted to specific scenarios, *i.e.*, video multi-label classification datasets THUMOS14 and ActivityNet1.3 focus on actions, and im-

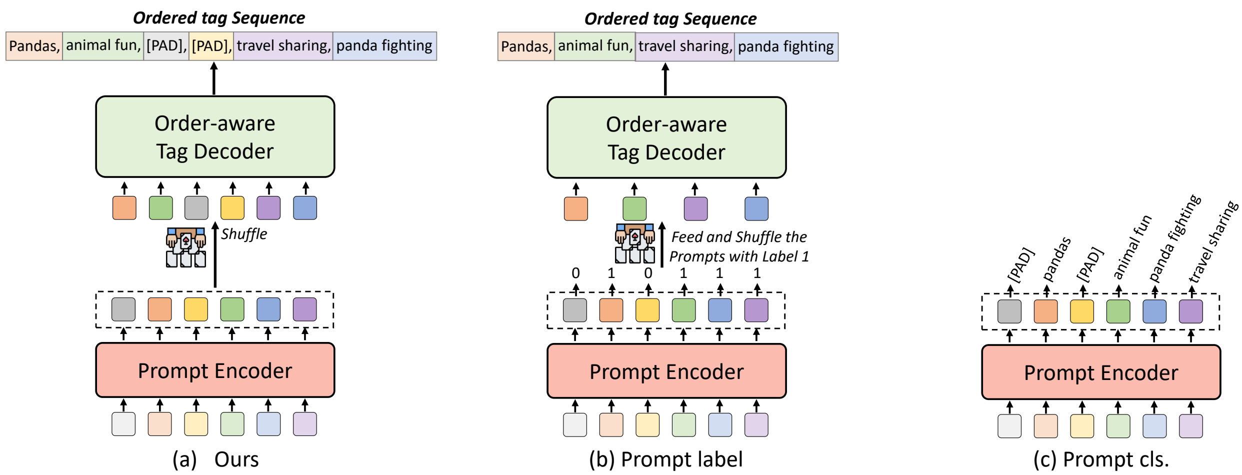


Figure 2. Architectures of different models: (a) Our OP-TSG. (b) Model F in Table 7 of the main paper, which uses prompt labels for handling meaningless prompts. (c) Model G in Table 7 of the main paper, which uses prompts for classification rather than generation.

age multi-label classification datasets focus on object categories (MSCOCO and NUS-WIDE) or human attributes (WIDER Attribute and PA-100k). (2) Video/image tagging datasets contain tens of thousands of unique tags, while multi-label classification datasets contain only tens or hundreds of unique labels.

3. Implementation details

3.1. Training schedules of OP-TSG

The training schedules of OP-TSG on CREATE-tagging and Pexel-Tagging are as follows: (1) For CREATE-tagging, we pre-train the model for 20 epochs with a batch size of 1024 on 16 NVIDIA V100 GPUs, followed by 30 epochs of fine-tuning with a batch size of 512 on 8 NVIDIA V100 GPUs. (2) For Pexel-tagging, we train the model for 20 epochs with a batch size of 512 on 8 NVIDIA V100 GPUs. Other methods used for comparison in Table 1 and Table 2 of the main paper also adhere to the same training schedules.

3.2. Comparisons of Model Architectures

The architectures of our OP-TSG, model F and model G in Table 7 of the main paper are shown in Figure 2 (a), Figure 2 (b) and Figure 2 (c), respectively. OP-TSG adopts the pre-defined [PAD] tags to assign to the meaningless order prompts. Model F trains a binary classification head that takes the order prompts as input, and predicts label 1 for the prompts aligned with meaningful tags and label 0 for the prompts aligned with [PAD] tags. Then prompts with label 1 are fed into the tag decoder and the target sequence is the concatenation of the aligned tags of the prompts. Model G directly attaches a multi-classification head on the order prompts, and trains the multi-classification head to predict the aligned tags of the prompts.

4. Inference Time Measurement

We evaluate the inference time of different models on Pexel-tagging using a single NVIDIA V100 GPU with a test batch size of 8, and the results are presented in Table 4. The classification model ASY runs the fastest but has the worst performance. Compared with the generation model open-book, our method improves the F1 score by 2.4% on all tags and expands the inference time to 1.7 times, the reason for the increase in inference time is that our method needs to generate longer tag sequences containing [PAD] tags.

5. More Visualizations

5.1. Tag inference results on CREATE-tagging

In Figure 3, we show examples of tag inference results of different methods on CREATE-tagging benchmark. We can observe that: (1) Our method is better at generating a more comprehensive tag set. As shown in Figure 3 (a), the classification method misses the tags “appetizers” and “cold noodles” and the generation model misses the tags “appetizers” and “summer noodles”, while our method provides a complete set of tags. (2) Our method is able to infer the wrong and meaningful tag that is outside the annotations but present in the training data and consistent with the video content, *e.g.*, the tag “western dessert” in Figure 3 (b). (3) Our method may also produce some wrong and meaningless tags, such as the tags “amateur billiard”, “top skills” and “waiters” in Figure 3 (c). (4) Our method can generate novel meaningful tag that is beyond the annotations and not present in the training data, such as the tag “Tianmen Mountain” in Figure 3 (d).

5.2. Tag inference results on Pexel-tagging

We also present examples of tag inference results of different methods on Pexel-tagging benchmark in Figure 4,

Title: Spicy and appetizing cold noodles, teach you how to make it at home, especially suitable for summer.

GT:	pasta, food tutorial, appetizers, summer food, cold noodles
Cls.	pasta, food tutorial, summer food
Cap.	pasta, food tutorial, cold noodles
Ours	pasta, food tutorial, appetizers, summer food, cold noodles

(a) Generate the complete tag set

Title: The legendary cheese waterfall really deserves its reputation, and the saliva fell all over the floor before eating.

GT:	food snapshots, food production, cheese
Cls.	food snapshots, food production, cheese
Cap.	food snapshots, cheese
Ours	food snapshots, food production, western dessert, cheese

(b) Generate wrong but meaningful tag

Title: Amazing my sister, but your cooperation is really seamless.

GT:	folk master, unique skill
Cls.	show
Cap.	folk master, unique skill
Ours	amateur billiard, folk masters, top skills, waiters, unique skills

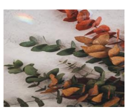
(c) Generate wrong and meaningless tag

Title: The manual interference in Tianmen Mountain in Zhangjiajie is too serious, and it has been automated.

GT:	travel news, amusement facilities, travel real shots, Zhangjiajie, Hunan tour
Cls.	travel news, amusement facilities, travel real shots, architectural landscape, Zhangjiajie
Cap.	travel news, amusement facilities, travel real shots, Zhangjiajie
Ours	travel news, amusement facilities, travel real shots, Zhangjiajie, Tianmen Mountain

(d) Generate novel and meaningful tag

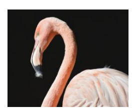
Figure 3. Examples of tag inference results from multiple methods on CREATE-tagging. "Cls." and "Gen." indicates the classification method Asy and generation method Open-Book, respectively. The tags in black, green, red, and purple are common tags, rare tags, incorrect tags, and novel tags, respectively.



Title: Leaves with Autumn Colors.

GT:	dried, green leaves, autumn colors, brown leaves, red leaves
Cls.	dried, leaves, season
Cap.	dried, green leaves, autumn colors, red leaves
Ours	dried, green leaves, autumn colors, brown leaves, red leaves

(a) Generate the complete tag set



Title: A Close-Up Shot of a Pink Flamingo.

GT:	animal, animal photography, avian, bird, feathers, flamingo
Cls.	animal, avian, bird
Cap.	animal, avian, bird, flamingo
Ours	animal, animal photography, avian, bird, flamingo

(b) Miss the tag

Title: A Man in a Hoodie Playing Drums.

GT:	hobby, leisure, musical instrument, musician, drummer, drums
Cls.	playing, hobby, musician, drummer, drums
Cap.	hobby, leisure, musical instrument, musician, drummer, drums
Ours	hobby, leisure, musical instrument, musician, drummer, drums, drumsticks

(c) Generate wrong but meaningful tag

Title: Close-up Photo of Dried Leaf on an Opened Book.

GT:	dried leaves, eyeglasses, eyewear, open book, soft focus, tilt shift, spectacles
Cls.	eyewear, overhead, aesthetic
Cap.	dried leaves, eyeglasses, eyewear, open book, soft focus, tilt shift, spectacles
Ours	dried leaves, eyeglasses, eyewear, open book, soft focus, tilt shift, spectacles, eye level shot, aesthetic

(d) Generate wrong and meaningless tag

Figure 4. Examples of tag inference results from multiple methods on Pexel-tagging. "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, rare tags, incorrect tags, and novel tags, respectively.

and we make the following observations: (1) Our method is able to generate an accuracy and complete tag set, *e.g.*, in 4 (a), the classification method provides the wrong tags “leaves” and “season” and the generation model misses the tag “brown leaves”, while our method generates all tags accurately. In addition, when our method misses tags, other methods will tend to miss more tags as well. As shown in Figure 4 (b), the classification method, the generation method and our method miss three, two and one tags, respectively. (2) Our method has the ability to infer wrong and meaningful tag that is outside the annotations but present in the training data and consistent with the image content, such as the tag “drumsticks” in Figure 3 (c). (3) Our method may also make the mistake of generating some wrong and meaningless tags, such as the tags “eye level shot” and “aesthetic” in Figure 3 (d).

References

[1] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 961–970. IEEE, 2015.

[3] Yu-Gang Jiang, Jingen Liu, Amir R. Zamir, G. Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. THUMOS challenge: Action recognition with a large number of classes. 2014.

[4] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 684–700. Springer, 2016.

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[6] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 350–359, 2017.

[7] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021.

[8] 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.

[9] 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.