Clover ☘: Towards A Unified Video-Language Alignment and Fusion Model
Supplementary Materials

A. Downstream tasks

We evaluate the proposed method on multiple downstream tasks, including three retrieval tasks for both zero-shot and fine-tuning settings, and eight video question answering tasks. The details of the datasets are as follows:

Text-to-Video Retrieval. (a) **MSRVTT** [11] contains 10K videos with 200K descriptions. For fine-tuning, we use the 1k-A split [12] which has 9K videos to train and 1K videos for test. (b) **DiDeMo** [1] contains 10K Flickr videos with 40K text sentences, where the test set contains 1,000 videos. We follow [4, 5] and evaluate Clover on the paragraph-to-video retrieval, where text sentences for each video are concatenated together as one text query. We do not use the ground-truth proposal for fair comparison with previous work. (c) **LSMDC** [8] consists of 118,081 video clips sourced from 202 movies. The validation set contains 7,408 clips and evaluation is done on a test set of 1,000 clips.

Video Question Answering. We conduct experiment on two kinds of video question answering tasks: multiple-choice QA and open-ended QA. Multiple-choice QA is to select the correct answer from multiple candidate answers given the question and video. Open-ended QA is to answer the question with free form answers, and we build answer candidates from the answers in the training set in each dataset.

Multiple-choice QA. (a) **TGIF-Action** [3] contains 18,428 GIF-question pairs for training and 2,274 GIF-question pairs for test, each GIF-question pair has 5 choices and one is true. (b) **TGIF-Transition** [3] contains 47,434 GIF-question pairs for training and 6,232 GIF-question pairs for test, each GIF-question pair has 5 choices and one is true. (c) **MSRVTT-MC** [12] contains 6,513 videos for training and 2990 videos for test. Each video has 5 text queries and one query is true. (d). **LSMDC-MC** [9] contains 101,079 clips, 7,408 clips and 10,054 clips for training, validation and test respectively. Each video clip have 5 text descriptions and one is true.

Open-Ended QA. (a) **TGIF-Frame** [3] contains 35,453 GIF-question pairs for training and 13,691 for test, with 1,540 answer candidates. (b) **MSRVTT-QA** [10] contains 149,075 video-question pairs for training and 72,821 for test, with 1,500 answer candidates. (c) **MSVD-QA** [10] contains 29,883 video-question pairs for training and 13,157 for test, with 1,000 answer candidates. (d) **LSMDC-FiB** [8] is a Fill-in-the-blank task. Given a video clip and a sentence with blank in it, the model need to predict a single correct word for the blank. The train set contains 296,960 clip-sentence pairs and test set contain 30,349 clip-sentence pairs. The answer candidates number is 908.

A.1. Details on Transferring to Downstream Tasks

For text-to-video retrieval, we use Recall at K (R@K) as the evaluation metric. For video question answering, we evaluate our Clover using accuracy. We use AdamW [6] to fine-tune Clover for each downstream task with betas of (0.9, 0.98), and we list the hyper-parameter settings for all downstream experiments in Tab. 1. The video frame size is 224 $\times$ 224 during pre-training and fine-tuning. Following [2], we use 8 frames during the finetune stage for every dataset except the DiDeMo retrieval benchmark. For DiDeMo, in which videos are relatively longer than other datasets, we find that use 64 frames for finetuning as in [7] can achieve higher performance. When frame number is set to 8 in DiDeMo, the recall@1/5/10 achieved by Clover are 46.1/74.8/82.7; when frame number is set to 64, the recall@1/5/10 are 50.1/76.7/85.6. We note that our Clover even outperforms the CLIP-pretrained method, e.g., CLIP4CLIP [7] in DiDeMo and LSMDC datasets, which demonstrates the superiority of our method.

B. Additional experiments

The effects of Clover on different modalities. To validate the effects of our Clover on video modality and text modality, we report results of model trained with *Clover w/o T* and *Clover w/o V* in Tab. 2. *Clover w/o T* represent replacing M_{T_m} anchored TMA with simple contrastive objective between $< V_e, T_e >$, and vice versa for *Clover w/o V*. The results show that applying Clover to either video modality or text modality brings about performance improvements, while the complete Clover performs the best.

Effect of exclusive-NCE. We present the results for different objective function designs for Tri-modal alignment in Tab. 3. Taking tri-modal alignment task for example,

Tasks	Fine-tuning Hyper-parameters				
	learning rate	batch size	train epochs	warmup epochs	weight decay
MSRVTT-Ret	1.2e-5	128	100	10	0.01
DiDeMo-Ret	1.2e-5	256	50	5	0.01
LSMDC-Ret	1.2e-5	256	50	5	0.01
MSRVTT-MC	1.2e-5	256	100	10	0.01
MSRVTT-QA	1.2e-5	256	20	2	0.01
LSMDC-MC	1.2e-5	128	20	2	0.01
LSMDC-FiB	1.2e-5	128	20	2	0.01
MSVD-QA	1.2e-5	128	40	4	0.01
TGIF-Action	5e-6	128	100	10	0.01
TGIF-Transition	5e-6	128	50	5	0.01
TGIF-Frame	1.2e-5	128	20	2	0.01

Table 1. Fine-tuning hyper-parameters of different tasks. *Ret* is short for retrieval task.

Method	MSRVTT				DiDeMo				LSMDC-MC	TGIF-Frame
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	Acc	Acc
Baseline	19.8	41.7	51.1	10	22.6	47.9	58.2	7	78.8	68.9
Clover w/o T	20.9	42.8	**52.9**	9	26.1	49.6	59.3	5	79.3	68.9
Clover w/o V	22.2	42.2	52.5	9	26.3	50.1	60.3	5	79.0	69.0
Clover	**23.4**	**43.3**	52.4	**9**	**26.4**	**51.1**	**61.3**	**5**	**80.7**	**69.7**

Table 2. The effects of Clover on different modalities for zero-shot retrieval and fine-tuning based video QA.

Method	MSRVTT-Zeroshot				TGIF-Frame
	R@1	R@5	R@10	MedR	Acc
InfoNCE	21.8	42.4	51.7	9	69.2
exclusive-NCE	**23.4**	**43.4**	**52.4**	**9**	**69.7**

Table 3. Effect of exclusive-NCE

Method	Batch Size	GPUs	GPU Hours
Clover	1024	64	1920
MCQ	800/2048	40	1000
VIOLET	-	-	2240
OA-Trans	1024	64	7680
All-in-one	2048	128	12288
MERLOT	1024	1024	30720

Table 4. Comparison of computing resources

one method is to use three InfoNCE objectives to operate the contrastive learning on $< V_e, T_e >, < V_e, T_m >, < V_e, M_{V_m f} >$ separately and then average the results. Another one is to use our proposed exclusive-NCE objective. Compared with InfoNce, our exclusive-NCE achieves better performance, which demonstrates the effectiveness of the proposed exclusive-NCE.

Comparison of computing resources In this paper, we have compared the computing resources of our proposed Clover with other models by analyzing three key factors:

Layers	#Params (Ratio)	MSRVTT-Zeroshot				TGIF-Frame
		R@1	R@5	R@10	MedR	Acc
3	28M (14%)	**23.4**	43.3	52.4	9	**69.7**
6	55M (27%)	22.2	43.4	**53.5**	8.5	69.2
12	110M (55%)	21.7	**43.6**	53.4	**8**	68.7

Table 5. Impact of # multi-modal layers.

batch size, the number of GPU, and the training GPU time. The results are presented in Tab. 4 Our Clover have a comparable batch size (1024) and less training time (1920 hours) than other models. This indicates that the superior performance of our Clover is due to the design of the method itself rather than simply having more resources or requiring longer training times.

Number of multi-modal encoder layers. We investigate the impacts of the number of multi-modal encoder layers in Tab. 5. We find that further increasing the size of multi-modal encoder does not bring significant performance improvement. Thus, we adopt a multi-modal encoder with 3 transformer layers for the trade-off between computation efficiency and performance.

Effect of Semantic Masking strategy and Focal Loss. We investigate the impacts of the Semantic Masking strategy and Focal Loss in Tab. 6. We find that both Semantic Masking strategy and Focal Loss imporve the Baseline model performance. We can see that only use Focal loss to

Method	MSRVTT-Zeroshot			
	R@1	R@5	R@10	MedR
Baseline	19.8	41.7	51.1	10
Baseline w/ SM	20.2	42.7	51.6	9
Baseline w/ Focal	19.9	42.3	51.7	9
Clover w/o Focal	22.9	43.0	**52.7**	9
Clover w/o SM	22.7	42.9	52.6	9
Clover	**23.4**	**43.3**	52.4	**9**

Table 6. Effect of Semantic Masking strategy and Focal loss

traditional MLM task bring marginal gain (compare line 3 with line 1), but with Semantic Masking strategy the performance gain is significant (compare line 4 with line 5), which means semantic masking strategy with focal loss did solve the class-inblanced problem in semantic word and improve the model performance.

C. More details of our training objectives.

In our paper, we introduced the training objective for tri-modal alignment $L_{TmA} = L_V + L_T$ and the detailed formulation of L_V. Similar to $L_V = L_v + L_{v'}$, we define the tri-modal alignment objective L_T w.r.t. the text modality. The objective are formulated as:

$$L_T = L_t + L_{t'}. \tag{1}$$

$$\begin{aligned} L_t = -\sum_{i=1}^{B} \Bigg[& \log \frac{e^{s(T_e^i, V_e^i)/\tau}}{e^{s(T_e^i, V_e^i)/\tau} + Z'} + \log \frac{e^{s(T_e^i, V_m^i)/\tau}}{e^{s(T_e^i, V_m^i)/\tau} + Z'} \\ & + \log \frac{e^{s(T_e^i, M_{Tmf}^i)/\tau}}{e^{s(T_e^i, M_{Tmf}^i)/\tau} + Z'} \Bigg], \text{ where} \\ Z' = \sum_{j \neq i}^{B} & \left[e^{s(T_e^i, V_e^j)/\tau} + e^{s(T_e^i, V_m^j)/\tau} + e^{s(T_e^i, M_{Tmf}^j)/\tau} \right], \end{aligned} \tag{2}$$

$$\begin{aligned} L_{t'} = -\sum_{i=1}^{B} \Bigg[& \log \frac{e^{s(V_e^i, T_e^i)/\tau}}{\sum_{j=1}^{B} e^{s(V_e^i, T_e^j)/\tau}} + \log \frac{e^{s(V_m^i, T_e^i)/\tau}}{\sum_{j=1}^{B} e^{s(V_m^i, T_e^j)/\tau}} \\ & + \log \frac{e^{s(M_{Tmf}^i, T_e^i)/\tau}}{\sum_{j=1}^{B} e^{s(M_{Tmf}^i, T_e^j)/\tau}} \Bigg], \end{aligned} \tag{3}$$

D. Addtional qualitative results of video-text retrieval and video question answering.

To illustrate the advantages of the Clover over the baseline method, we present some more examples of zero-shot video-text retrieval and video question answering in Fig. 1 and Fig. 2. These examples show that our clover method can focus on more fine-grained visual and text information, making it more precise to match or generate more accurate answers to the question.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 1

[2] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 1

[3] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 1

[4] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 1

[5] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019. 1

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 1

[7] 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. *arXiv preprint arXiv:2104.08860*, 2021. 1

[8] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893, 2017. 1

[9] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016. 1

[10] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 1, 4

[11] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 1, 4

[12] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 1

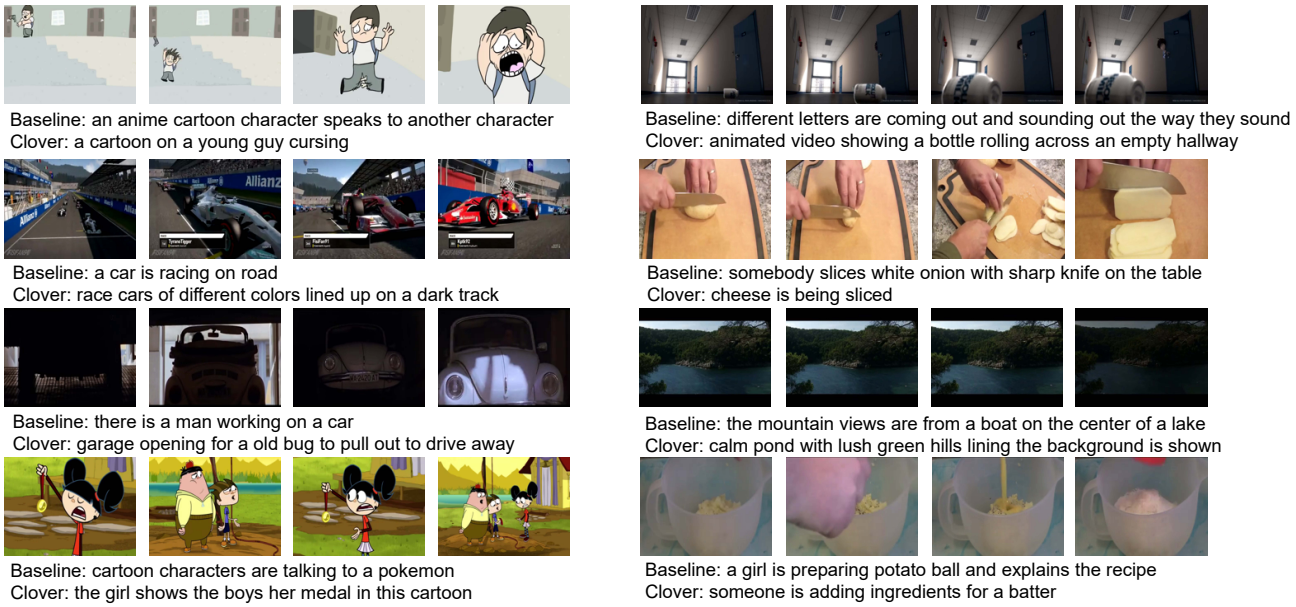

Figure 1. Qualitative results of zero-shot video-text retrieval results on MSRVTT [11].

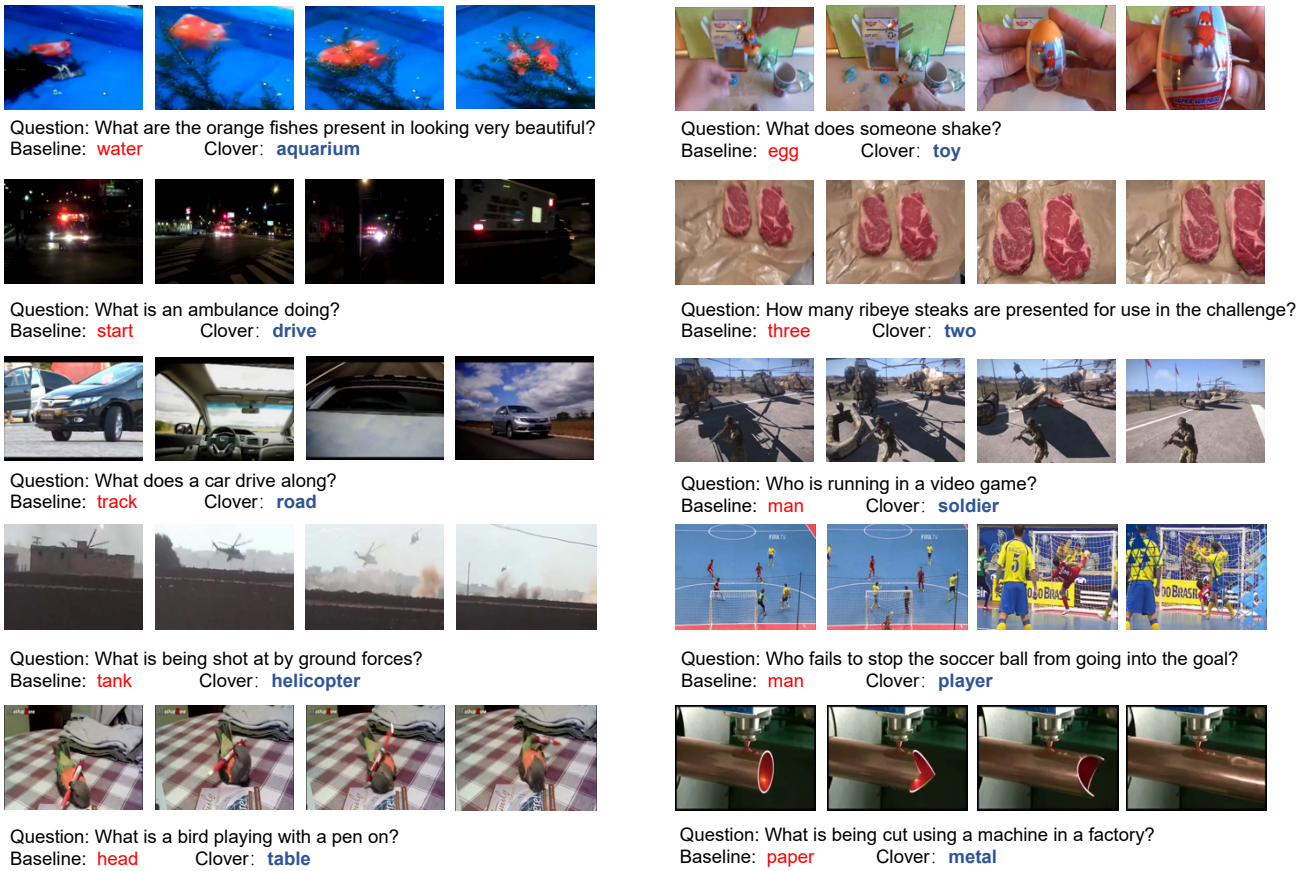

Figure 2. Qualitative results of video question answering results on MSRVTT-QA [10].