VLTinT: Visual-Linguistic Transformer-in-Transformer for Coherent Video Paragraph Captioning Supplementary Materials

Our proposed VLTinT consists of two main modules, i.e., VL Encoder and TinT Decoder and it is trained by the proposed VL contrastive loss function. The effectiveness of each module and VL contrastive loss have been quantitatively analyzed in the submitted main manuscript. In this supplementary, we first provide a qualitative analysis of each module and VL contrastive loss. We then present more qualitative results of VLTinT in video paragraph captioning (VPC).

The first module, VL Encoder, includes three modalities: (i) global visual environment, (ii) local visual main agents, and (iii) linguistic relevant scene elements. While the global visual environment feature is extracted by using C3D (Ji, Xu et al. 2010) backbone network pre-trained on Kinetics-400 (Kay, Carreira et al. 2017) as in other VPC approaches, our contribution towards the last two modalities, i.e., local visual main agents and linguistic relevant scene elements. The effectiveness of the last two modalities has been quantitatively analyzed in the submitted main manuscript (Tables 4 and 5). In this supplementary, we are going to provide further qualitative analysis of the last two modalities.

The second module, TinT Decoder, contains an inner transformer to model the intra-event coherency and an outer transformer to model inter-event coherency. The quantitative analysis of TinT Decoder has been included in the submitted main manuscript (Table 6), where we replace the outer transformer with an RNN-based network (Lei, Wang et al. 2020) to model the inter-event coherency. In this supplementary, we will provide some further qualitative analysis on the effectiveness of the inner transformer and outer transformer.

Besides qualitative analysis in the submitted main manuscript (Fig. 5), we further provide more qualitative VPC results conducted by VLTinT as in this supplementary.

To distinguish between the submitted main manuscript and the supplementary, Tables, Figures, and Equations in the submitted main manuscript will be mentioned with a bracket, i.e., ().

Analysis of Local Visual Main Agents

Our VLTinT utilizes Hybrid Attention Mechanism (HAM) to select main agents, who actually commit action in a video. Thus we investigate the effectiveness of HAM in VLTinT by comparing HAM with Soft-Attention (**?**) and Hard-Attention (Patro and Namboodiri 2018) as shown in Table 1.

Specifically, as in the submitted main manuscript (Eq.9), HAM is defined as follows:

$$\mathcal{H}_{\text{in}} = \mathcal{F}_{\text{in}} \oplus f_{\text{ref}} \tag{1a}$$

$$\mathcal{C} = \text{softmax}(||\mathcal{H}_{\text{in}}||_2) \tag{1b}$$

$$\mathcal{M} = \mathcal{C} > \frac{1}{N_{in}} \tag{1c}$$

$$f_{\text{out}} = g_\gamma(\mathcal{F}_{\text{in}} \odot \mathcal{M}) \tag{1d}$$

To conduct comparison in Table 1 we adjsut the above equations as follows:

For Soft-Attention, we remove Eq. 1a $\sim$ 1c and replace $\mathcal{M}$ in Eq. 1d by a vector of 1's, i.e., $f_{\text{out}} = g_\gamma(\mathcal{F}_{\text{in}})$.

For Hard-Attention, we replace $g_\gamma(\cdot)$ in Eq. 1d by an average pooling.

Furthermore, we illustrate the qualitative results of our proposed local visual main agents modality as shown in Fig.1. This example shows that our proposed modality, local visual main agents, can eliminate trivial agents while keeping the key agents who actually commit the action in the scene.

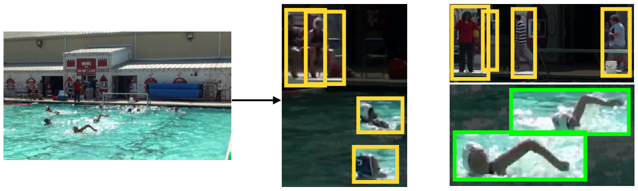
Table 1: Comparison between HAM and other attention mechanisms, i.e., soft attention (**?**) and hard attention (Patro and Namboodiri 2018), on ActivityNet Captions *ae-test*.

Attention	B@4↑	M↑	C↑	R↑	R@4↓
Soft-Att.	<u>14.34</u>	<u>17.85</u>	30.69	**36.74**	6.50
Hard-Att.	13.95	17.69	**31.13**	36.17	**4.21**
HAM (ours)	**14.50**	**17.97**	**31.13**	<u>36.56</u>	<u>4.75</u>

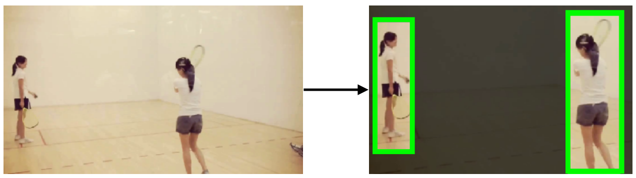
Analysis of Linguistic Relevant Scene Elements

In the linguistic relevant scene elements modality, the linguistic scene elements are first extracted by CLIP (Radford, Kim et al. 2021) and the most relevant ones are selected by HAM. Fig.2 first shows qualitative results from CLIP and then the most linguistic relevant scene elements by HAM. As shown in Fig.2, CLIP effectively captures both visual and non-visual scene elements. Among all scene elements captured by CLIP, part of them are actually relevant to the action; thus we utilize HAM to effectively select those scene elements.

Scene elements are often presented as objects in the scene. thus, we further compare the effectiveness of our CLIP &



(a)



(b)

(c)

Figure 1: Qualitative results of our local visual main agent modality. indicates main agents selected by our local main agent modality, and indicates eliminated trivial agents. Left: Input image. Right: selected and eliminated agents.

HAM against Mask R-CNN (He, Gkioxari et al. 2017) in extracting the most relevant scene elements. We observe that object detectors like Mask R-CNN can only extract a limited amount of visual scene elements, whereas CLIP provides much richer information on scene concepts including visual and non-visual scene elements. For example, given an image of people playing tennis as shown in Fig. 3, it is unfeasible to detect a small object such as a tennis ball using an object detector (He, Gkioxari et al. 2017). As shown in Fig. 3 (bottom), Mask-RCNN (He, Gkioxari et al. 2017) is only able to detect humans and a tennis racket while the tennis ball is not captured. Whereas, CLIP already encoded tennis scene elements including a tennis ball when modeling tennis games. As shown in Fig. 3 (top), CLIP captures a tennis ball and other related objects such as basket, court, fence, etc. Thus, we leverage CLIP (Radford, Kim et al. 2021) as a pre-trained model to extract linguistic information.

Analysis of TinT Decoder

Our TinT Decoder is designed as a nested transformer architecture where the inner transformer models the intra-event coherency and the outer transformer models the inter-event coherency. The submitted main manuscript (Table 6) shows quantitative analysis of TinT Decoder when replacing outer transformer with RNN-based network (Lei, Wang et al. 2020). The network architecture of captioning decoder in two cases, i.e., inter-event coherency is modeled by outer-transformer and inter-event coherency is modeled by RNN-based network is compared in Fig.4. With the same comparison settings, qualitative results are illustrated in Fig. 5. Besides some small captioning mistakes, the main issue with RNN-based inter-event coherency is repetitive patterns. That means the relationships between sentences cannot be addressed well by the RNN-based network. This also implies the advantages of our proposed TinT Decoder in modeling the inter-event coherency by the outer transformer and intra-event coherency by the inner transformer.

Qualitative Comparison

In this section, we present a qualitative analysis of VLTinT ActivityNet Captions as shown in Figure 6. For each sample video, we compare the descriptions generated from our VLTinT and ones generated by Vanilla Transformer (VTrans) (Zhou, Zhou et al. 2018) and MART (Lei, Wang et al. 2020). Overall, we observe our VLTinT can generate more descriptive captions such as ”He lassos a calf” in the first example and ”acoustic guitar” in the third example. We also noticed the accuracy of the caption generated by VLTinT. As in the second example, while VTrans and MART fail to capture the motion of taking contact lens out, VLTinT can correctly describe the scene.

Regarding to the caption repetitiveness, our model improved the inter-sentence diversity while maintaining a coherence. However, as shown in the first example, our model still suffers from some repetitive words and phrases within a sentence, suggesting further room for improvement on reducing the repetition in single sentence generation.

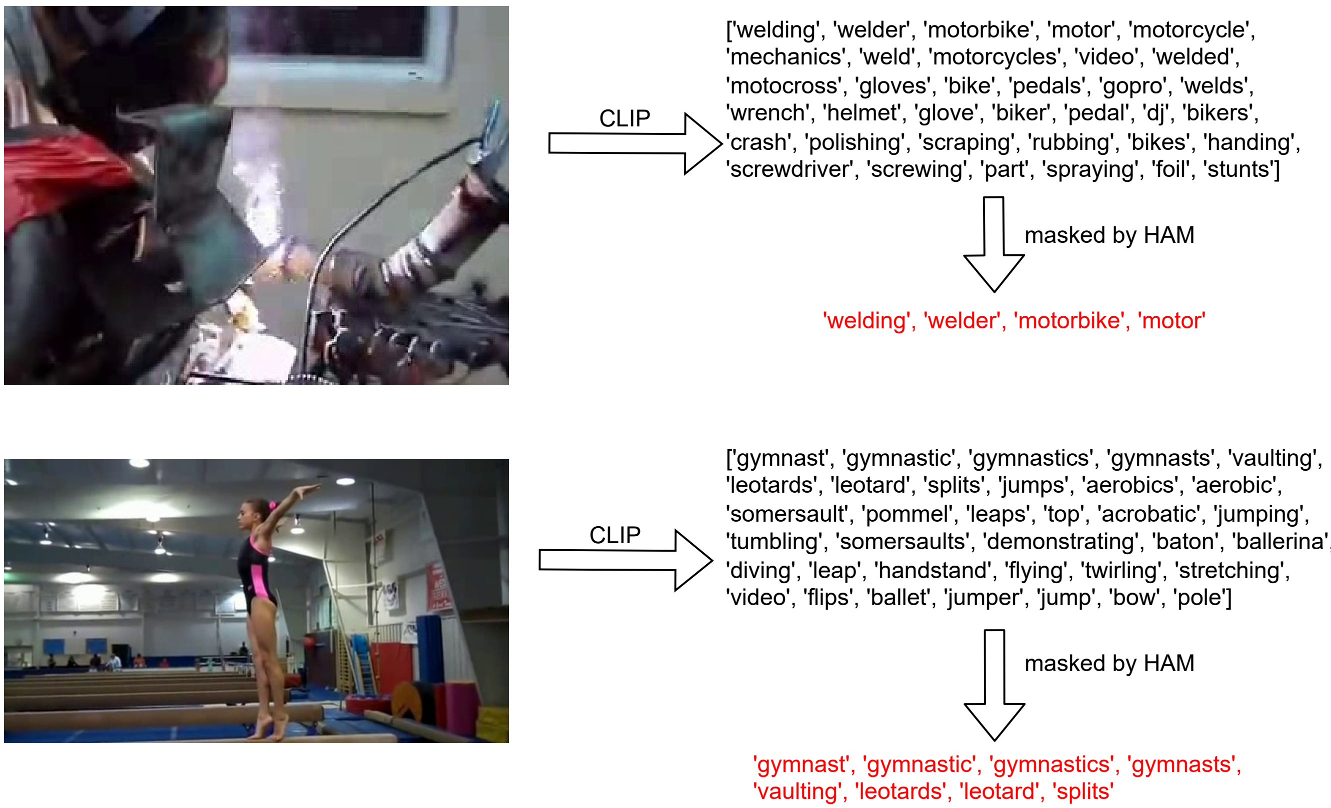


Figure 2: Qualitative examples of scene elements extracted by CLIP (black text) and then most relevant ones selected by HAM (red text).

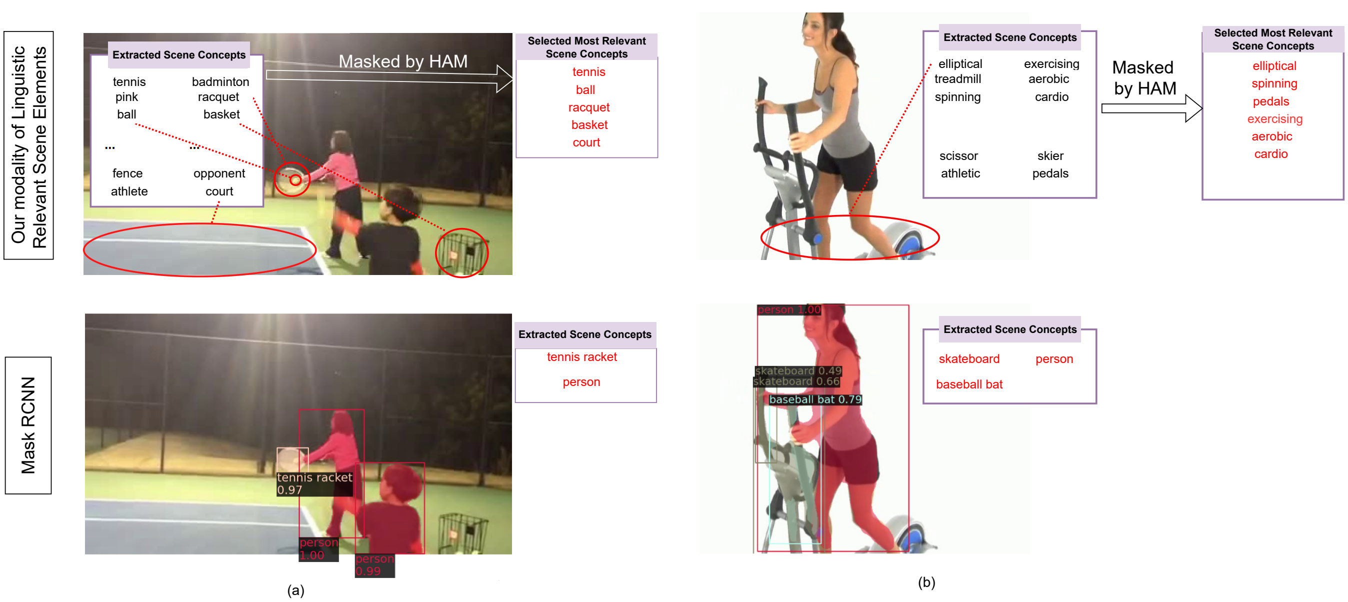


Figure 3: Qualitative examples of scene elements obtained by our proposed modality of linguistic relevant scenes element (top) vs. Mask-RCNN (He, Gkioxari et al. 2017) (bottom). In our proposed linguistic relevant scenes element (top), the scene elements obtained by CLIP (shown in black text) and then the most relevant ones selected by HAM (shown in red text).

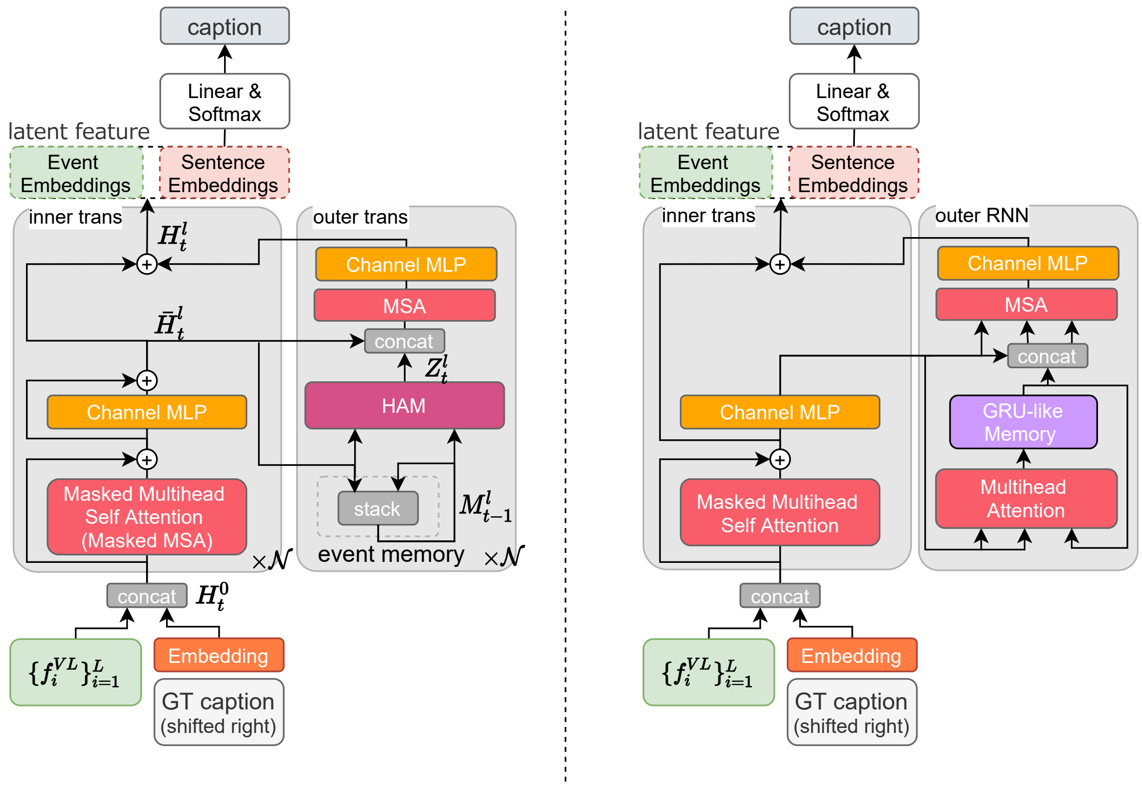


Figure 4: Architectural comparison of TinT Decoder in two cases: inter-event coherency modeled by the outer transformer (left) and by RNN-based network (Lei, Wang et al. 2020) (right)

References

Chen, T.; Kornblith, S.; et al. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Chen, Y.; Li, L.; et al. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV*, volume 12375, 104–120.

Chung, J.; Gulcehre, C.; et al. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *NIPS*.

Dai, Z.; Yang, Z.; et al. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *ACL*, 2978–2988.

Deng, C.; Chen, S.; et al. 2021. Sketch, Ground, and Refine: Top-Down Dense Video Captioning. In *CVPR*, 234–243.

Denkowski, M.; and Lavie, A. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Workshop on Statistical Machine Translation*, 376–380.

Dosovitskiy, A.; Beyer, L.; et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Ging, S.; Zolfaghari, M.; et al. 2020. COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning. In *NIPS*.

Han, K.; Xiao, A.; et al. 2021. Transformer in Transformer. In *NIPS*, volume 34, 15908–15919.

He, K.; Gkioxari, G.; et al. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Iashin, V.; and Rahtu, E. 2020. Multi-modal dense video captioning. In *CVPRW*, 958–959.

Ji, S.; Xu, W.; et al. 2010. 3D Convolutional Neural Networks for Human Action Recognition. In *ICML*, 495–502.

Kay, W.; Carreira, J.; et al. 2017. The kinetics human action video dataset. *ArXiv preprint*, abs/1705.06950.

Krishna, R.; Hata, K.; et al. 2017. Dense-Captioning Events in Videos. In *ICCV*, 706–715.

Lei, J.; Wang, L.; et al. 2020. MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning. In *ACL*, 2603–2614.

Li, Y.; Yao, T.; et al. 2018. Jointly Localizing and Describing Events for Dense Video Captioning. In *CVPR*, 7492–7500.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81.

Lin, T.-Y.; Maire, M.; et al. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Lupyan, G.; Abdel Rahman, R.; et al. 2020. Effects of Language on Visual Perception. *Trends Cogn Sci*, 24(11): 930–944.

v_laeOL4ipHck

RNN: A group of people are **on a beach playing volleyball**. They lob the ball back and forth over the net. They hit **the ball back and forth over the net**.

Trans: A group of people are **playing volleyball on a beach**. They lob the ball back and forth over the net. The game continues on with the ball, and the other teammates play.

GT: A group of girls are on a sandy beach. They are engaged in a game of volleyball. They lob the ball back and forth over the net.

v_aCknCFmU0sA

RNN: **A young woman** is seen sitting in front of a camera and begins brushing her hair. **She** then brushes **her hair** down and **begins brushing her hair**. **She** continues brushing the hair and looking off into the camera.

Trans: **A young man** is seen speaking to the camera while holding up a brush. **The man** then begins brushing **his hair** and **looking** back **to the camera**. **He** continues brushing his hair and looking off into the distance.

GT: A man with long hair is seen looking at the camera and begins brushing his hair. The man brushes his hair all around while still looking down at the camera. The man turns around to finish brushing his hair and ends by waving to the camera.

Figure 5: Qualitative analysis of inter-event modeling by RNN (the first row) and our outer transformer (the second row), whereas the groundtruth is shown in the last row. Red text indicates the captioning mistakes, purple text indicates repetitive patterns, and blue text indicates some distinct expressions.

Mun, J.; Yang, L.; et al. 2019. Streamlined Dense Video Captioning. In *CVPR*, 6588–6597.

Papineni, K.; Roukos, S.; et al. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*, 311–318.

Park, J. S.; Rohrbach, M.; et al. 2019. Adversarial Inference for Multi-Sentence Video Description. In *CVPR*, 6598–6608.

Pascanu, R.; Mikolov, T.; and Bengio, Y. 2013. On the difficulty of training recurrent neural networks. In *ICML*, volume 28, 1310–1318.

Pasunuru, R.; and Bansal, M. 2017. Multi-Task Video Captioning with Video and Entailment Generation. In *ACL*, 1273–1283.

Patashnik, O.; Wu, Z.; et al. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *ICCV*, 2065–2074.

Patro, B. N.; and Namboodiri, V. P. 2018. Differential Attention for Visual Question Answering. In *CVPR*, 7680–7688.

Radford, A.; Kim, J. W.; et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Rahman, T.; Xu, B.; and Sigal, L. 2019. Watch, Listen and Tell: Multi-Modal Weakly Supervised Dense Event Captioning. In *ICCV*.

Ren, S.; He, K.; et al. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, 91–99.

Shetty, R.; Rohrbach, M.; et al. 2017. Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training. In *CVPR*, 4155–4164.

Shi, B.; Ji, L.; et al. 2019. Dense Procedure Captioning in Narrated Instructional Videos. In *ACL*, 6382–6391.

v_PAGuZzrzSO4

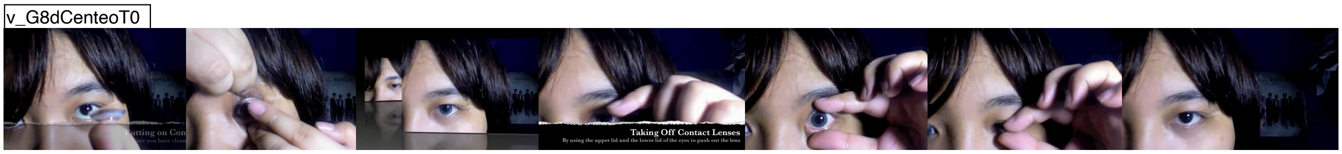
VTrans: A man is riding a horse down a river. The man then gets up and throws the calf down and grabs the horse and runs back to the horse. He gets back on his horse and gets back on his horse .

MART: A man is seen standing on a horse and throws a rope around. The man throws the calf down and the man chases after it. He ties the calf up and walks back to the horse .

VLTinT: A man is riding a horse in a rodeo ring. He lassos a calf. He ties the calf up and ties it up.

GT: A cowboy is riding a horse in a barn. He lassos a small calf. He dismounts, tying the calf and celebrating.

v_G8dCenteoT0



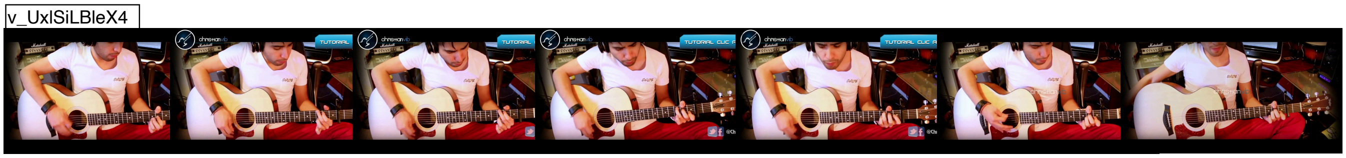
VTrans: The person then puts eye on the contact lens. The woman puts the contact lens in her eye. The person puts a contact lens in the eye.

MART: A woman is seen looking at the camera. She holds up a contact lens and puts it in her eye. She then puts the contact into the camera.

VLTinT: A close up of a eye is shown with a person's eye. A person is then seen putting a contact lens in her eye. The person then takes a contact lens out of her eye.

GT: A woman holds a contact lens on her finger. She puts the contact lens into her eye. She opens her eye with her fingers and takes the contact lens out.

v_UxlSiLBleX4



VTrans: A man is playing a guitar. He is playing the guitar. He stops playing the guitar .

MART: A man is seen sitting on a stool holding a guitar and playing a guitar. The man continues playing the guitar while the camera captures his movements. The man finishes the song and smiles .

VLTinT: A man is sitting down playing an acoustic guitar. He is playing the guitar. He finishes playing the guitar and smiles .

GT: A man is sitting down in a chair. He begins to play an acoustic guitar. He finishes playing the guitar and standing up.

Figure 6: Qualitative comparison on ActivityNet Captions *ae-test* split between our VLTinT and VTrans(Zhou, Zhou et al. 2018), MART (Lei, Wang et al. 2020). At each video, captioning from VTrans is in the 1^{st} row, MART is in the 2^{nd} row, our VLTinT is in the 3^{rd} row, and groundtruth (GT) is in the 4^{th} row. Red text indicates the captioning mistakes, purple text indicates repetitive patterns, and blue text indicates some distinct expressions. We compared our model with Vanilla Transformer (VTrans) and MART as baselines. GT indicates the groundtruth captioning.

Song, Y.; Chen, S.; and Jin, Q. 2021. Towards Diverse Paragraph Captioning for Untrimmed Videos. In *CVPR*, 11245–11254.

Vaswani, A.; Shazeer, N.; et al. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*, 4566–4575.

Vo, K.; Joo, H.; et al. 2021. AEI: Actors-Environment Interaction with Adaptive Attention for Temporal Action Proposals Generation. *BMVC*.

Vo, K.; Le, N.; et al. 2021. Agent-Environment Network for Temporal Action Proposal Generation. In *ICASSP*, 2160–2164.

Vo, K.; Yamazaki, K.; et al. 2021. ABN: Agent-Aware Boundary Networks for Temporal Action Proposal Generation. *IEEE Access*, 9: 126431–126445.

Wang, T.; Zhang, R.; et al. 2021. End-to-End Dense Video Captioning with Parallel Decoding. In *ICCV*, 6827–6837.

Wang, T.; Zheng, H.; et al. 2020. Event-centric hierarchical representation for dense video captioning. *IEEE Trans Circuits Syst Video Technol*, 31(5): 1890–1900.

Wang, X.; Wu, J.; et al. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language. In *ICCV*, 4580–4590.

Wu, Z.; Xiong, Y.; et al. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 3733–3742.

Xiong, Y.; Dai, B.; and Lin, D. 2018. Move Forward and Tell: A Progressive Generator of Video Descriptions. In *ECCV*, volume 11215, 489–505.

Yang, B.; and Zou, Y. 2021. CLIP Meets Video Captioners: Attribute-Aware Representation Learning Promotes Accurate Captioning. *ArXiv preprint*, abs/2111.15162.

Zhang, Z.; Zhang, H.; et al. 2022. Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding. In *AAAI*, volume 36, 3417–3425.

Zhou, L.; Kalantidis, Y.; et al. 2019. Grounded Video Description. In *CVPR*, 6578–6587.

Zhou, L.; Xu, C.; and Corso, J. J. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. In *AAAI*, 7590–7598.

Zhou, L.; Zhou, Y.; et al. 2018. End-to-End Dense Video Captioning With Masked Transformer. In *CVPR*, 8739–8748.