Align and Attend: Multimodal Summarization with Dual Contrastive Losses
Supplementary Material

Bo He[1]*, Jun Wang[1], Jielin Qiu[2], Trung Bui[3], Abhinav Shrivastava[1], Zhaowen Wang[3]

[1]University of Maryland, College Park [2]Carnegie Mellon University [3]Adobe Research

{bohe,abhinav}@cs.umd.edu, junwong@umd.edu, jielinq@andrew.cmu.edu, {bui,zhawang}@adobe.com

Sec. 1 elaborates more details about the collection process of BLiSS dataset. Sec. 2 provides more dataset-specific implementation details and hyper-parameters for training and testing. We also present more qualitative results in Sec. 3. Finally, we discuss the limitation and some future work of our paper in Sec. 4.

1. BLiSS Dataset

Data Collection

We collected 674 livestream videos from behance.net, including the corresponding English transcripts and meta data. Audio tracks are also available in the videos; however, we don't use them in our study since most information from audio modality can be captured by transcripts. Meta data from the website are annotated by creators which include title, overall text description, creative fields, creative tools, streamer, cover image and animation.

The transcripts are first segmented by sentence, each with corresponding time stamps. We remove transcripts with very short duration which are likely caused by broken words and speech recognition failures. Based on the transcript segments, we further divide each video into 5-minute long clips, so that each clip has its corresponding frames and aligned transcript sentences.

Annotators were instructed to watch the entire clip, read the transcript sentences, and select keywords from the sentences representing the important content in the clip. Each clip has about 5 to 10 keywords. The sentences containing keywords are regarded as key sentences for extractive text summarization. The annotators were also asked to write a summary of the whole clip in their own words, which can be used for abstractive text summarization.

For each video, we extract all the frames from its thumbnail animation. For each frame g in the thumbnail, we select the most similar frame f in the video as the key-frame.

Corpus Statistics of BLiSS Dataset In Table 5, we compare the statistics of our collected BLiSS dataset with other datasets including standard video summarization datasets (SumMe [1] and TVSum [2]), multimodal datasets (CNN [3] and Daily Mail [3]) and the transcript summarization dataset ([4]). We can see that our BLiSS dataset has a much larger scale than all the other datasets. Specifically, the BLiSS dataset has 1,109 hours of total video duration. The total number of text tokens of the BLiSS dataset is 5.5M, much larger than the Daily Mail and StreamHover datasets.

Example We show one example of the annotated sample in the BLiSS dataset in Figure 5. We visualize the uniformly sampled video frames, annotated keyframes, sentence-level transcripts, and the abstractive text summary. Note that the extractive text summary is formed by the key sentences, where the ground-truth keywords in the key sentences are marked in blue color.

2. Experiment Details

On multimodal summarization datasets (Daily Mail and CNN), we train our A2Summ with a batch size of 4, a learning rate of 2e-4, weight decay of 1e-5, training epochs of 100, L = 2, the ratio controlling hard-negative samples $r = 16$, the balancing weights for dual contrastive losses β of 0.01 and 0, λ of 0.01 and 0 for the Daily Mail and CNN datasets, respectively.

On standard video summarization datasets (SumMe and TVSum), we train our A2Summ with a batch size of 4, a learning rate of 1e-3, weight decay of 1e-3 and 1e-5, training epochs of 300, number of transformer layers L = 2, the ratio controlling hard-negative samples $r = 16$ the balancing weights for dual contrastive losses β of 0.1, λ of 3 and 1 for the SumMe and TVSum datasets, respectively.

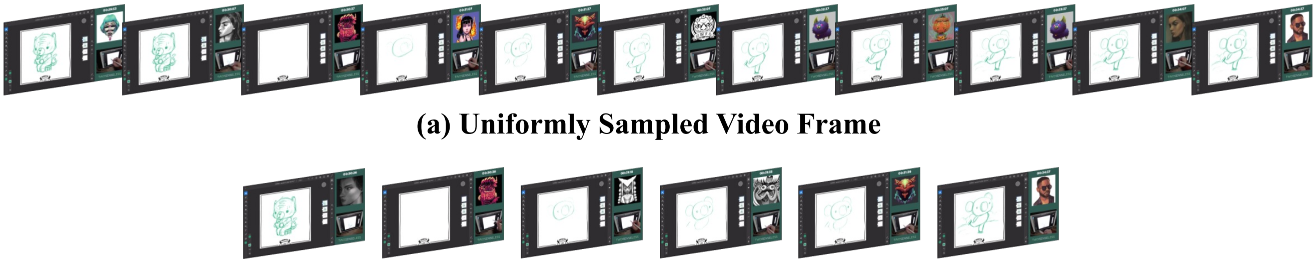
On the BLiSS dataset, we set a batch size of 64, a learning rate of 1e-3, weight decay of 1e-5, training epochs of 50, transformer layers L of 6, the ratio controlling hard-negative samples $r = 16$, the balancing weights for dual contrastive losses β of 0.001 and λ of 0.001.

We set the expansion size for both sides of key-frames and key-sentences in the contrastive pair selection procedure as 4 on all the datasets.

*Part of this work was done when Bo was an intern at Adobe Research.

Table 5. Statistics comparison of BLiSS dataset with other datasets.

	SumMe	TVSum	CNN	Daily Mail	StreamHover	BLiSS
Number of Data	25	50	203	1970	5421	13303
Total Video Duration (Hours)	1.0	3.5	7.1	44.2	452	1109
Total Number of Text Tokens	–	–	0.2M	1.3M	3.1M	5.5M
Avg. Video Summary Length	44	70	–	2.9	–	10.1
Avg. Text Summary Length	–	–	29.7	59.6	79	49



(a) Uniformly Sampled Video Frame

(b) Annotated Keyframe

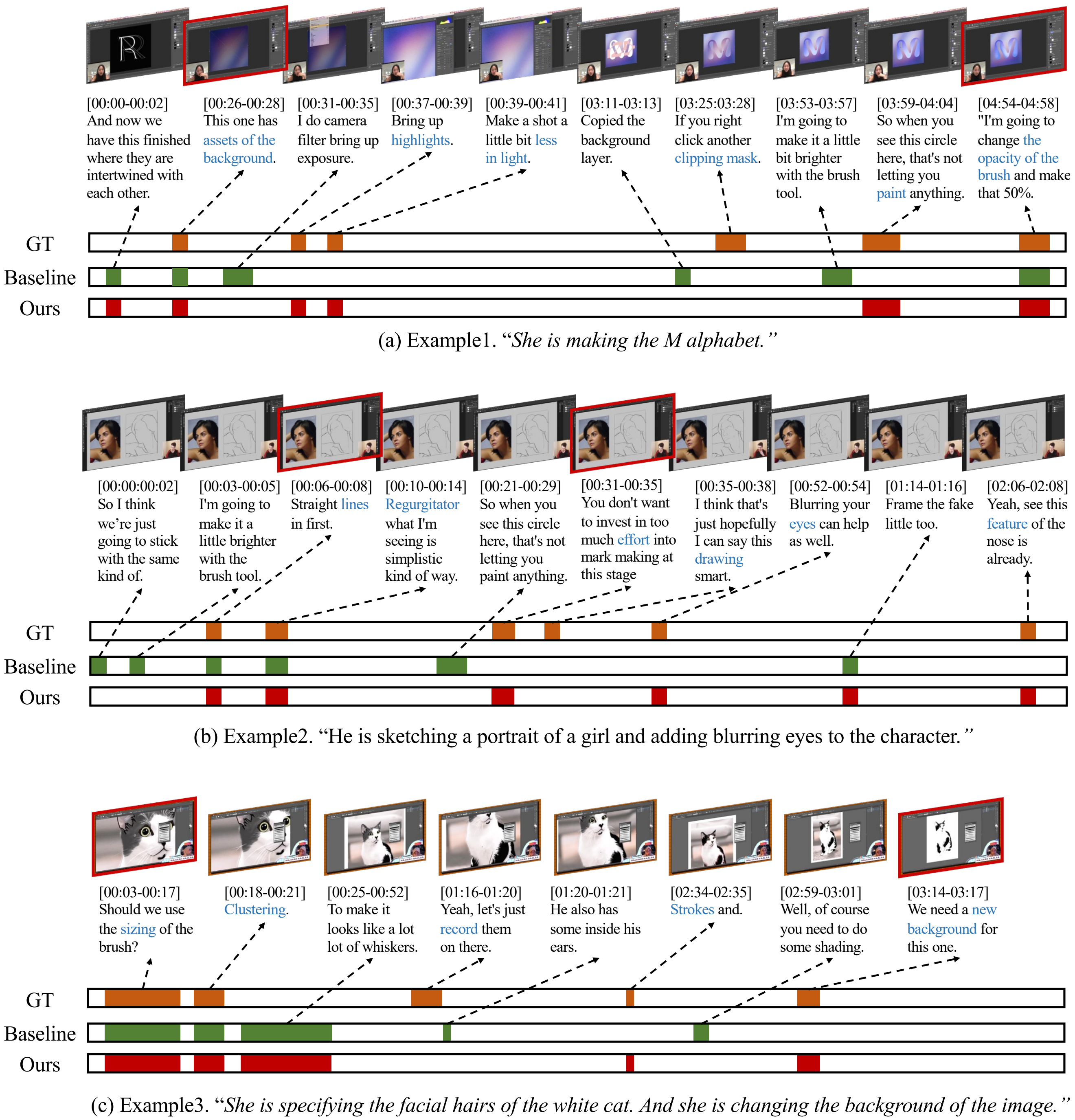
[00:12-00:15] That down so he has complete.
[00:15-00:24] I know you guys can't see the screen, but sometimes it's nice to see what I'm doing.
[00:24-00:34] Alright, let's see, let's see.
[00:38-00:42] Alright.
[00:42-00:44] I'm gonna draw koala.
[00:44-00:46] From memory is going to be terrible.
[00:48-00:49] Koala look like.
[00:51-00:52] Big head.
[00:53-00:55] Big nose.
[01:32-01:47] OK, so.
[01:47-01:52] Now actually the bigger question is how do I draw someone skating?
[01:52-02:40] Oh gosh.
[02:41-02:43] Wait, did you mean skating?
[02:43-02:46] Did you mean ice skating?
[02:46-02:48] I don't know why I first thought of ice skating when he said skating.
[03:24-04:08] Wait, I don't know if you're telling the truth.
[04:08-04:14] Turns out I have Frisco in my account, nice.
[04:14-04:17] It's still, um, they're still working on a lot of stuff.
[04:17-04:21] Cause there's I'm still part of the beta program.
[04:21-04:23] And they are continuing to add new features so.
[04:24-04:26] Still a baby program.
[04:29-04:34] But honestly, I I just love sketching and it just pencil is really nice.
[04:46-04:50] And it is nice if you use Photoshop and the other Adobe products.
[04:50-04:54] If you have creative Cloud, it saves automatically saves your creative cloud.
[04:55-05:00] So I can literally go into my computer on Photoshop and just pull it up and start.

(c) Sentence-level Transcript

She starts drawing an image of a koala with a big head and nose. She does this in a pencil-looking sketch.

(d) Annotated Abstractive Summary

Figure 5. Example of one data sample from the BLiSS dataset. Here, we visualize the uniformly sampled video frames, annotated keyframes, sentence-level transcript, and abstractive text summary. Note that the extractive text summary is formed by the key sentences, where the ground-truth keywords are marked with blue color. Best viewed in color.



[00:00-00:02] And now we have this finished where they are intertwined with each other.

[00:26-00:28] This one has assets of the background.

[00:31-00:35] I do camera filter bring up exposure.

[00:37-00:39] Bring up highlights.

[00:39-00:41] Make a shot a little bit less in light.

[03:11-03:13] Copied the background layer.

[03:25:03:28] If you right click another clipping mask.

[03:53-03:57] I'm going to make it a little bit brighter with the brush tool.

[03:59-04:04] So when you see this circle here, that's not letting you paint anything.

[04:54-04:58] "I'm going to change the opacity of the brush and make that 50%.

(a) Example1. "*She is making the M alphabet.*"

[00:00:00:02] So I think we're just going to stick with the same kind of.

[00:03-00:05] I'm going to make it a little brighter with the brush tool.

[00:06-00:08] Straight lines in first.

[00:10-00:14] Regurgitator what I'm seeing is simplistic kind of way.

[00:21-00:29] So when you see this circle here, that's not letting you paint anything.

[00:31-00:35] You don't want to invest in too much effort into mark making at this stage

[00:35-00:38] I think that's just hopefully I can say this drawing smart.

[00:52-00:54] Blurring your eyes can help as well.

[01:14-01:16] Frame the fake little too.

[02:06-02:08] Yeah, see this feature of the nose is already.

(b) Example2. "He is sketching a portrait of a girl and adding blurring eyes to the character.*"*

[00:03-00:17] Should we use the sizing of the brush?

[00:18-00:21] Clustering.

[00:25-00:52] To make it looks like a lot lot of whiskers.

[01:16-01:20] Yeah, let's just record them on there.

[01:20-01:21] He also has some inside his ears.

[02:34-02:35] Strokes and.

[02:59-03:01] Well, of course you need to do some shading.

[03:14-03:17] We need a new background for this one.

(c) Example3. "*She is specifying the facial hairs of the white cat. And she is changing the background of the image.*"

Figure 6. Visualization of multimodal summarization results for the BLiSS dataset. The ground-truth text summary, predictions from the baseline model and our A2Summ are shown for each video. "Baseline" denotes our A2Summ without the proposed alignment module and dual contrastive losses. The ground-truth keywords from key sentences are marked with blue color. We also show the corresponding video frames for each transcribed sentence where the frames with red boxes represent some of the predicted key-frames from our A2Summ. The title for each video clip is the annotated abstractive summary.

3. More Qualitative Results

In Figure 6, we show three different examples of multi-modal summarization results on the BLiSS dataset. We can see that, compared to the baseline method, our A2Summ can predict the key sentences more accurately and faithfully for the extractive text summarization task. It proves the effectiveness of the proposed alignment module and dual contrastive losses for the text modality. For the video summarization task, because livestream videos change slowly over time, their video frames generally share similar visual content. However, our predicted key-frames can still capture the important scenes from the input video qualitatively.

4. Limitation and Future Work

The main limitation is that our A2Summ is based on the Transformer [5] architecture with the self-attention operation, which suffers from heavy computation cost due to the quadratic computation complexity with respect to the input sequence length. Although there are a series of works [6–9] trying to design computation efficient transformer models to handle long sequences, it is out of the scope of our paper and we still follow the basic transformer design. In addition, the data annotation process for the video and text summaries is laborious. More research on unsupervised or self-supervised multimodal summarization tasks would be a good direction for future work.

References

[1] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014. 1

[2] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015. 1

[3] Xiyan Fu, Jun Wang, and Zhenglu Yang. Mm-avs: A full-scale dataset for multi-modal summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5922–5926, 2021. 1

[4] Sangwoo Cho, Franck Dernoncourt, Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, and Fei Liu. StreamHover: Livestream transcript summarization and annotation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6457–6474, 2021. 1

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[6] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 4

[7] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020. 4

[8] Gonçalo M Correia, Vlad Niculae, and André FT Martins. Adaptively sparse transformers. *arXiv preprint arXiv:1909.00015*, 2019. 4

[9] Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. Fast transformers with clustered attention. *Advances in Neural Information Processing Systems*, 33:21665–21674, 2020. 4