GIF Thumbnails: Attract More *Clicks* to Your Videos Supplementary File

Ablation Study

Effect of hyperparameter β.

We evaluate the effectiveness of hyper-parameter β in Eq. 5 via Monte Carlo sampling (10,000 samples per example). As shown in Fig. 1, a large β achieves better PME and PS on the training set, but leads to worse results on the validation/test set due to overfitting. On the contrary, model trained with a small β ignores the influence of latent variable z, and thus is hard to converge with multiple GIFs in the train set. Specifically, our method degenerates to a sequence-to-sequence model with an additional input noise z when $\beta = 0$. It can be observed from the 3rd figure in Fig. 1 that there are few modes of generated GIFs among 10,000 samples when $\beta = 0$, which means latent variable z has little impact on the model outputs. Based on the above discussion, we carefully choose $\beta = 4e^{-4}$ in the baseline comparison experiments.

Effect of learned prior.

A drawback for the model with fixed prior is that the generated GIFs ignore data-dependent GIF patterns, because all latent variable zs fed into *GIF Generator* are drawn from a fixed Gaussian $\mathcal{N}(0, 1)$. Indeed as shown in *Tab. 2 in the full paper*, the performance of the fixed prior is much worse than our method. Specifically, the model with a learned prior ($\lambda = 0$) outperforms that with a fixed prior by more than 27% improvement in terms of PME on both validation and test set.

Next, we try to analyze the model adaptability via the memorized pattern capacity of latent space for ground-truth GIFs in the training set. However, it is unreasonable to resolve the solution of z and the likelihood of ground-truth GIFs by freezing the network parameters and doing back-propagation as used in (Kohl et al. 2018). Instead, we adopt a dense grid search in the central area of latent space to traverse the potential GIFs with high probability. Concretely, we draw samples z from -3σ to 3σ with the stride of σ in all dimensions, which leads to 7^{10} samples for each video. We then define the memorized ratio as the ratio of ground-truth GIFs that can be recovered by one sample among all zs. The memorized ratio over videos with respect to different number of ground-truth GIFs is reported in Fig. 2. As can be seen, GIFs whose corresponding video has less ground-truth GIFs are easier to be memorized in latent space. Compared with fixed prior, learned prior fits the ground-truth distribution better due to the connection capability between GIF patterns and the video semantics.

Analysis of the latent space.

In this subsection, we show the guiding effect of latent variable z in selecting shots of videos. To demonstrate this, we draw the $z_0 - z_1$ plane of latent space where the generated GIFs are embedded. According to Fig. 3 and Fig. 5, there are six different kinds of GIF thumbnails. We can observe that samples in the left part prefer selecting shots with a guy in red, while samples in the right favor the guy in green. These cases indicate the latent variable z acts as a selector for video, whose location in latent space encodes guidance for GIF generation. Additional examples are described in Fig. 6.

Case Analysis

We qualitatively compare generated GIF thumbnails with image thumbnails in Fig. 4. Fig. 4(a) illustrates four image thumbnails created by video owners, and Fig. 4(b) shows some corresponding GIFs generated by our model.

1. The first one in Fig. 4(a) is a typical image thumbnail about delicious food sharing, which utilizes a close-up image of food to attract viewers. The corresponding GIF in the first row of Fig. 4(b) not only presents the close-up in the 3rd shot, but also displays the key process of food cooking.
2. Instead of selecting a frame in a video as a thumbnail directly, the 2nd video owner designed the image thumbnail by blending three fashionable outfits and putting a textual introduction. For generated GIF thumbnail, it takes advantage of the particular ability to tell the whole context across multiple shots. Specifically, the generated GIF thumbnail consists of multiple dress-ups in shots and gives the viewers a glance at the video content. For these videos, GIF thumbnails can achieve competitive performance against manually designed image thumbnails.
3. The third case is a short film, called *Facing it*. The generated GIF describes a scene that the guy's tongue falls on the grill. It is easy to raise audiences' sympathy and curiosity comparing with the raw image thumbnail.

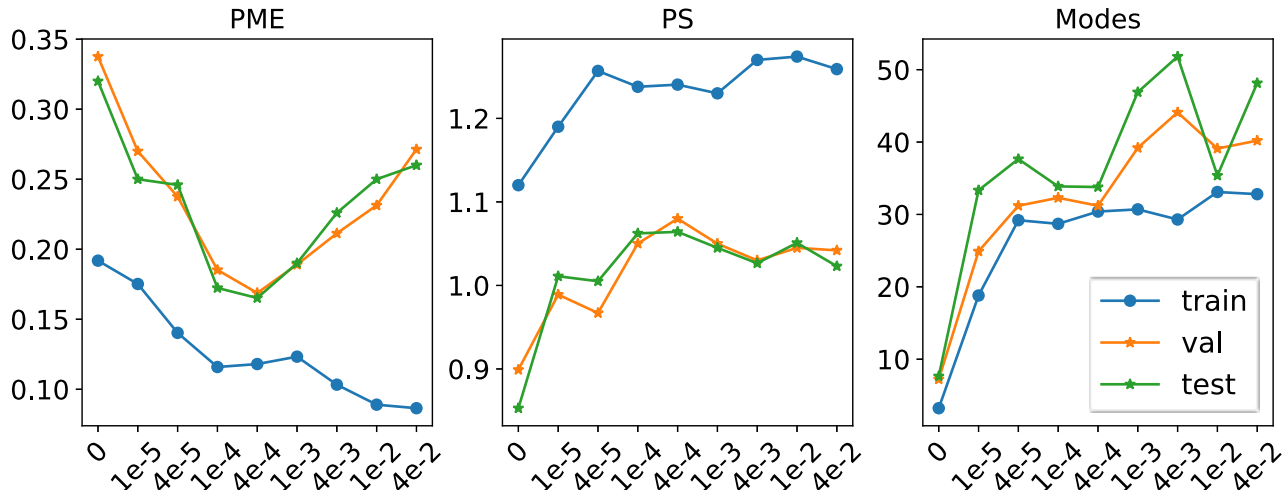


Figure 1: The effectiveness of β on GIF Thumbnail dataset. The first and second figures are the results of PME and PS metrics over different β on train/validation/test set. The third figure depicts the average distinct generated GIFs over 10,000 samples.

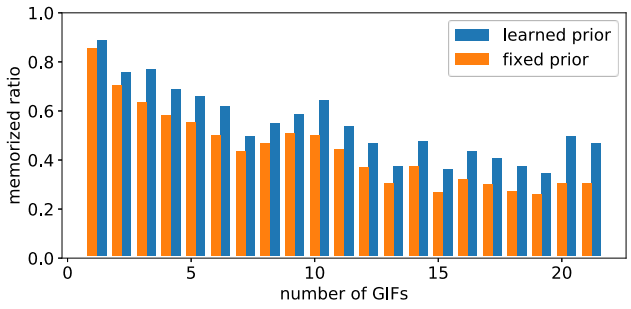


Figure 2: Ratios of ground-truth GIFs that are memorized in latent space after training phrase.

4. The fourth one is a type of video that most GIF proposals work better than the image thumbnails due to poor information capacity of images. A similar situation usually occurs in videos about dancing, because a series of actions are more persuasive than a single image.

In summary, these cases demonstrate the practical potential of GIF thumbnail generation and validate the advantage of our model over manually created image thumbnails. More cases can be found in the "Demo.mp4" file.

Pilot Application

Setting

During the application running period, we allocate similar web traffic of the platform for evaluated methods, and obey the following rules of the platform and its recommendation system:

1. We randomly select one of the videos in one page as a *test video* due to the high cost of network bandwidth for GIF thumbnail. Other videos always keep the original image

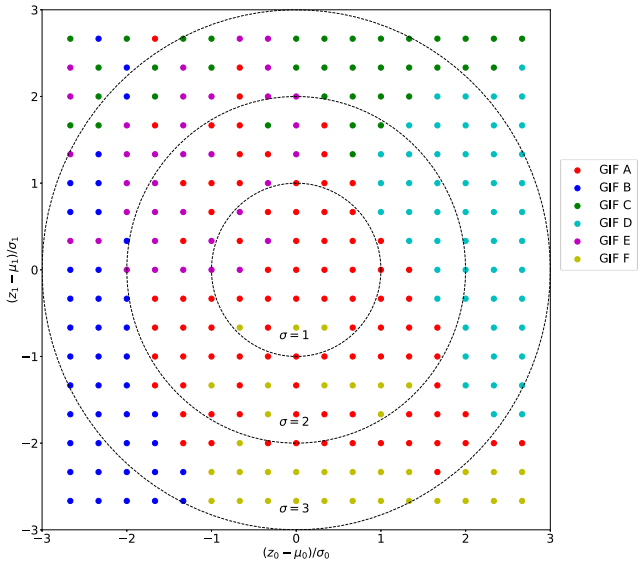


Figure 3: Visualization of $z_0 - z_1$ plane in latent space for the generated GIF. Each point denotes a generated GIF with corresponding z, and points in the same color corresponds to the same GIF. The dashed circles denote deviations from the mean. Detailed generated GIFs are listed in Fig. 5.

Method	PS
BeautThumb (Song et al. 2016)	0.941
Hecate (Song et al. 2016)	0.937
RankNet (Gygli, Song, and Cao 2016)	0.949
re-SEQ2SEQ (Zhang, Grauman, and Sha 2018)	0.963
GEVADEN, fixed prior	0.986
GEVADEN, $\lambda = 0$	1.025
GEVADEN, $\lambda = \lambda^*$	**1.039**

Table 1: Running results of a pilot application with an extended test set.

thumbnails, and PV and Click of them are not included in the experimental results.

2. Image thumbnail or all generated GIF thumbnails have an equal probability of being assigned to a *test video*;
3. Each video is recommended to a viewer at most once during the application running period.
4. Due to complex online strategies and various user preferences, some videos may be assigned more times than the others to the users.

Pilot Application Interface

We record the video platform interface on mobile devices in the "Demo.mp4" file.

Experimental Results

Besides CTR results, we also crawl human interaction information (time-sync comments) from Bilibili to evaluate the PS metrics for all methods. Experimental results are shown in Tab. 1. Results in Tab. 1 demonstrate that GEVADEN achieves the state-of-the-art performance of PS metric in the additional test set. Specifically, GEVADEN achieves at least 7.8% improvement over all baselines. High PS depicts that trimmed shots in generated GIF thumbnails attract more interest of viewers. In conclusion, both CTR and PS on the extended test set demonstrate the practical potential of our task and the proposed model.

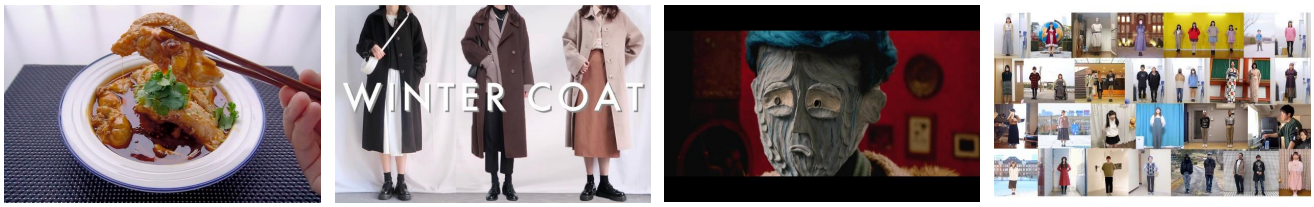
References

Gygli, M.; Song, Y.; and Cao, L. 2016. Video2gif: Automatic generation of animated gifs from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1001–1009.

Kohl, S.; Romera-Paredes, B.; Meyer, C.; De Fauw, J.; Ledsam, J. R.; Maier-Hein, K.; Eslami, S. A.; Rezende, D. J.; and Ronneberger, O. 2018. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 6965–6975.

Song, Y.; Redi, M.; Vallmitjana, J.; and Jaimes, A. 2016. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM)*, 659–668.

Zhang, K.; Grauman, K.; and Sha, F. 2018. Retrospective encoders for video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 383–399.

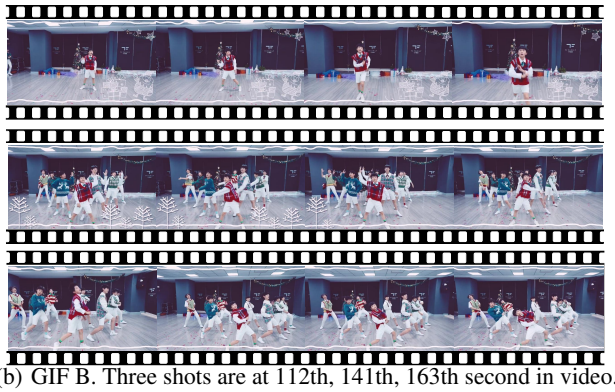


(a) Image thumbnails

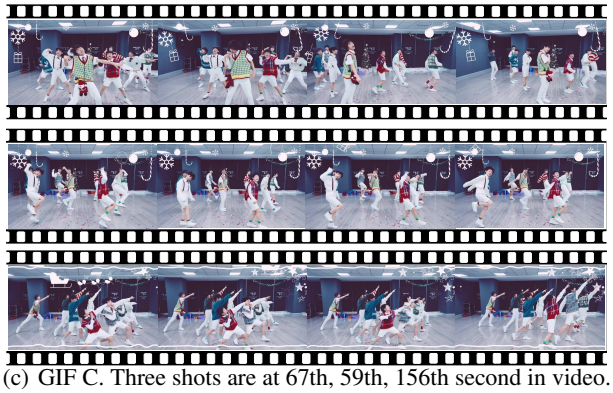
(b) GIF thumbnails generated by GEVADEN

Figure 4: Case study of image thumbnails and GIF thumbnails generated by GEVADEN.

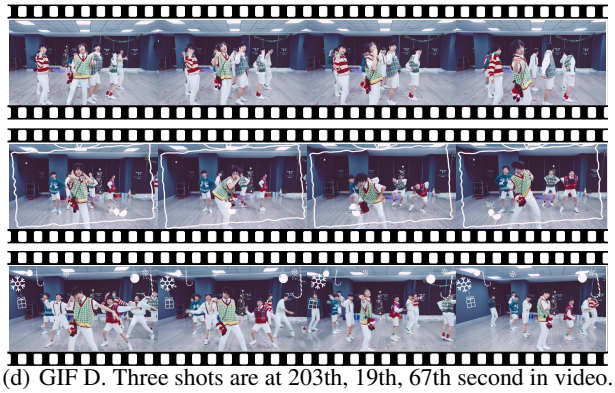
(a) GIF A. Three shots are at 16th, 67th, 156th second in video.



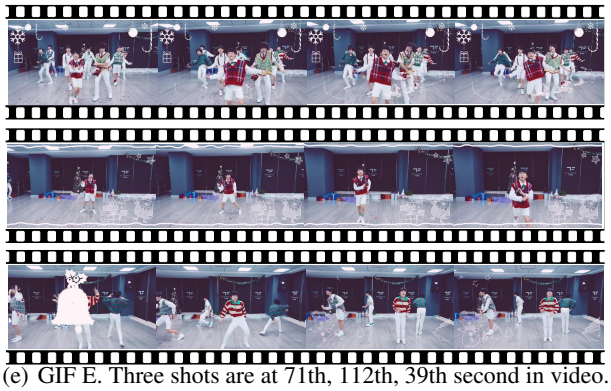
(b) GIF B. Three shots are at 112th, 141th, 163th second in video.



(c) GIF C. Three shots are at 67th, 59th, 156th second in video.



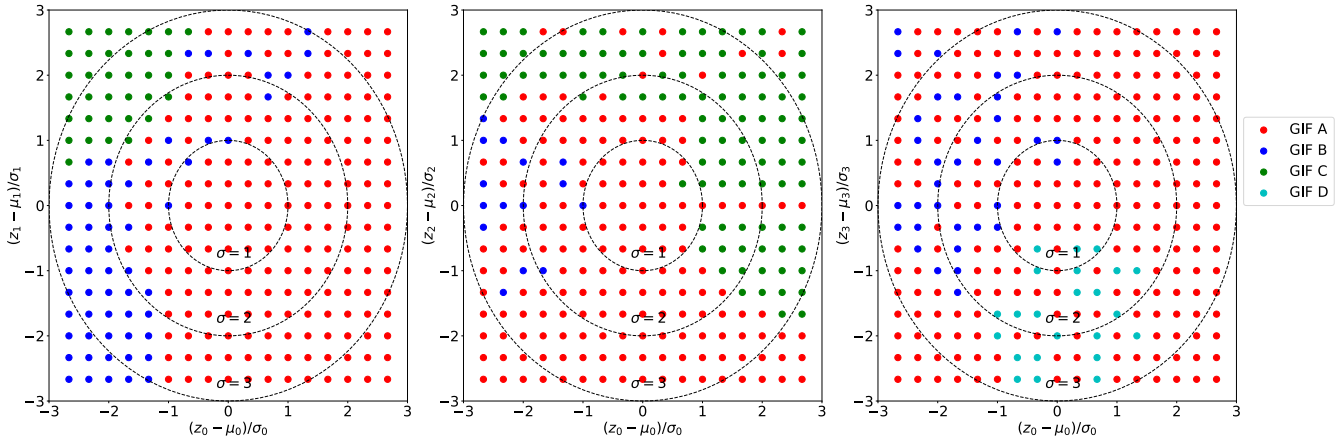
(d) GIF D. Three shots are at 203th, 19th, 67th second in video.



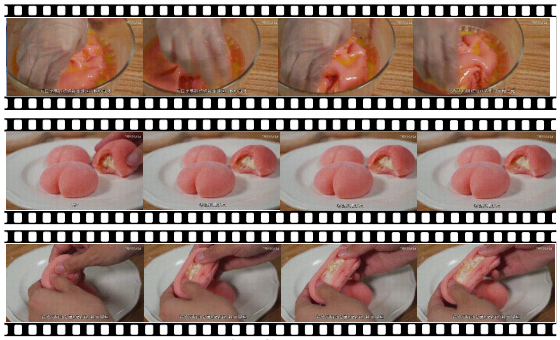
(e) GIF E. Three shots are at 71th, 112th, 39th second in video.

(f) GIF F. Three shots are at 16th, 156th, 71th second in video.

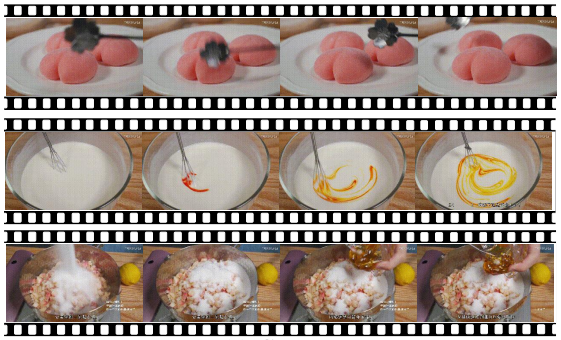
Figure 5: Case illustration for Fig. 3. Each GIF is described in three rows, and each row corresponds to a shot in GIF.



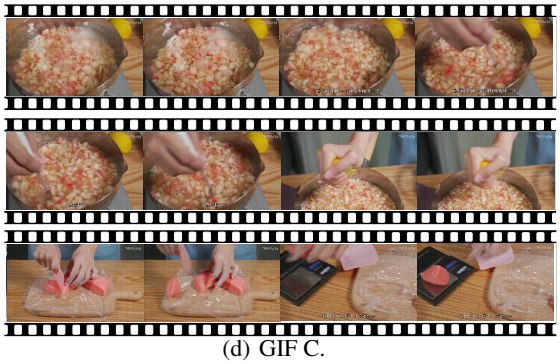
(a) Visualization of $z_0 - z_i$ plane in latent space for the generated GIF. The dashed circles denote deviations from the mean.



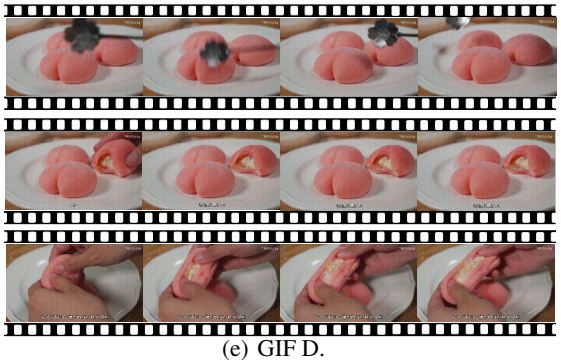
(b) GIF A.



(c) GIF B.



(d) GIF C.



(e) GIF D.

Figure 6: Case illustration for Latent space.