Supplementary:Context Guided Transformer Entropy Modeling for Video Compression

Junlong Tong[1,2] Wei Zhang[2] Yaohui Jin[1] Xiaoyu Shen[2]
[1]Shanghai Jiao Tong University
[2]Ningbo Key Laboratory of Spatial Intelligence and Digital Derivative, Institute of Digital Twin, EIT

A. Network Structure

Image encoder, decoder In this paper, we focus on conditional entropy modeling based on a well-designed frame codec. The frame codec employed in this work follows a contextual encoding backbone [3, 8], comprising a frame encoder, decoder, and a temporal context mining module. The frame encoder employs 4 strided convolutional layers achieving a 16× downsampling factor, while the frame decoder utilizes 4 strided transposed convolutions for upsampling to reconstruct the input frame. The temporal context mining module learns multi-scale temporal contexts directly from the propagated feature generated by the previously decoded frame $\hat{x}_{t-1}$ and the optical flow $\hat{v}_t$. Multi-scale contexts are used as a reference for both the encoder and decoder. The frame codec map the input frame to a sequence of tokens with dimension of 96 in latent space. Consistent with previous works [5, 9], we employ Anchor model [2] as the encoder for I-frames and set the GoP to 32. We train the frame encoder E and decoder D with a hyper-prior approach [1]. The well-trained model will serve as the basis for the context-guided transformer entropy model.

To demonstrate the generalization capability of our conditional entropy model, we additionally employ DCVC-DC, proposed by Li et al. [6], as the frame encoder backbone. The primary distinction of DCVC-DC with the above frame codec lies in the improvements to the context network, which differentiate it from our main experimental setup. For further details, please refer to Li et al. [6].

Transformer entropy model As described in the main text, the CGT model consists of temporal context resampler (TCR) and an attention-guided masked model. The attention-guided mask model consists of an encoder and a decoder based on a teacher-student network architecture. To maintain a concise and reusable structure, both the TCR and decoder employ the same architecture. The detailed structure is illustrated in Fig 1, where the core of these modules is the swin-transformer block. The swin-transformer block alternates between using window multi-head self-attention (Window-MSA) and shift-window multi-head self-attention (S-Window-MSA) layers. For TCR and the decoder, we introduce the window multi-head cross-attention (Window-MCA) and shift window multi-head cross-attention (S-Window-MCA) into the original Swin Transformer block to handle queries and key-value pairs from different sources. The cross-attention and self-attention mechanisms are applied alternately.

The input frame with $(256, 256, 3)$ is mapped to latent space with $(16, 16, 96)$ at first using the frame codec. Then the latents y_{t-1}, hyper-prior y_{h_p}, and temporal-prior y_{t_p} are utilized as temporal contexts. The temporal context such as y_{t-1} is resampled through TCR to obtain a corresponding compact representation. The scale of resampling is controlled by the shape of learnable quires, which is set to (1, 256, 768) in this paper. Then these temporal contexts are concatenated with the swin-transformer encoder to generate joint tokens, which are served as the key and value inputs of swin-transformer decoder.

B. Training and Inference

Training details The loss function of the CGT model is

$$\mathcal{L}_{\mathrm{RD}} = \underbrace{R\left(\hat{\boldsymbol{y}}_t\right) + R\left(\hat{\boldsymbol{z}}_t\right) + R\left(\hat{\boldsymbol{v}}_t\right)}_{\text{bit-rate}} + \lambda \cdot \underbrace{d\left(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\right)}_{\text{distortion}}, \quad (1)$$

where R is the bit-rate term, d is distortion term. We set coefficient λ to $256, 512, 1024, 2048$ for RD trade-off. We use straight-through estimator (STE) to enable gradient propagation through quantization operations during training.

Due to observed instability when training from scratch, this work adopts a three-stage training approach in this work. First, we train the frame encoder-decoder using adjacent frames consisting of I-frames and P-frames. This process follows the standard training procedure for image codec, utilizing hyperprior training with the RD loss function. This stage focuses primarily on the reconstruction performance of the model. We initialize part of parameters with the weights of DMC [5] and then fine-tune for 500K

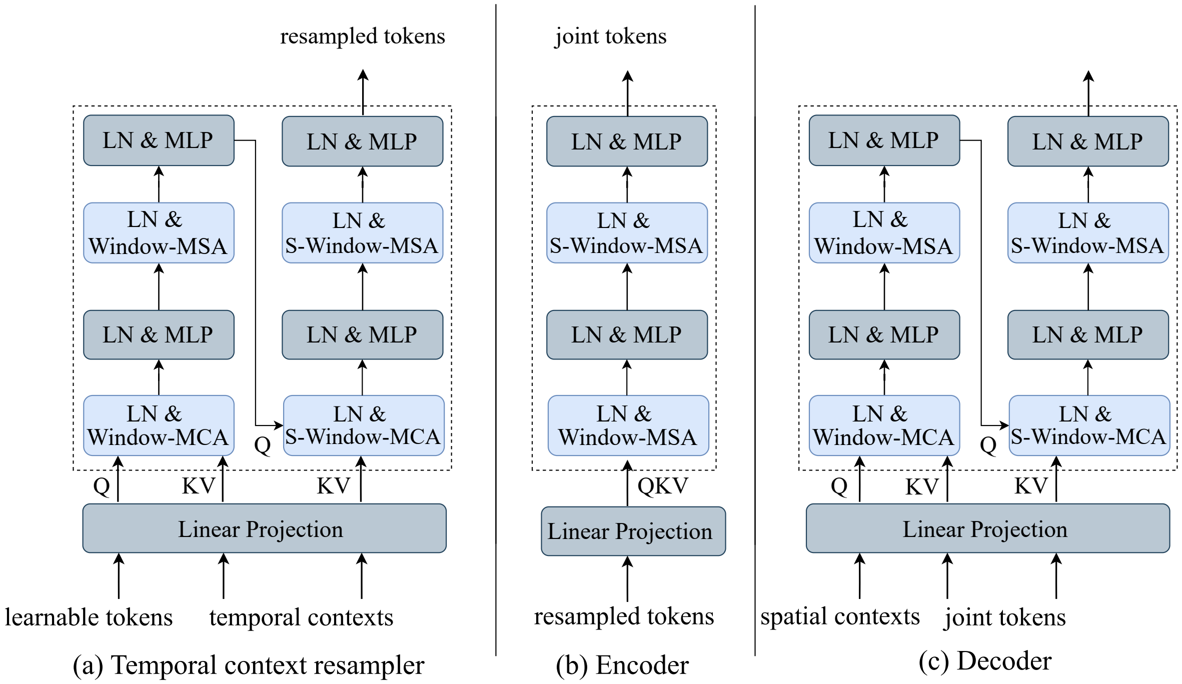

Figure 1. Detailed structure of the components in our CGT model. The TCR and decoder employ the same architecture for concise. In practice, the number of the temporal context resampler is set as 1, the swin transformer encoder have 2 blocks, and the swin transformer decoder have 4 blocks.

steps. This stage emphasizes the distortion performance of the model. In the second stage, we use three consecutive frames as model input and freeze the frame codec parameters, training only the CGT conditional entropy model for 1M steps. This stage continues to train the model using adjacent frames consisting of one I-frame and two P-frames. Finally, in the third stage, we incorporate the full video data and perform end-to-end fine-tuning to reduce the cumulative error between P-frames.

Analysis training paradigm In the main text, we analyzed the decoding processes of existing conditional entropy models and identified a common limitation: the lack of explicit modeling of spatial context dependencies. In this section, we further examine the training paradigms of different conditional entropy models, highlighting why our CGT model effectively captures and explicitly models the dependency within spatial context.

For autoregressive-based [7] and checkerboard-based [4] methods, the training and inference processes are strictly aligned, ensuring a consistent decoding strategy. However, this rigid structure prevents the model from differentiating the importance of contextual information at either stage, as it treats all available contexts equally without explicitly prioritizing more informative regions. MIMT [9] adopts a minimum-entropy principle for decoding and trains the model using a random masking proxy task. Since the model passively adapts to random masking during training while actively selecting the optimal path during inference, this approach, despite increasing training diversity, struggles to fully cover all possible minimum-entropy paths at inference. This limitation is particularly pronounced in complex videos with dynamically changing local features, where the optimal entropy-minimizing trajectory varies significantly.

We draw inspiration from the aforementioned training paradigms and introduce a teacher-student network along with a soft top-k strategy within the proxy task. During training, the teacher model selects the most contextually relevant regions for prioritized decoding, which then serve as the context for the student network's decoding process. This approach not only explicitly models the importance of spatial context but also ensures consistency between training and inference, bridging the gap between the two stages.

Training and inference of CGT model The training process of our CGT conditional entropy model is illustrated in Algorithm 1. And the inference process of our CGT conditional entropy model is illustrated in Algorithm 2.

Algorithm 1 Training process of CGT model

```
Input: $y_t$, temporal context $y_{t_p}, y_{h_p}, \hat{y}_{t-1}$.
Output: $\mu_t, \sigma_t$.
Resample $y_t^p, y_h^p$, and $\hat{y}_{t-1}$ using TCR;
Fuse the resampled temporal context using swin transformer encoder;
Apply random masking to current latent:$y_t^S = y_t + M$;
Perform cross-attention between the fused context and $y_t^S$ using teacher network;
Cross-attention calculation between fused context and $y_t^S$ using teacher network;
Generate attention map $A$ and entropy map $H$ based on cross-attention output, and perform dependency-weighted processing $\alpha H + (1 - \alpha)A$;
Apply soft top-k selection and remove the mask at selected positions;
Decode the entire latent according to the initially decoded context;
Obtain $\mu_t, \sigma_t$ through linear mapping.
```

Algorithm 2 Decoding process of CGT model

```
Input: Context $y_{t_p}, y_{h_p}, \hat{y}_{t-1}$, bit-stream, decoding step $n$.
Output: $\hat{x}_t$.
Decoded $\hat{v}_t$ and $\hat{z}_t$ from bit-stream;
Decoded the temporal-prior $y_{t_p}$ and the hyperprior $y_{h_p}$;
while $i < n$ do
    Resample $y_{t_p}, y_{h_p}$, and $\hat{y}_{t-1}$ using TCR;
    Fuse the resampled temporal context using the swin transformer encoder;
    Perform cross-attention between the fused context and $y_t^S$ using the swin transformer decoder;
    Generate attention map and entropy map based on cross-attention output;
    Perform dependency-weighted processing and apply top-k selection;
    Decode distribution of the top-k position;
end while
Decode the current frame using entire distribution parameters.
```

References

[1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[2] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7939–7948, 2020. 1

[3] Haifeng Guo, Sam Kwong, Dongjie Ye, and Shiqi Wang. Enhanced context mining and filtering for learned video compression. *IEEE Transactions on Multimedia*, 2023. 1

[4] Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14771–14780, 2021. 2

[5] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1503–1511, 2022. 1

[6] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22616–22626, 2023. 1

[7] Fabian Mentzer, George D Toderici, David Minnen, Sergi Caelles, Sung Jin Hwang, Mario Lucic, and Eirikur Agustsson. Vct: A video compression transformer. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:13091–13103, 2022. 2

[8] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *IEEE Transactions on Multimedia*, 2022. 1

[9] Jinxi Xiang, Kuan Tian, and Jun Zhang. Mimt: Masked image modeling transformer for video compression. In *International Conference on Learning Representations(ICLR)*, 2023. 1, 2