A Attention Analysis

To explore the attention mechanism of *dual-channel attention*, we visualize (1) the attention distribution in the temporal channel and (2) the scale factor α controlling the ratio between the spatial and temporal channel in equation 2.

Figure 8 visualizes the distribution among frames and texts in sequential generation (stage 1) with heat maps, where only 24 of 48 attention heads in 6 layers are shown for display purposes. The attention patterns can be broadly classified into the following categories:

- Most of the attention is on the text. E.g. the attention heads in violet.
- Most of the attention is on a certain frame. E.g. the attention heads in pink focus mainly on the previous frame, and the attention heads in blue focus mainly on the first frame besides the text, while the attention heads in yellow focus mostly on the frame itself.
- Attention is spread over several frames. E.g. the attention heads in green.

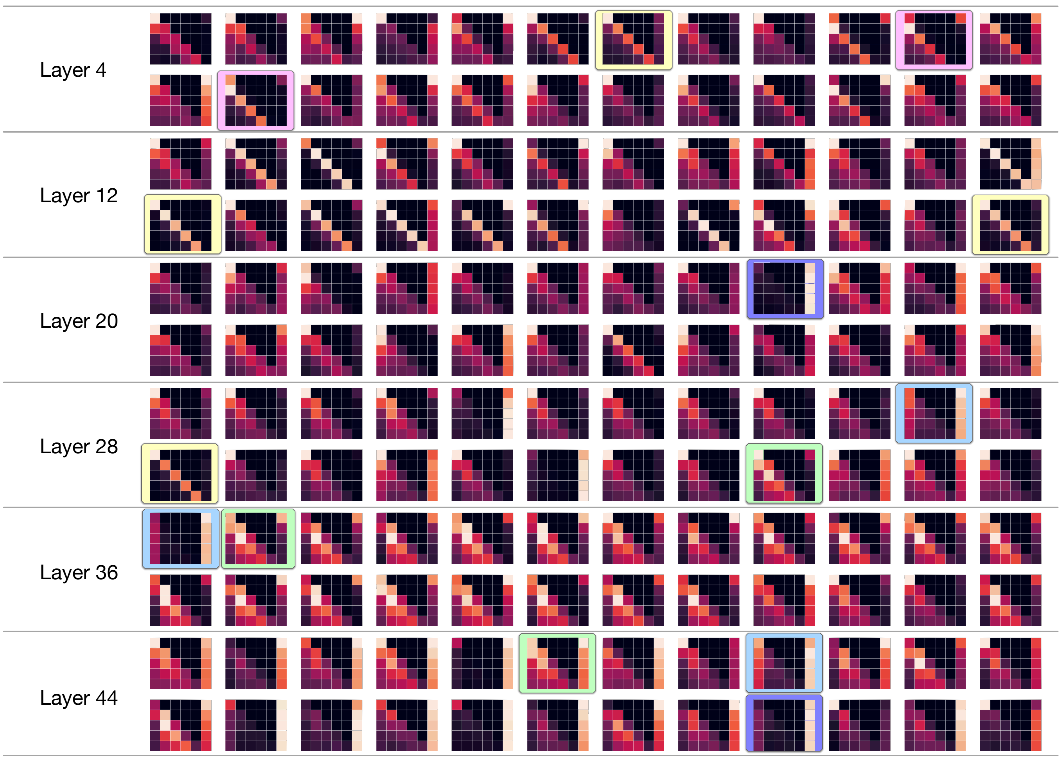

Figure 8: The attention distribution among frames and texts in sequential generation (stage 1). Only 24 of 48 attention heads in 6 layers are selected for display purpose. Each attention head is visualized with a heat map of size 5×6, where lighter color represents larger value. The 5×5 block on the left indicates the sum of attention scores (after softmax) between each pair of frames, and the rightmost column indicates the sum of the attention score of each frame to text. I.e. the grid in row i column j ($j \leq 5$) represents $\sum_{x \in F_i, y \in F_j} \mathrm{attn}_{x,y}$, and the grid in row i column 6 represents $\sum_{x \in F_i, y \in T} \mathrm{attn}_{x,y}$, where F_i, T denotes the set of tokens in the i-th frame and text respectively, and $\mathrm{attn}_{x,y}$ denotes the attention score of token x to y.

Some attention heads exhibit a single pattern, while others may exhibit a mixture of them. Attention heads in the same layer tend to show similar patterns. In lower layers (e.g. layer 4, 12) the heads tend to allocate attention according to position, while in higher layers more attention is allocated to text (e.g. layer 44) or spread over multiple frames. One possible explanation is that there are more high-level features in higher layers such as video semantics, by which the model can interact among more frames and texts to make high-level feature analysis.

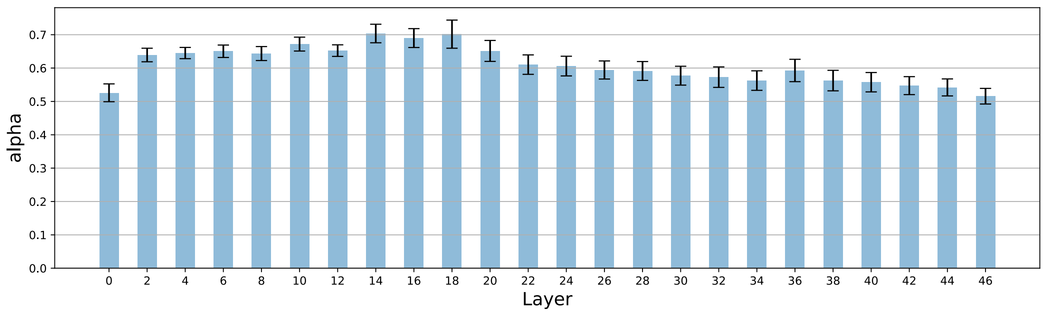

Figure 9: The scale factor α controlling the ratio between the spatial and temporal channel in equation 2 in dual-channel attention. Only α in half of the layers are shown for display reasons. As α is a vector of dimension 3072, we show the mean and variance among all of its dimensions in this figure.

It is worth noting that many heads do not allocate much attention to the frame itself which is important for inference, especially in higher layers. This shows that the CogVideo performs a certain degree of decoupling in the analysis of temporal and spatial features. While the spatial channel is in charge of feature analysis within the frame, the temporal channel can allocate more resources to explore relationships among different frames. We further illustrate this perspective with Figure 9, which shows that features calculated by CogView2 in the spatial channel are heavily relied on.

B Generated Video Samples

Thanks to the recursive interpolation model in stage 2, CogVideo is able to generate relatively high-frame-rate videos, as shown in Figure 10. We provide further examples generated by CogVideo in Figure 11. *The generated videos in mp4 format can be found in supplementary material, with filename "CogVideo_samples.mp4".* The length and the frame rate of provided videos are 4 seconds and 8 fps, respectively.

Figure 10: A 4-second video sample generated by CogVideo, which is firstly sequentially generated at 1 fps then recursively interpolated for 3 iterations.

C Training Details

CogVideo consists of two models corresponding to two stages, i.e. sequential generation and recursive interpolation. Both models have 7.7 billion parameters while 6 billion of them are fixed to CogView2, thus CogVideo has 9.4 billion different parameters in total.

CogVideo is trained on a dataset of 5.4 million captioned videos with a spatial resolution of 160×160 (can be upsampled to 480×480 by CogView2). Each model is pretrained separately. The stage-1 model is first pretrained for 76,000 iterations on video clips with a minimum frame rate of 0.25 fps, then trained for 15,000 iterations with a minimum frame rate of 1 fps. The stage-2 model is pretrained for 78,500 iterations with the frame rate of 2, 4, and 8 fps. Both models are trained in FP16 with batch size 416, and optimized by Adam with max learning rate $= 2 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay $= 1 \times 10^{-2}$.

D Details about Human Evaluation

In this section, we introduce more details about the human evaluation for measuring generation quality. The conduction of our human evaluation generally follows previous works including Ramesh et al. [19], Ding et al. [5]

We randomly extract 30 classes from UCF101 for video generation, using corresponding video samples in the dataset as ground truth items in the evaluation. Based on captions of selected classes, we generate video samples from models including TGANv2, VideoGPT, and our model, CogVideo. To further illustrate the effectiveness of hierarchical multi-frame-rate generation, we also include a 1-stage version of CogVideo model fine-tuned on Kinetics-600 which is described in § 5.3. For TGANv2, we use the official source code to train an unconditional generation model under the same setting as that in Saito et al. [21]. For VideoGPT, we use the official unconditional pretrained model to generate samples. To assign unconditionally generated samples into corresponding categories, we choose TSM[13] as the action recognition model and only select samples with confidence > 80%. A randomly selected subset of samples is displayed in Figure 12.

For each sample of the video mentioned above, we ask evaluators to give scores between 1 and 5 (5 indicates the best while 1 indicates the worst) from three aspects including frame texture, motion realism, and semantic relevance. Then the evaluators are required to give a general score of quality for each sample between 1 and 10, where a higher score indicates better quality. After video samples from each caption are all evaluated, the evaluators are asked to select the best one from them. We show snapshots of the evaluation website in Figure 13

Throughout the process of human evaluation, we invited nearly one hundred anonymous evaluators, while 90 of them completed the whole evaluation and were counted in the final results. None of the questions in the evaluation have any time limit. We offer each evaluator 75RMB as a reward for the evaluation. Results of the human evaluation, including the average score and standard deviation for each group, have already been introduced in Figure 5 in the main body. As ground truth samples take an absolute predominance in the best selection question, we have removed the part of ground truth samples in the selection pie plot for clearer model comparison.

阳光下，男孩在向日葵地里奔跑。
A boy is running in the sunflower field in the sunshine.

一个女人在街道上喝奶茶。
A woman is drinking milk tea in the street.

伤心哭泣的女人。
A sad woman who is crying.

一对夫妻在吵架。
A couple are quarreling.

两个快乐的朋友在打视频电话，谈笑风生。
Two happy friends are making video calls, talking and laughing.

一个男人背着包登山。
A man is hiking with a backpack.

一个男孩在海上冲浪。
A boy is surfing in the sea.

一个女人在海里骑马。
A woman is riding a horse in the sea.

一个男青年在喝咖啡。
A young man is drinking coffee.

一个男子在吃披萨。
A man is eating pizza.

一个愤怒的年轻女人在电话里尖叫。
An angry young woman is screaming on the phone.

一个男人戴着耳机跳舞。
A man is dancing with headphones on.

一只狗在草坪上奔跑。
A dog is running on the lawn.

一颗燃烧的心。
A burning heart.

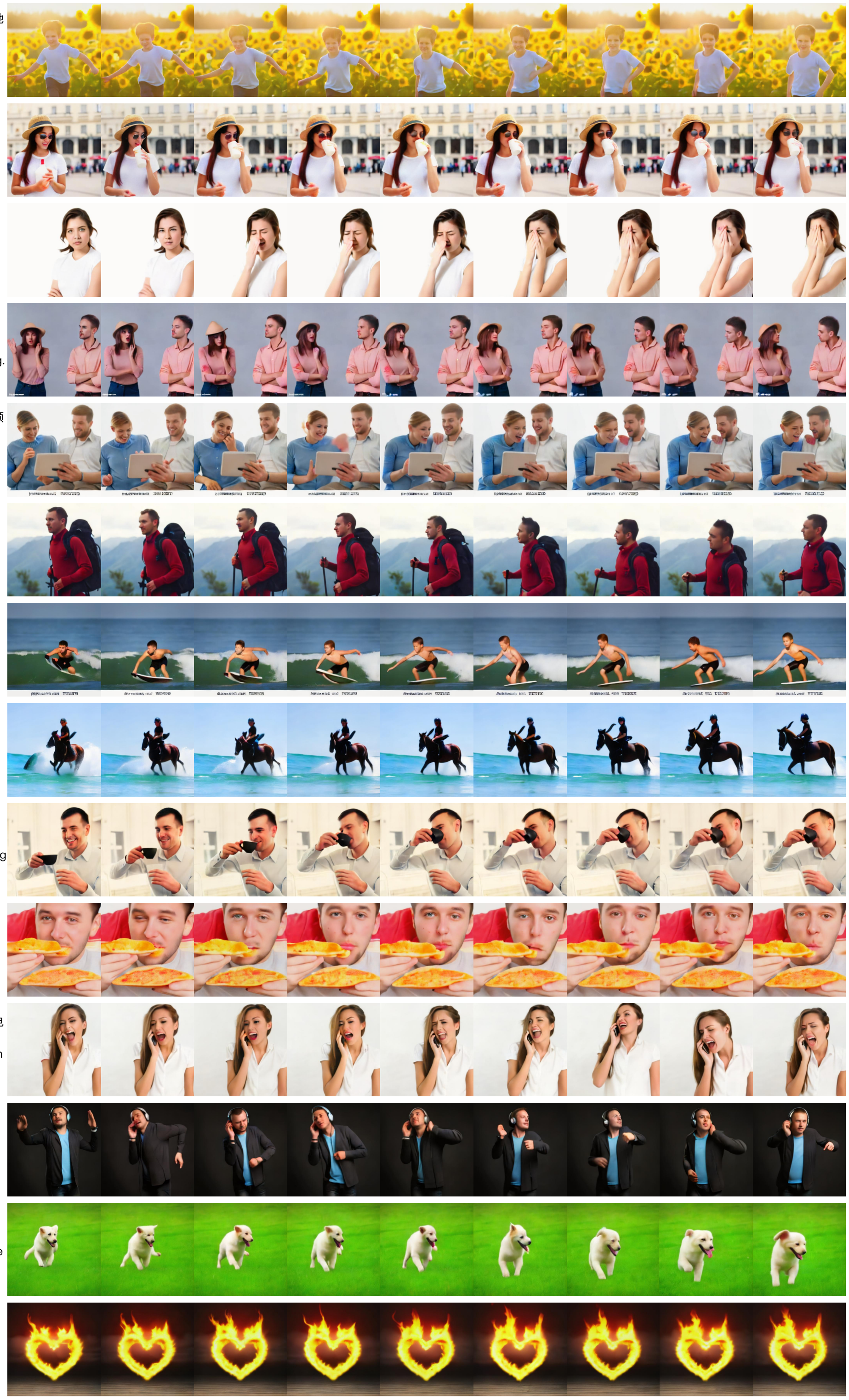

Figure 11: Further samples generated by CogVideo. The actual text inputs are in Chinese. Each sample is a 4-second clip of 32 frames, and here we sample 9 frames uniformly for display purposes.

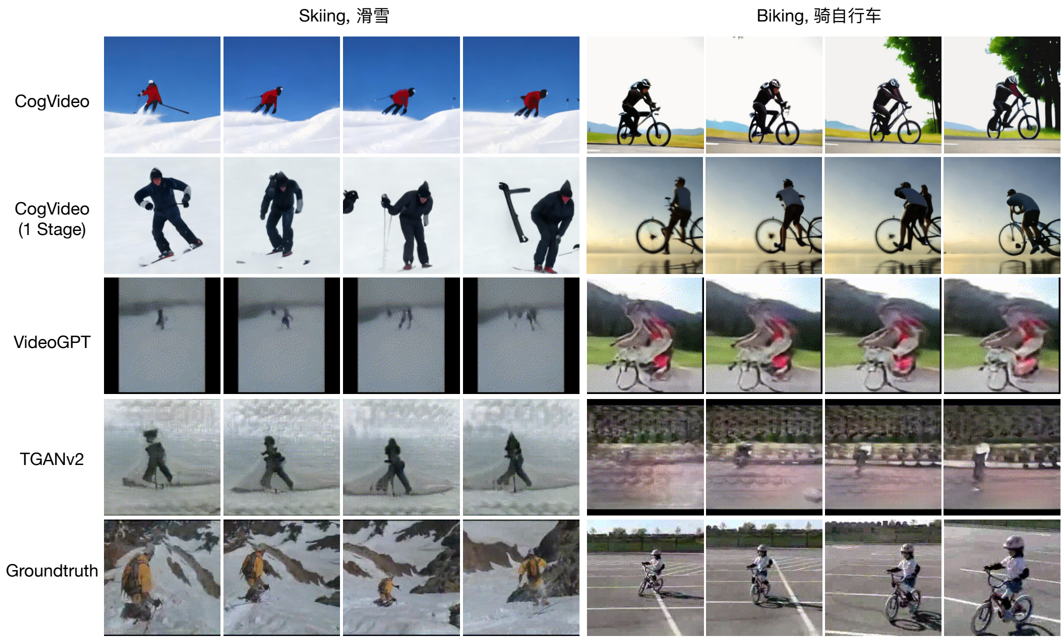

Figure 12: A subset of human evaluation samples. The captions are randomly selected from UCF-101. The original samples are clips of 16 frames, which are downsampled to 4 frames uniformly for display purposes.

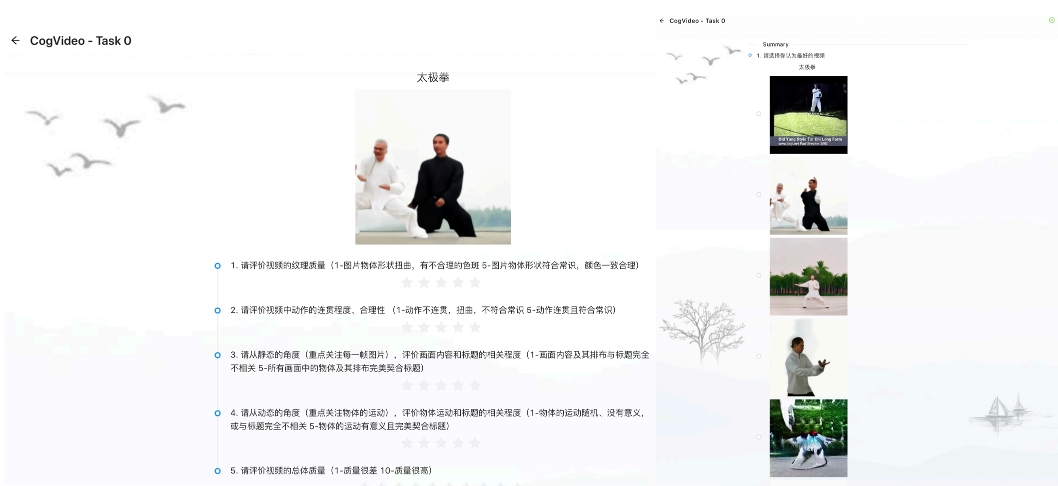

Figure 13: Snapshots of the evaluation website.