## Week 8 Thursday Paper 1: Sequential Modeling Enables Scalable Learning for Large Vision Models

## Explained by Sequential Modeling Enables Scalable Learning for Large Vision Models-CSDN博客

## Supplemental: Sequential Modeling Enables Scalable Learning for Large Vision Models

**1. What is the main idea presented in the paper?**
Ans: The paper introduces a novel approach to constructing Large Vision Models (LVMs) by using sequential modeling, enabling learning from visual data alone without relying on linguistic information. This method represents various types of visual data, including raw images and videos, semantic segmentations, and depth reconstructions, as "visual sentences" in a common format. The LVM is trained to predict the next token in a sequence, allowing it to understand and perform a range of vision tasks by creating suitable visual prompts at test time

**2. What is meant by image sequence?**
Ans: An image sequence refers to a sequence of images, such as frames from a video or different views of a 3D object, formatted as visual sentences. These sequences are used to train the model on temporal or spatial relationships within the data, enhancing its understanding of visual contexts

**3. How are single images handled in the paper?**
Ans: In the paper, single images are treated as the simplest form of visual sentences, consisting of an image followed by an EOS token. This format allows the model to be trained on individual images without additional context, making up the largest portion of the training data

**4. What dataset was used? What is the input to the network?**
Ans: The dataset used is called Unified Vision Dataset (UVDv1), containing 1.64 billion images or frames represented as 420 billion tokens. The input to the network are these images and sequences converted into a series of discrete tokens using a learned tokenizer, which maps each image to a string of 256 vector-quantized tokens

**5. Explain in detail the model architecture.**
Ans: The model architecture is a large transformer-based structure trained autoregressively on the visual sentences. It involves a two-stage approach where individual images are first converted into a sequence of visual tokens using a VQGAN encoder, and then these tokens are concatenated and fed into the transformer model to predict the next token in the sequence

**6. What is meant by image tokenization? Why is it important?**
Ans: Image tokenization is the process of converting each image into a sequence of discrete tokens using a visual tokenizer, specifically a VQGAN model in this paper. This step is crucial for transforming high-dimensional image data into a format suitable for sequential processing by the transformer model

**7. What is meant by sequential prompting, visual prompting, and analogy prompting?**
Ans: These terms refer to different methods of using visual prompts to guide the model in performing specific tasks,

**Sequential Prompting**: Using a sequence of images to predict the next image in the sequence.

**Visual Prompting**: Constructing a visual sentence that defines a task at test time, enabling the model to generate an appropriate output.

**Analogy Prompting**: Using a set of visual examples to form an analogy, prompting the model to apply learned patterns to new, unseen images or tasks

**8. What are the failure cases highlighted in the paper? How would you address them?**
Ans: The paper identifies several failure cases, such as task confusion, task entanglement, wrong instance prediction, and tokenizer limitations. These problems suggest that the model sometimes struggles with correctly interpreting the tasks or the visual content. Addressing these issues could involve improving the training data diversity, refining the tokenization process, and enhancing the model's ability to distinguish between different tasks and contexts

## Week 8 Thursday Paper 2: Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets

## Explained by

## Stable Diffusion and Friends - Generative Modeling in Latent…

## Supplemental: Stable Video Diffusion

**1. What is the central idea presented in this paper?**
Ans: The central idea of this paper is the development of Stable Video Diffusion, a latent video diffusion model designed for high-resolution, state-of-the-art text-to-video and image-to-video synthesis. The authors propose a systematic method for curating large datasets of videos to improve the quality of generated videos. They introduce a three-stage training regime for generative video models: text-to-image pretraining, video pretraining, and high-quality video finetuning, demonstrating that pretraining on well-curated datasets leads to significant improvements in video generation performance

**2. Explain in your own words how the input data for stable video diffusion was curated. Does this improve the performance?**
Ans: The curation of input data involves several steps aimed at improving the performance of the video diffusion model. Initially, a large dataset of videos is collected and processed to detect and eliminate cuts and fades. Each video clip is then annotated with captions using different methods. Further, the dataset is filtered based on metrics such as motion (using optical flow), text presence (using OCR), and aesthetic value (using CLIP embeddings). The paper argues that this meticulous data curation is crucial for training effective video models, as it ensures that the training data consists of high-quality, relevant video clips, leading to better performance of the final model

**3. Why is generative modeling done in the latent space?**

Ans: Generative modeling is done in the latent space to reduce computational complexity and improve the efficiency of the training process. In latent space, the high-dimensional input data is transformed into a lower-dimensional representation, making it easier to manipulate and learn from. This approach allows for faster training times and requires less memory, making it feasible to train models on large datasets

**4. What are the baselines used in the paper?**
Ans: The paper mentions various baselines used for comparison, including state-of-the-art methods like CogVideo, Make-A-Video, Video LDM, MagicVideo, and PYOCO. The performance of these models is compared based on metrics like FVD, showcasing the improvements made by the proposed Stable Video Diffusion model over these existing approaches

**5. What are the evaluation metrics used? Explain each metric in your own words.**
Ans: The evaluation metrics used include User Preference-comparing the quality and prompt alignment of generated videos, FVD-Fréchet Video Distance, a measure of similarity between the distribution of generated videos and real videos, and additional domain-specific metrics for tasks like multi-view generation. These metrics help assess the quality, relevance, and realism of the generated videos

**6. What are the three stages of video model training?**
Ans: The three stages of video model training are image pretraining, video pretraining, and video finetuning. Image pre-training equips the model with a strong visual representation. Video pre-training involves training the model on a large, curated video dataset, and video finetuning refines the model's performance on a smaller dataset of high-quality videos. This structured approach allows for incremental learning and improves the model's performance on video generation tasks

**7. What is the difference between GANs, Autoregressive models, and Diffusion Models?**
Ans: GANs involve training a generator to produce data that is indistinguishable from real data by a discriminator. Autoregressive models generate data sequentially, predicting each part of the data given the previous parts. Diffusion models, on the other hand, generate data by starting from noise and gradually refining it towards a sample from the data distribution through a process that reverses diffusion. These methods differ in their approaches to generating new data, with diffusion models typically offering a balance between the realism of GANs and the control of autoregressive models.

**8. What questions do you have about the paper?**

Ans: Is it ethical to publish these models for open use considering all kinds of potential issues it can create- criminal, modesty, etc.