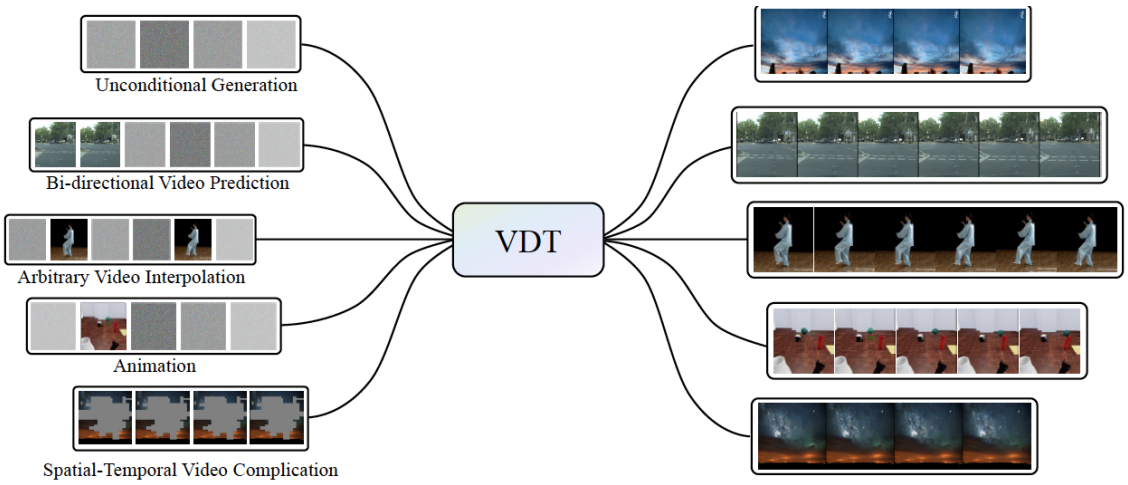
VDT: GENERAL-PURPOSE VIDEO DIFFUSION TRANSFORMERS VIA MASK MODELING

- アブスト
  - Video Diffusion Transformer (VDT)と呼ばれる拡散モデルベースの動画生成モデルにTransformerを導入したモデルを提案
  - temporal (時系列), spatial (空間的)なアテンションモジュールを導入することによって、これらの表現をよく学習できるようにした
  - spatial-temporal maskを導入することによって、様々な動画生成タスクの学習を可能にした
- イントロ
  - Transformerは拡散モデルで使用されるU-Netなどに比べると、長期の時系列情報をアテンションモジュールによって保持できる
  - また、Transformerはスケールする：スケールさせることで精度がどんどん良くなる
  - spatial-temporal maskによって、以下の動画生成タスクを可能にした



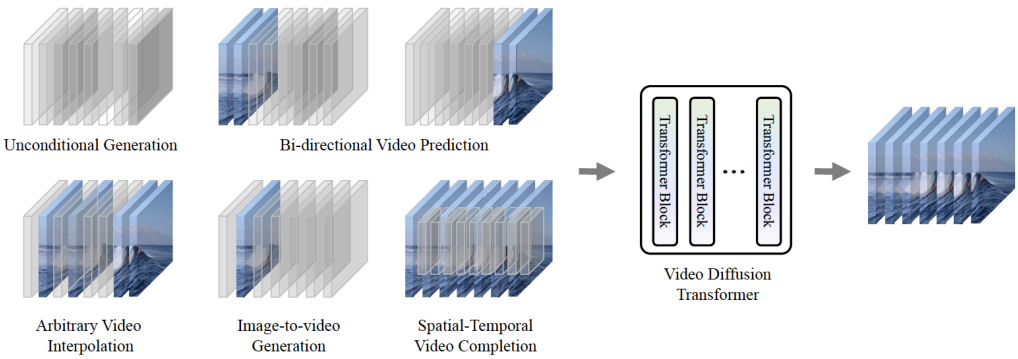
- 提案手法
  - 動画をそのままVDT入れると計算量が膨大になるので、学習済みのVAEによって特徴量してからVDTに入れる
    - H/8, W/8にする
  - ViTと同様に空間方向にN×Nのoverlapしないパッチに分ける
    - spatial and temporal positional embeddingを入れる
  - Transformer Blockに時系列情報を入れるために、Layer Normalization層に時系列情報を入れる

$$adaLN(h,t) = t_{scale} LayerNorm(h) + t_{shift},$$

  - VDT modelを動画予測用にする
    - 条件付けフレームをLayer Normalization層に入れる

$$adaLN(h,c) = c_{scale} LayerNorm(h) + c_{shift},$$

    - 条件付けフレームをCross Attentionで導入
      - 条件付けフレームをkeyとvalueに使用し、ノイズが乗っかったフレームをqueryとして使用する
      - 条件付けフレームとノイズが乗っかったフレームにspatial and temporal maskを入れることで以下の動画生成を実現



- 実験

- データセット
  - Unconditional Generation
    - UCF101, TaiChi, Sky Time-Lapse
  - Video Prediction
    - CityScapes, Physion
- 評価指標
  - FVD, SSIM, PSNR
- 条件付け方法
  - token concatenationが一番良かったので、それを使用

Table 2: Video prediction on Physion (128 × 128) conditioning on 8 frames and predicting 8. We compare three video prediction schemes.

Methods	FVD ↓	SSIM ↑	PSNR ↑
Ada. LN	270.8	0.6247	16.8
Cross-Attention	134.9	0.8523	28.6
Token Concat	129.1	0.8718	30.2

- 結果

Table 4: Unconditional video generation results on UCF-101. * means the model trained on the full dataset (train + test).

Method	Resolution	FVD ↓
GAN:		
TGANv2 (Saito et al., 2020)	16×128×128	1209.0
MoCoGAN* (Tulyakov et al., 2018)	16×128×128	838.0
DIGAN (Yu et al., 2022)	16×128×128	655.0
DIGAN* (Yu et al., 2022)	16×128×128	577.0
TATS (Ge et al., 2022)	16×128×128	420.0
Diff. based on U-Net with Pre:		
VideoFusion* (Luo et al., 2023)	16×128×128	220.0
Make-A-Video* (Singer et al., 2022)	16×256×256	81.3
Diff. based on U-Net:		
PVDM* (Yu et al., 2023)	16×256×256	343.6
MCVD (Voleti et al., 2022)	16×64×64	1143.0
VDM* (Ho et al., 2022b)	16×64×64	295.0
Diff. based on Transformer:		
VDT	16×64×64	**225.7**

Table 5: Video prediction on Cityscapes (128 × 128) conditioning on 2 and predicting 28 frames.

Cityscapes	FVD↓	SSIM↑
SVG-LP (Denton & Fergus, 2018)	1300.3	0.574
vRNN 1L (Castrejon et al., 2019)	682.1	0.609
Hier-vRNN (Castrejon et al., 2019)	567.5	0.628
GHVAE (Wu et al., 2021)	418.0	0.740
MCVD-spatin (Voleti et al., 2022)	184.8	0.720
MCVD-concat (Voleti et al., 2022)	141.4	0.690
VDT	**142.3**	**0.880**

Table 6: VQA accuracy on Physion-Collide.

Model	Accuracy
Object centric:	
Human (upper bound)	80.0
SlotFormer (Wu et al., 2022b)	69.3
Scene centric:	
PRIN (Qi et al., 2021)	57.9
pVGG-lstm (Bear et al., 2021)	58.7
pDEIT-lstm (Bear et al., 2021)	63.1
VDT (Ours)	**65.3**