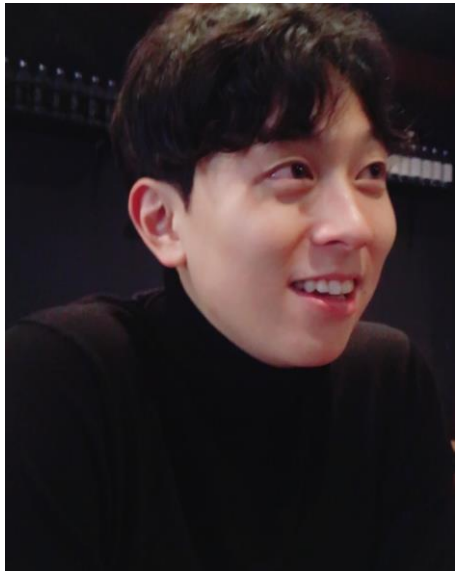
# STREAM: Spatio-TempoRal Evaluation and Analysis Metric for Video Generative Models

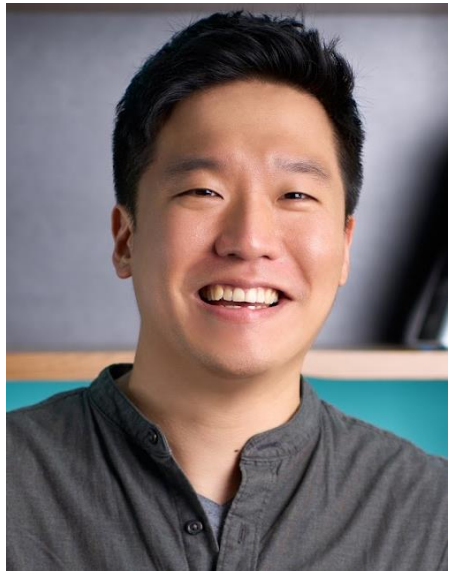



Pum jun Kim

UNIST

Seojun Kim

UNIST



Jaejun Yoo

UNIST


LAB. OF ADVANCED IMAGING TECHNOLOGY

UNIST AIGS ARTIFICIAL INTELLIGENCE GRADUATE SCHOOL

# Video Generative Models and Evaluation Protocol

## (a) Real video

## (b) Generated video

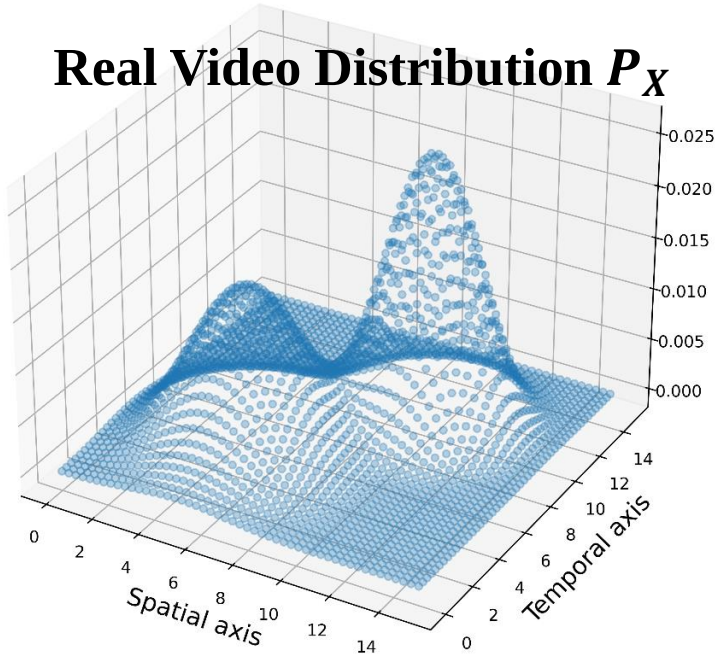
# Video Generative Models and Evaluation Protocol

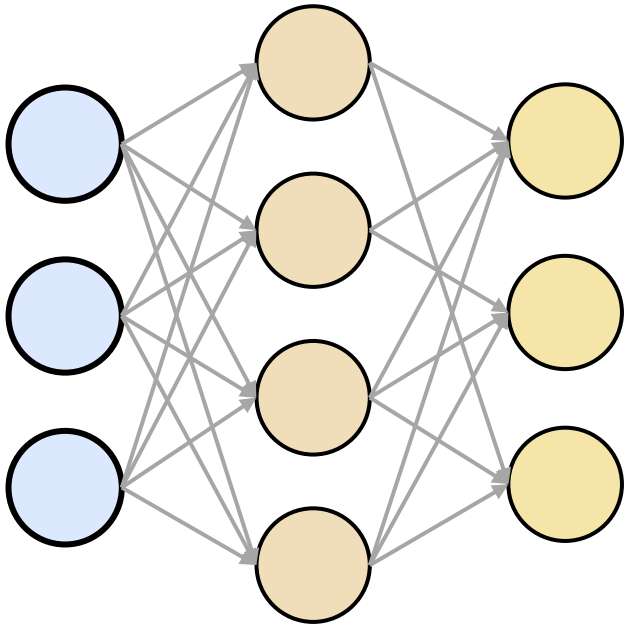
## (a) Real video

## (b) Generated video

# Video Generative Models and Evaluation Protocol

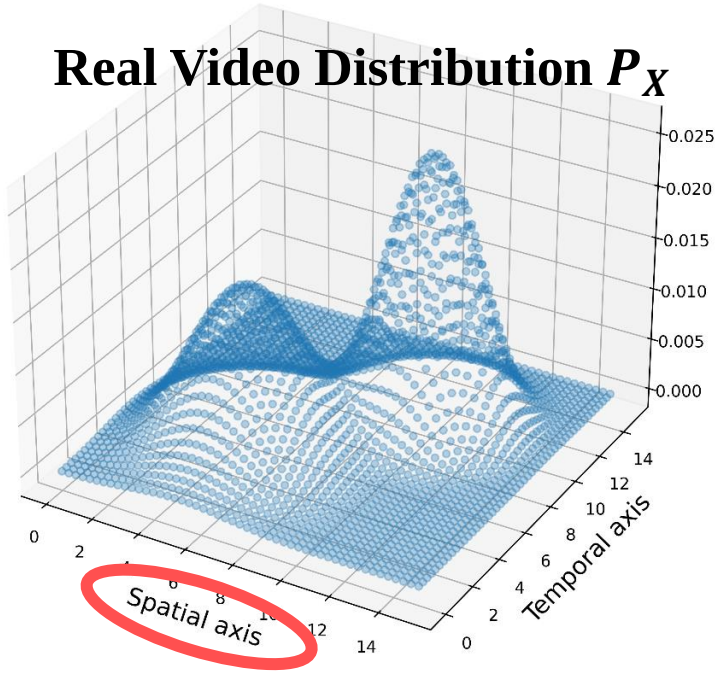
**Real Video Distribution $P_X$**



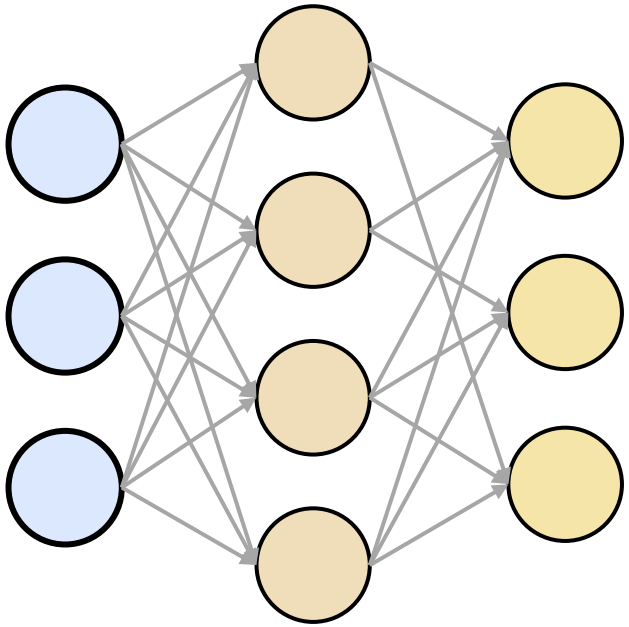



**Ideal Video Generator**

# Video Generative Models and Evaluation Protocol

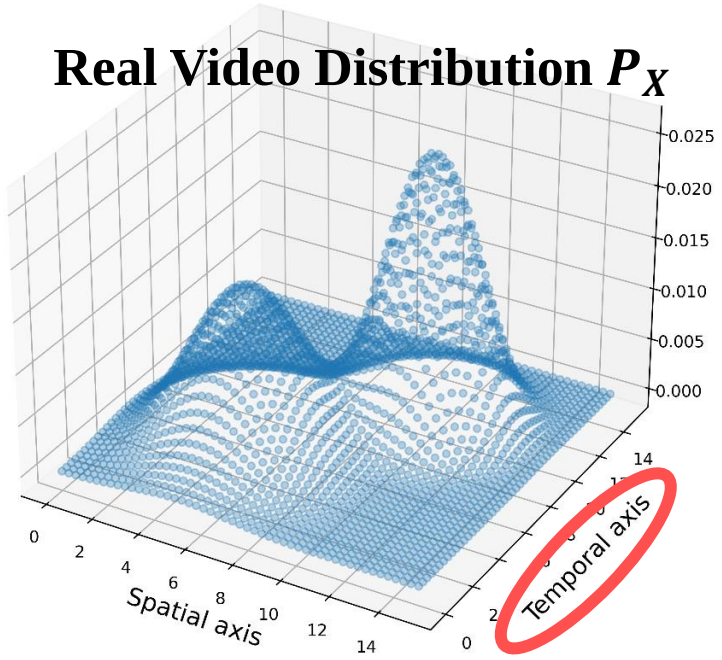



**Spatial aspect**

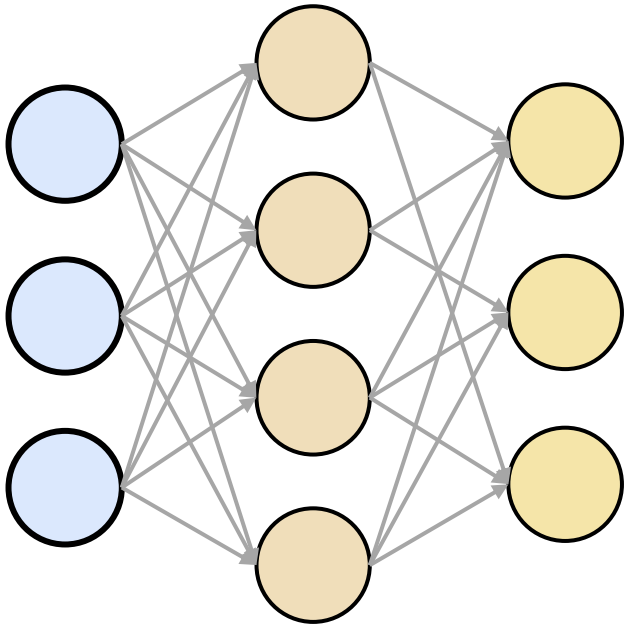


**Ideal Video Generator**

# Video Generative Models and Evaluation Protocol



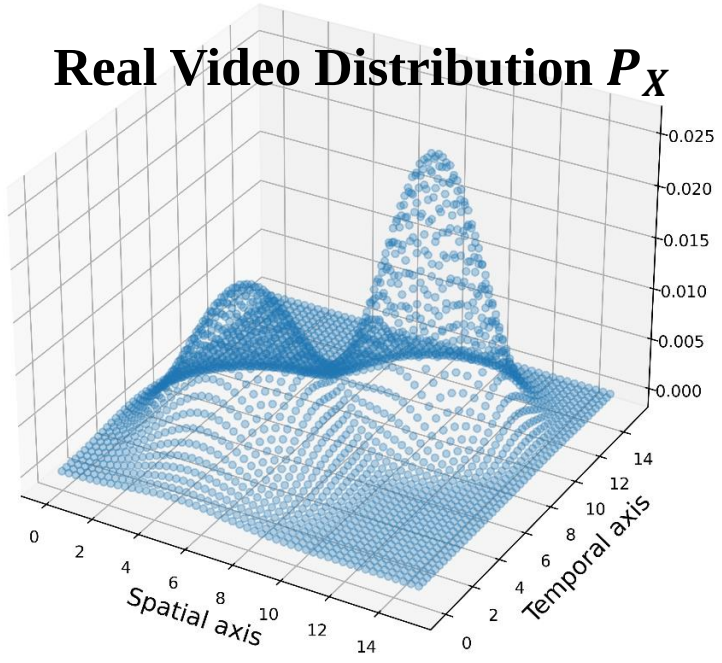

**Temporal aspect**

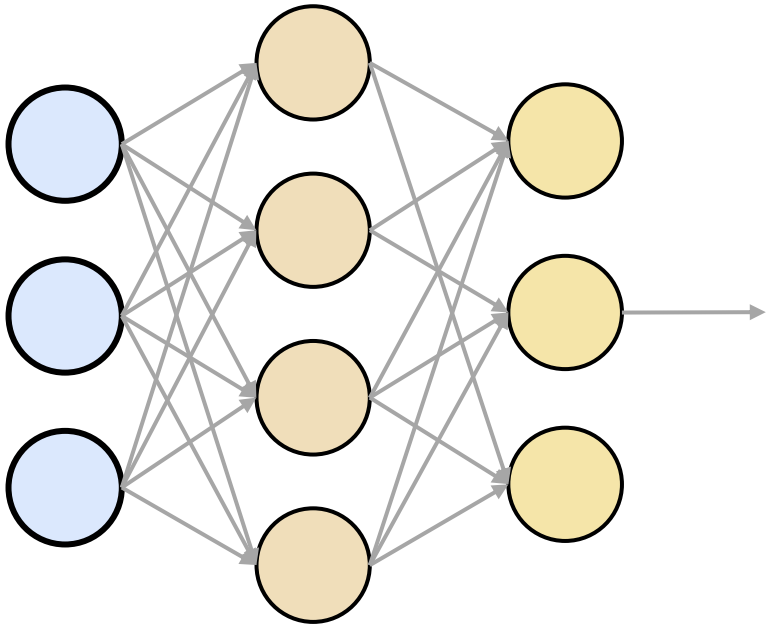


**Ideal Video Generator**

# Video Generative Models and Evaluation Protocol

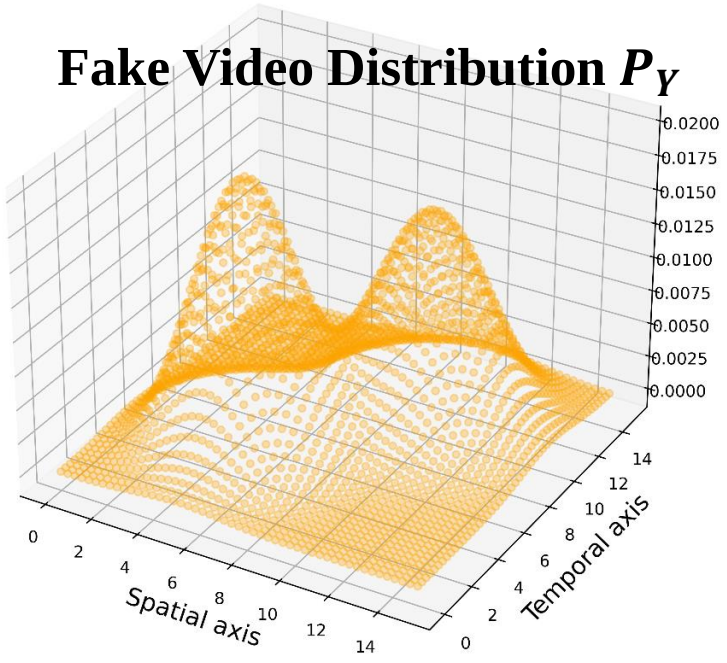
**Real Video Distribution $P_X$**



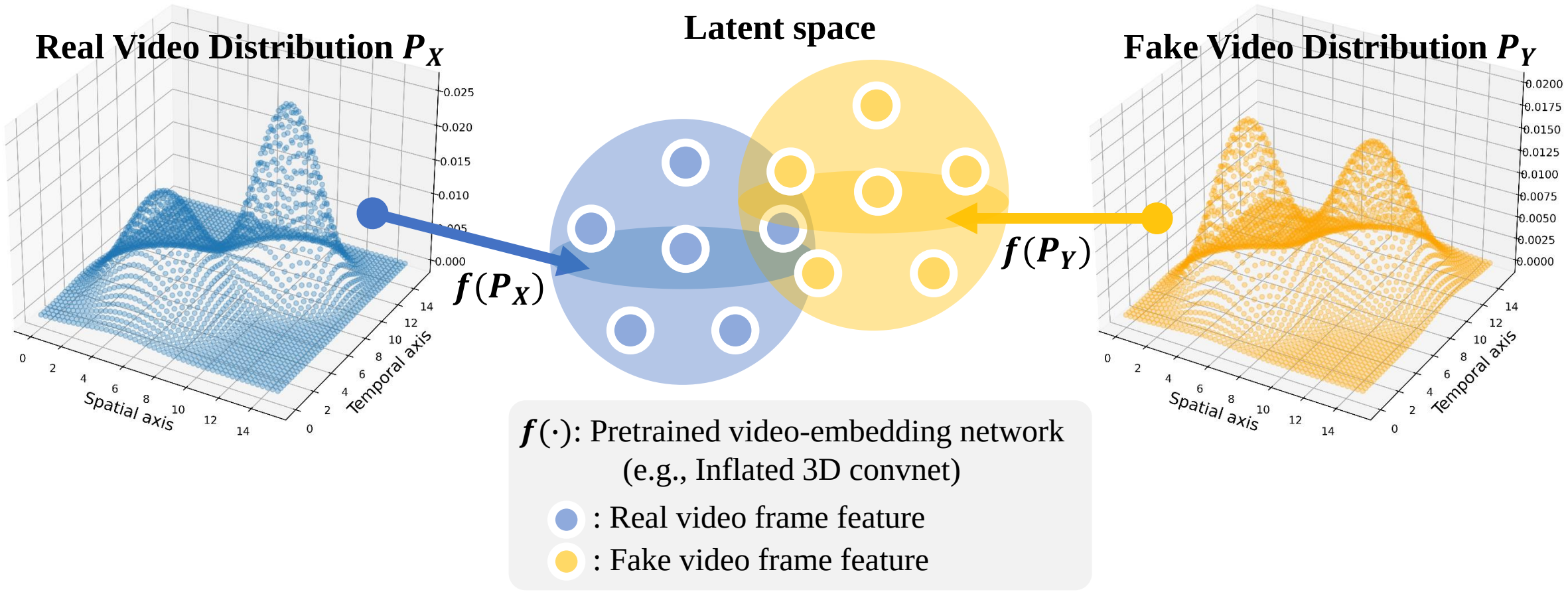



**Ideal Video Generator**

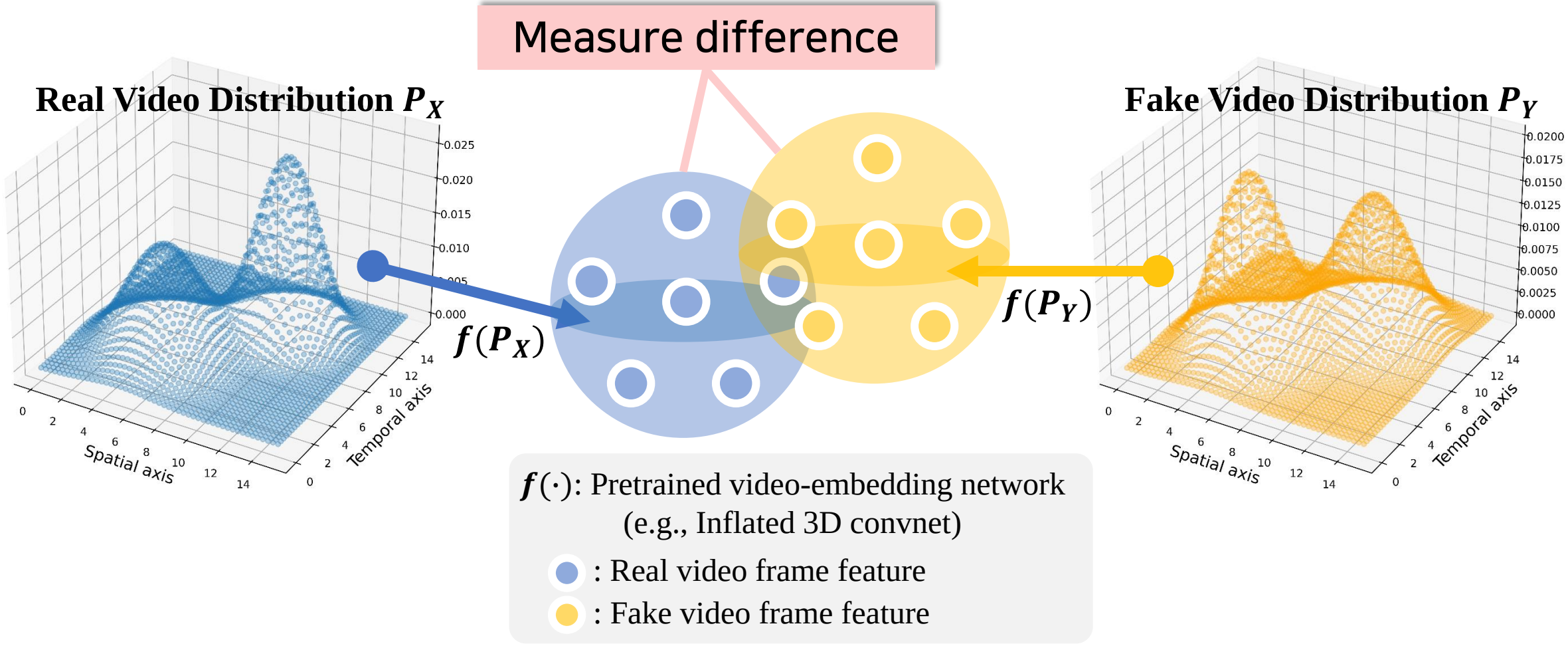
**Fake Video Distribution $P_Y$**



# Video Generative Models and Evaluation Protocol

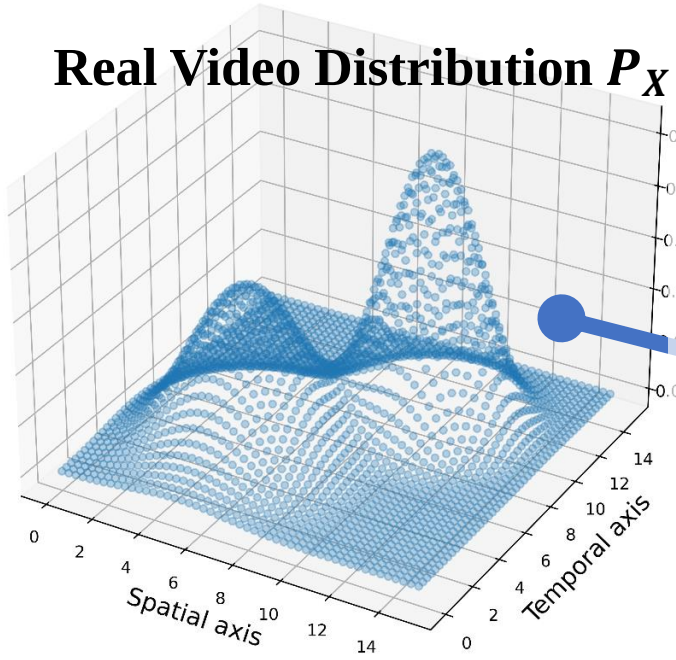


# Video Generative Models and Evaluation Protocol



# Video Generative Models and Evaluation Protocol

**Real Video Distribution $P_X$**



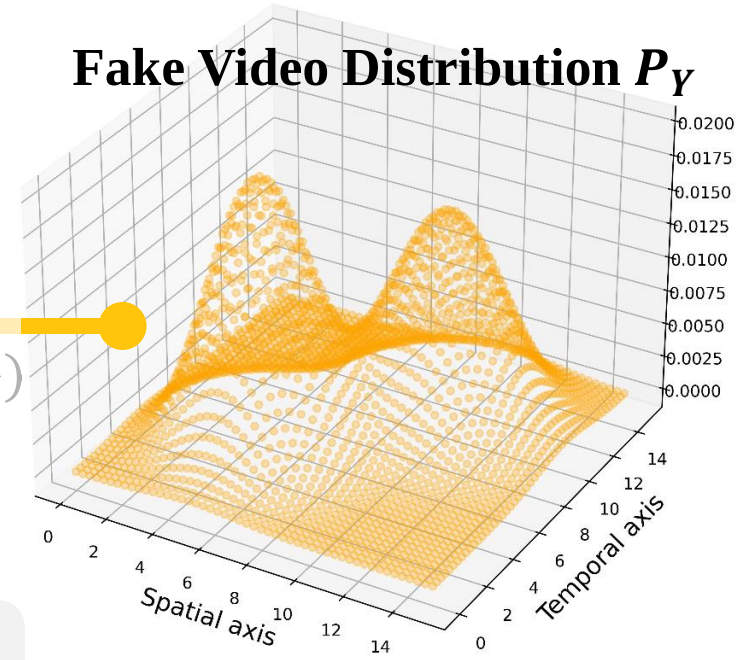

**(1)** Cannot account for long video frames

**(2)** Classifier embedding network achieves its performance based on the spatial information

**(3)** Metrics giving the single score are only useful to compare or rank models

$f(P_X)$ $f(P_Y)$

**Fake Video Distribution $P_Y$**

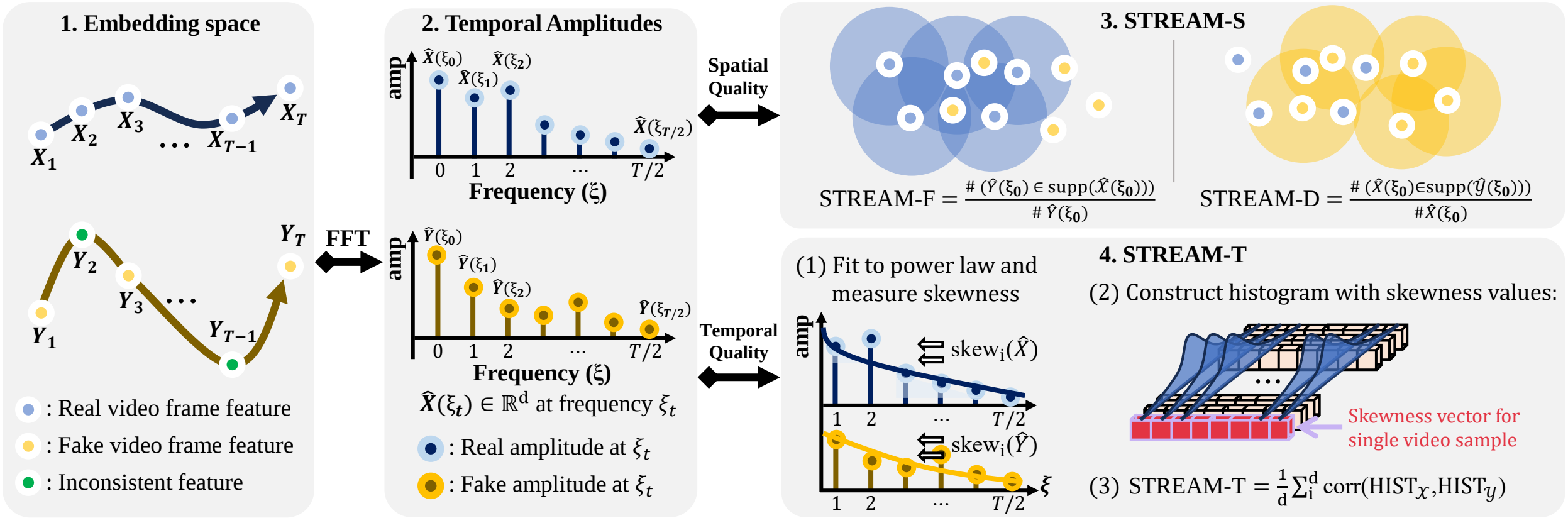



$f(\cdot)$: Pretrained video-embedding network (e.g., Inflated 3D convnet)

: Real video frame feature

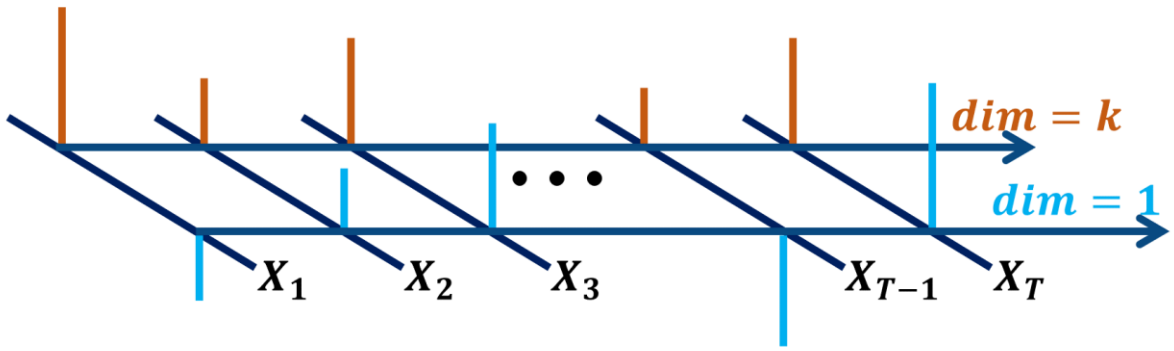
: Fake video frame feature

# Overview of STREAM



# STREAM-T: Evaluating the Temporal Flow of Videos

## (1) Vector representation of temporal flow



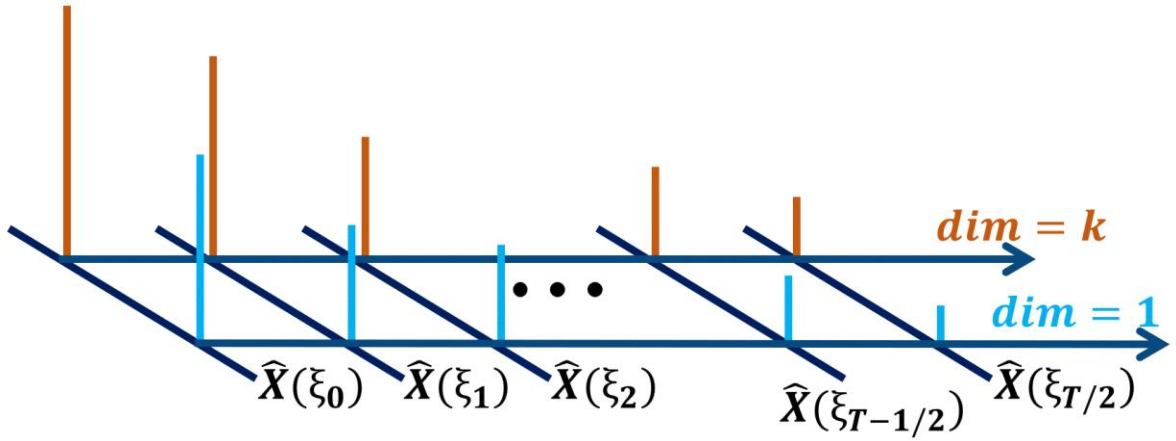

: Feature value at **1st** dimension

: Feature value at **kth** dimension

: Temporal flow at **1st** dimension

: Temporal flow at ***k*th** dimension

## (2) Fourier series of temporal flow




: Amplitude value at **1st** dimension

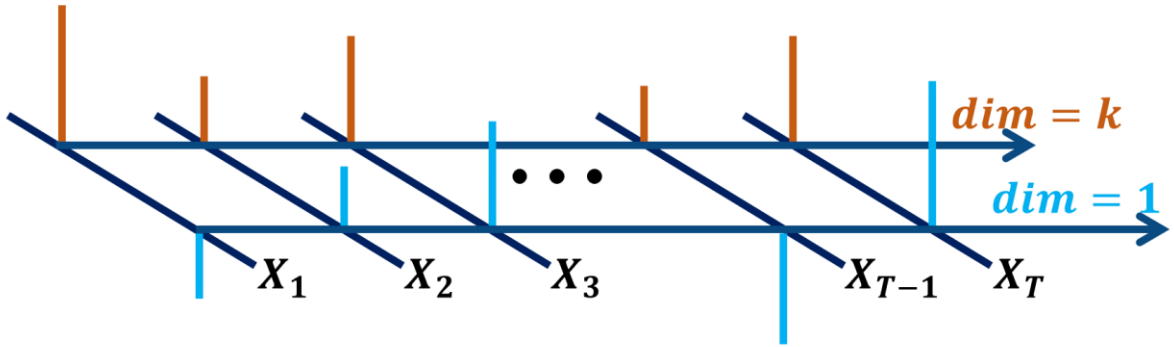
: Amplitude value at **kth** dimension

: Change of amplitude at **1st** dimension

: Change of amplitude at ***k*th** dimension

# STREAM-T: Evaluating the Temporal Flow of Videos

## (1) Vector representation of temporal flow



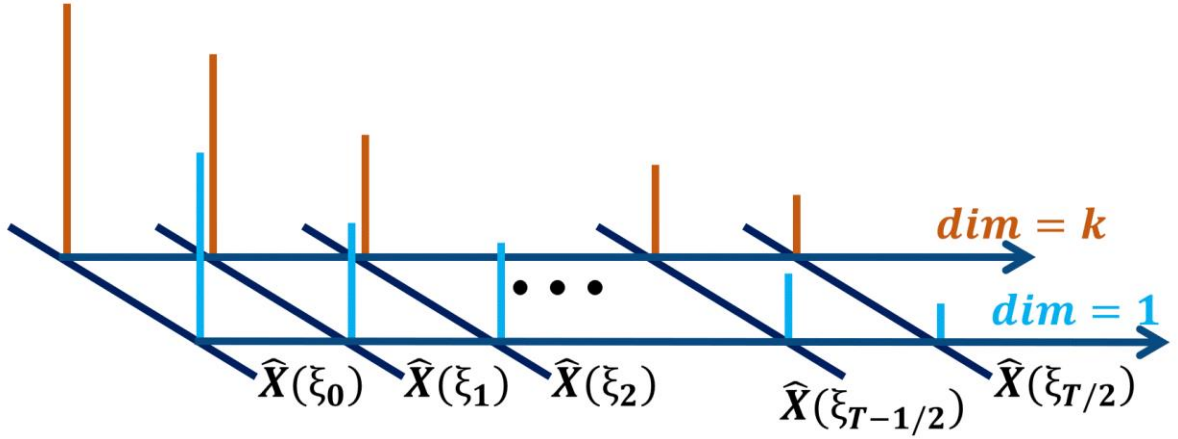

: Feature value at **1st** dimension

: Feature value at **kth** dimension

: Temporal flow at **1st** dimension

: Temporal flow at ***k*th** dimension

## (2) Fourier series of temporal flow




: Amplitude value at **1st** dimension

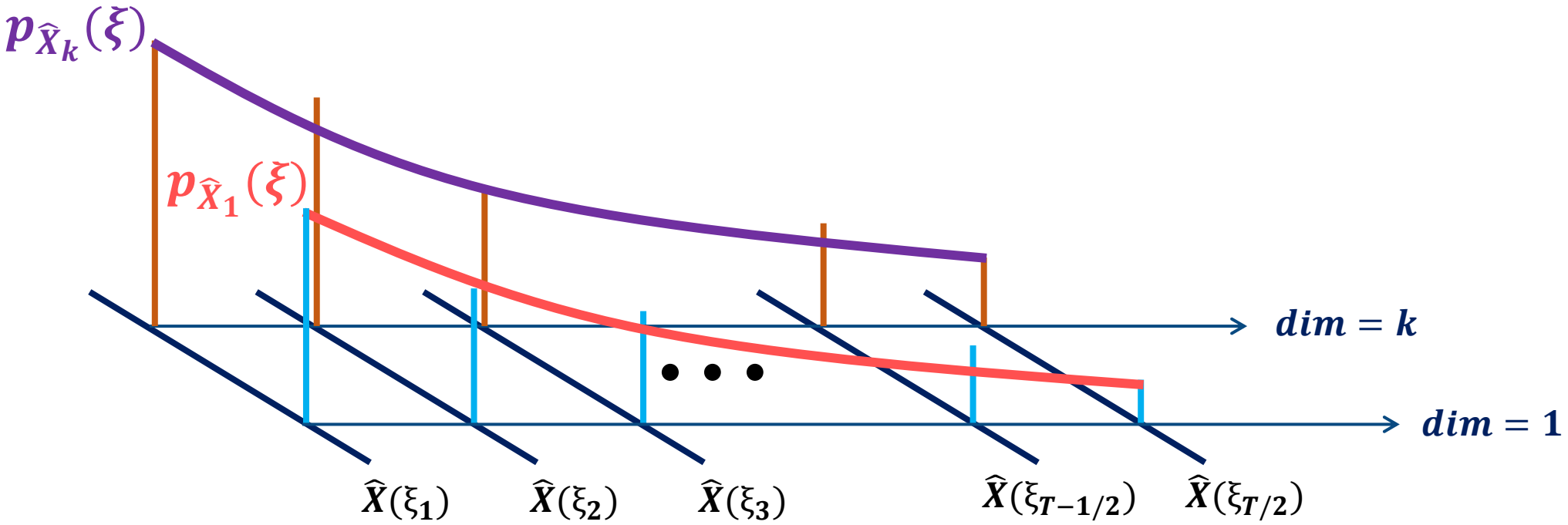
: Amplitude value at **kth** dimension

: Change of amplitude at **1st** dimension

: Change of amplitude at ***k*th** dimension

# STREAM-T: Evaluating the Temporal Flow of Videos

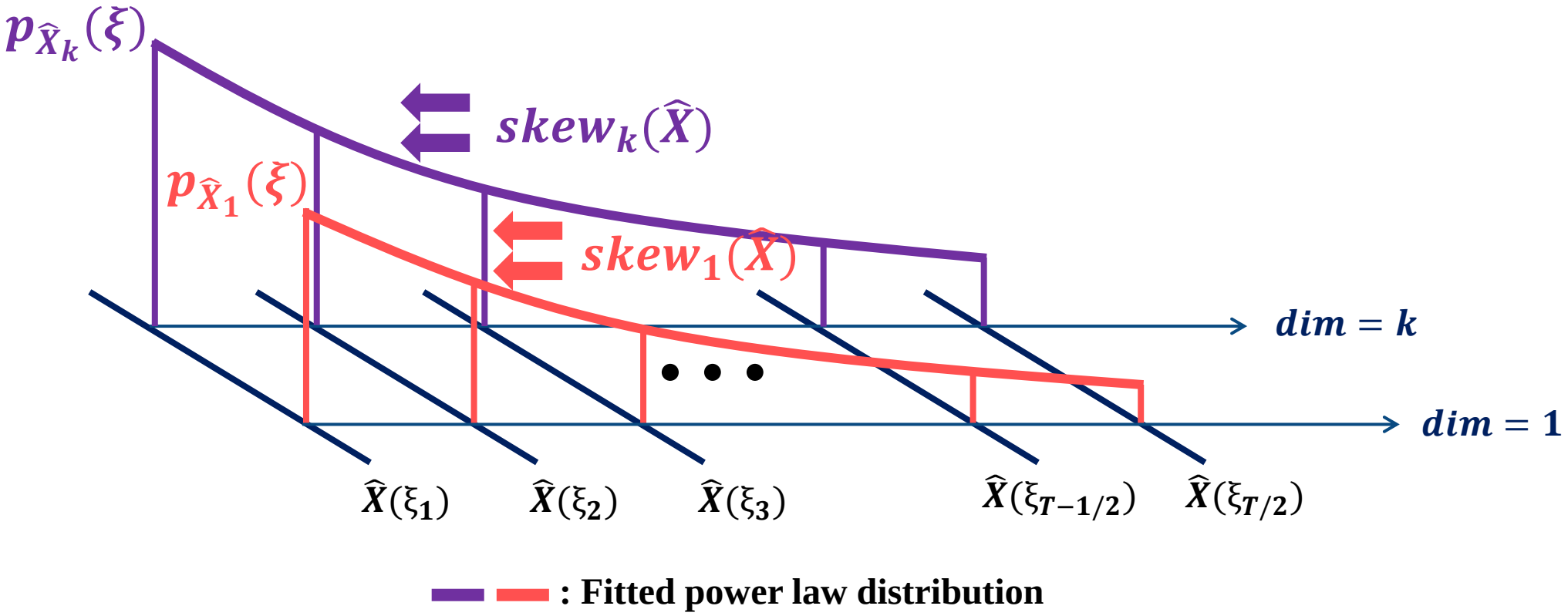
## (3) Fit Fourier series to the power law distribution




: Amplitude value at **1st** dimension  : Amplitude value at ***k*th** dimension  **: Fitted power law distribution**

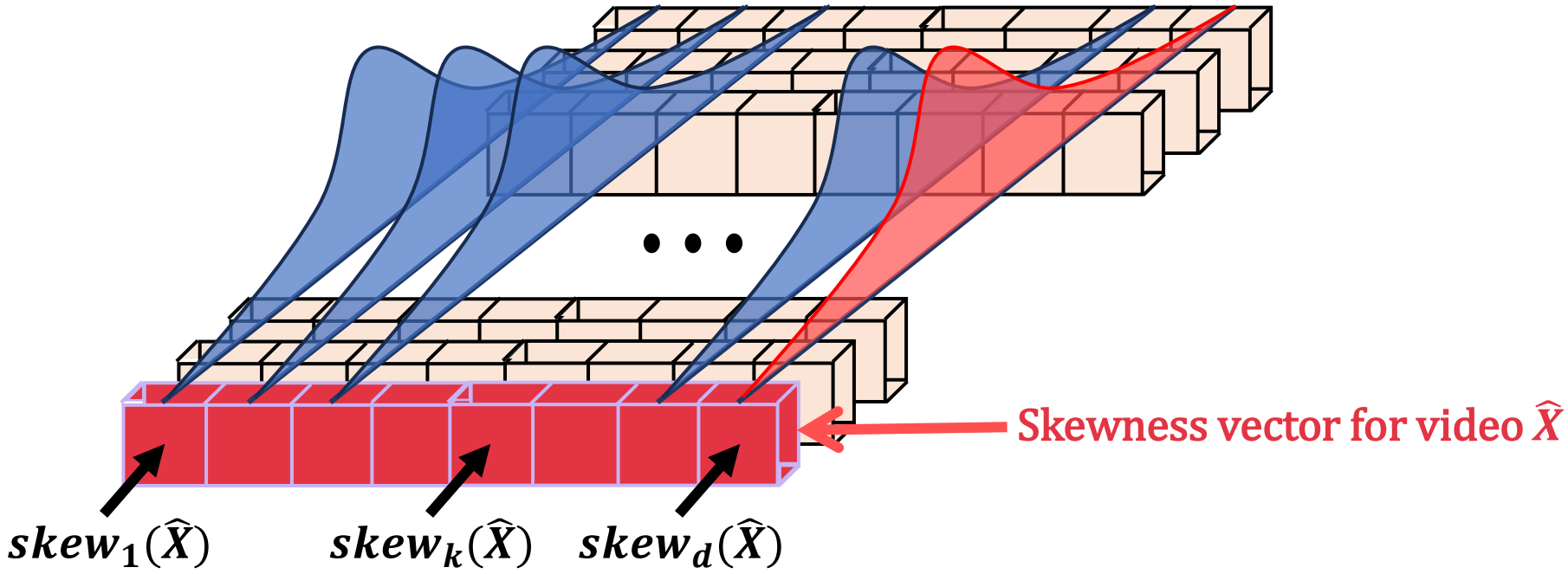
# STREAM-T: Evaluating the Temporal Flow of Videos

**(4) Calculate "skewness" from the power law distribution**



# STREAM-T: Evaluating the Temporal Flow of Videos

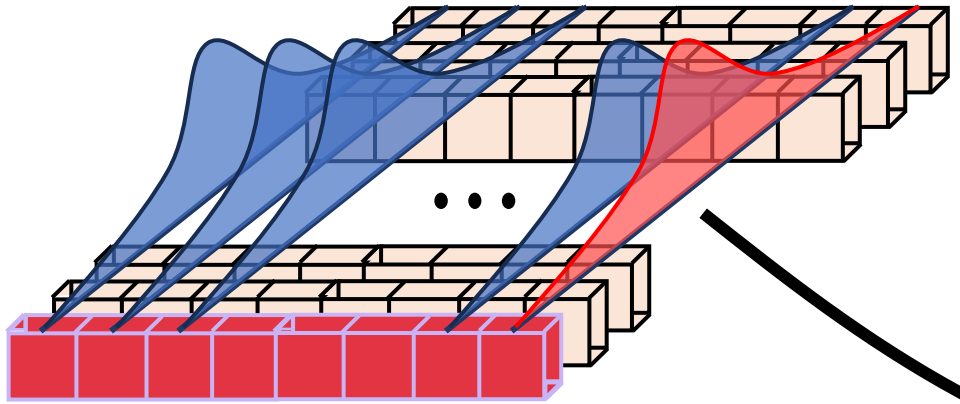
**(5) Construct "histogram" for each skewness values at each feature dimension**



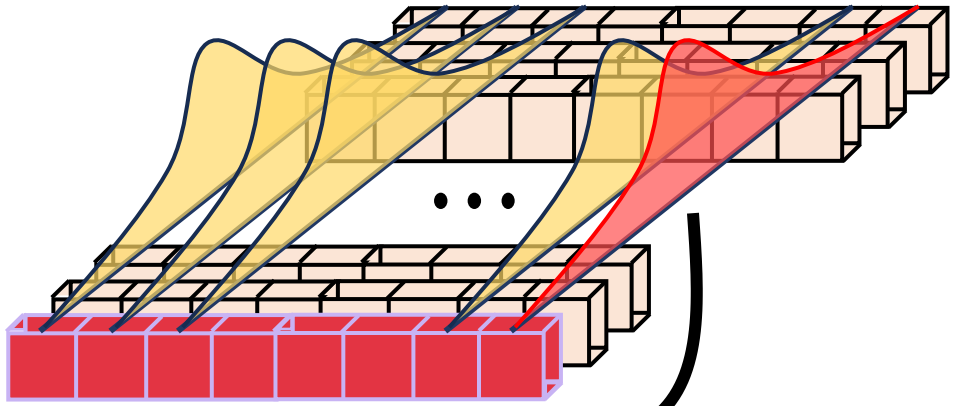
# STREAM-T: Evaluating the Temporal Flow of Videos

## (5) Calculate STREAM-T: mean correlation between real and fake histograms

Set of histograms for real dataset: $HIST_X$

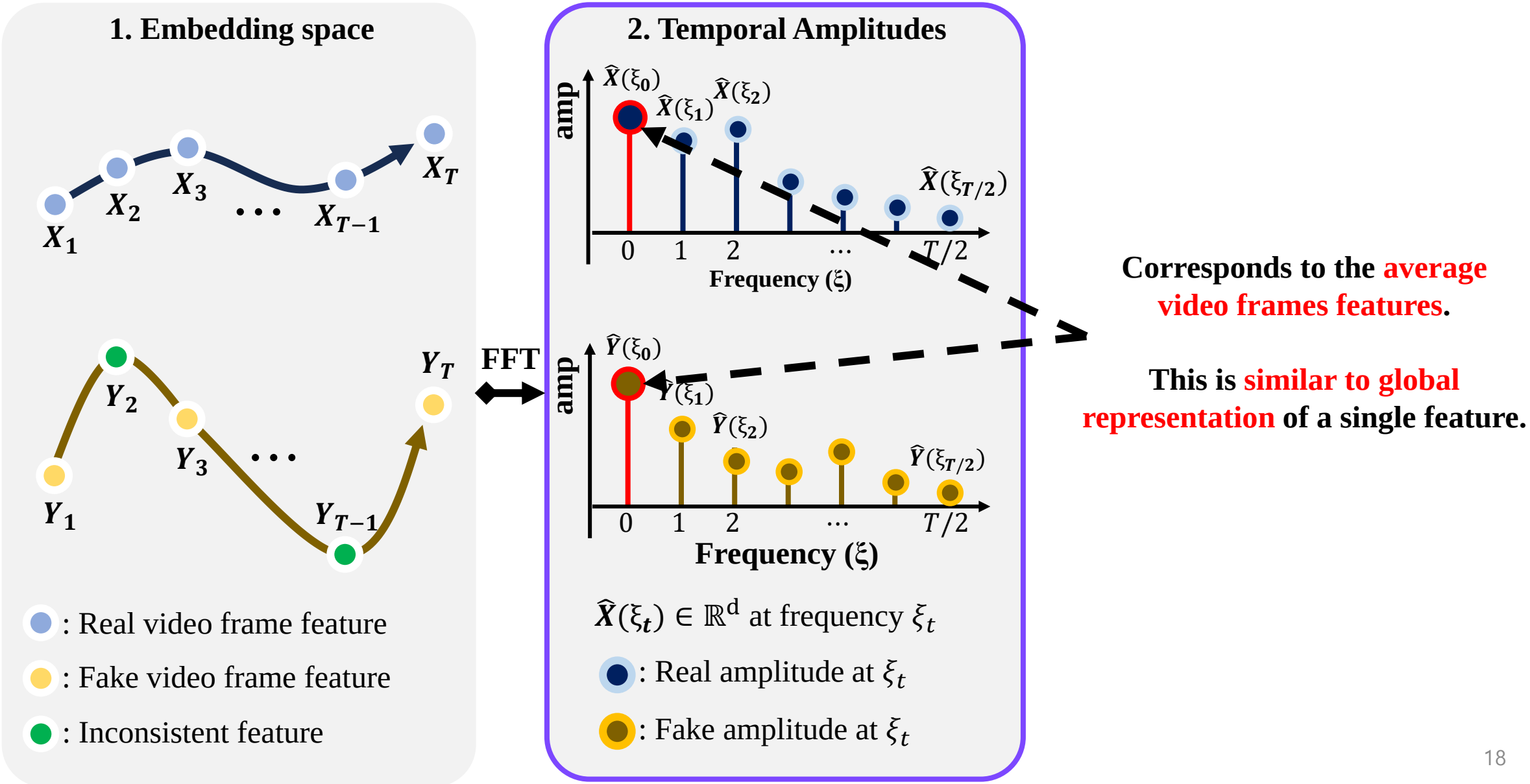


Set of histograms for fake dataset: $HIST_Y$



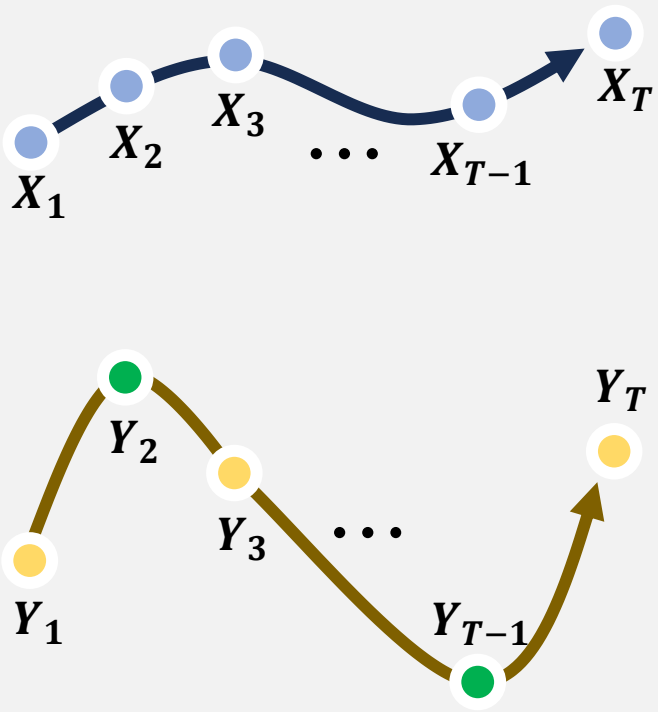
$$\text{STREAM-T} = \frac{1}{d}\sum_{i}^{d} \text{corr}(\text{HIST}_X, \text{HIST}_Y)$$

# STREAM-S: Evaluating the Spatial Quality of Videos



# STREAM-S: Evaluating the Spatial Quality of Videos

## 1. Embedding space



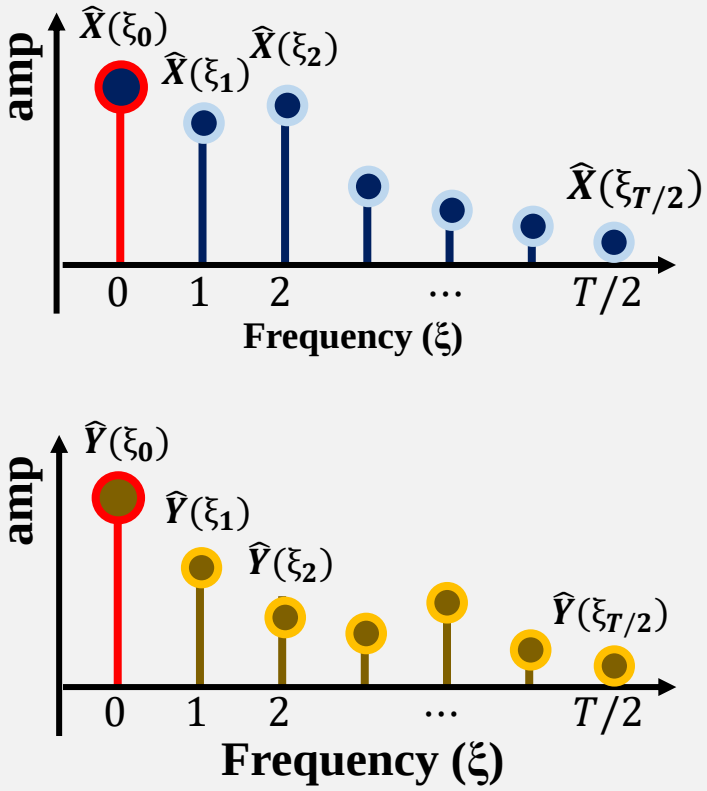

: Real video frame feature

: Fake video frame feature

: Inconsistent feature

FFT

## 2. Temporal Amplitudes

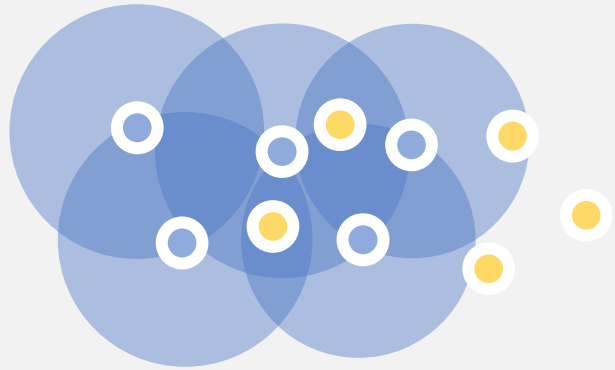



$\hat{X}(\xi_t) \in \mathbb{R}^d$ at frequency $\xi_t$

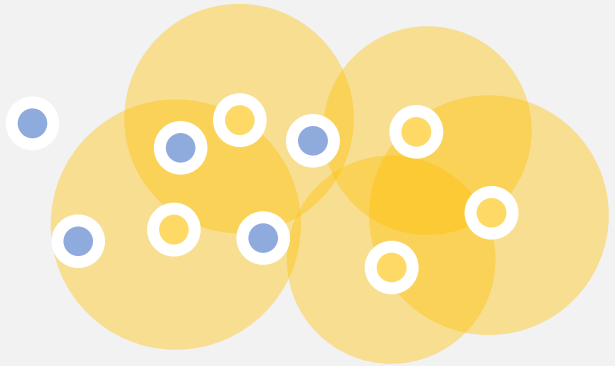
: Real amplitude at $\xi_t$

: Fake amplitude at $\xi_t$

## 3. STREAM-S



$$\text{STREAM-F} = \frac{\#\left(\hat{Y}(\xi_0) \in supp(\hat{X}(\xi_0))\right)}{\#\,\hat{Y}(\xi_0)}$$

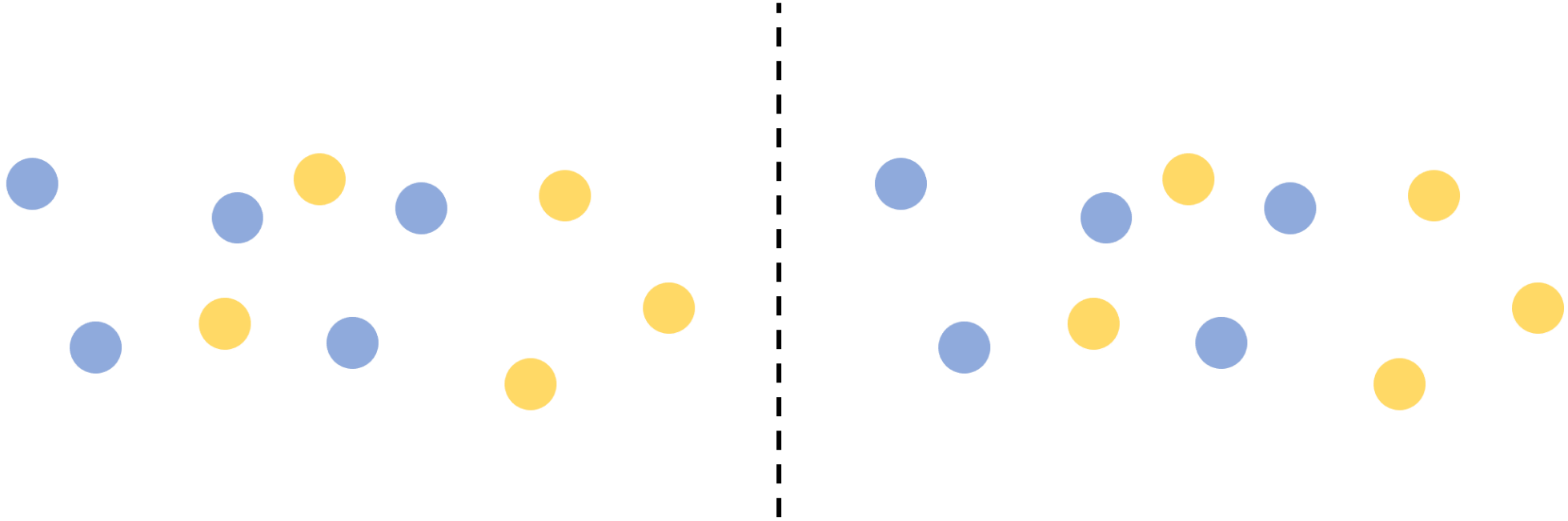


$$\text{STREAM-D} = \frac{\#\left(\hat{X}(\xi_0) \in supp(\hat{Y}(\xi_0))\right)}{\#\hat{X}(\xi_0)}$$

# STREAM-S: Evaluating the Spatial Quality of Videos

## (a) Fidelity

## (b) Diversity



: Mean amplitude feature of a single real video

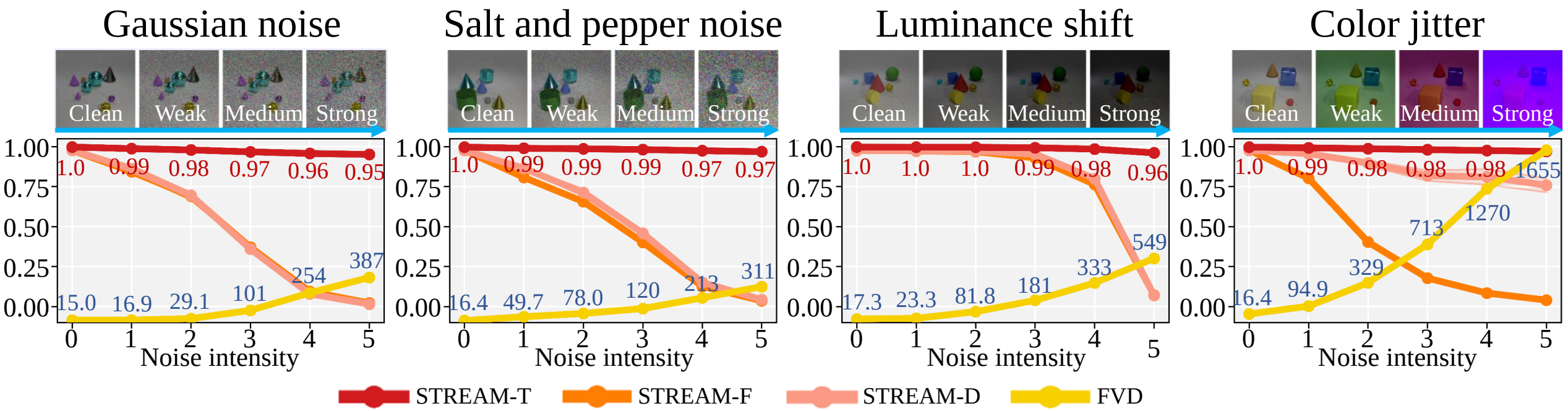
: Mean amplitude feature of a single fake video

$$\text{STREAM-F} = \frac{\#\ inlying\ fake}{\#\ fake} = \frac{3}{5}$$

$$\text{STREAM-D} = \frac{\#\ inlying\ real}{\#\ real} = \frac{4}{5}$$

# Experiments & Results – toy data experiment

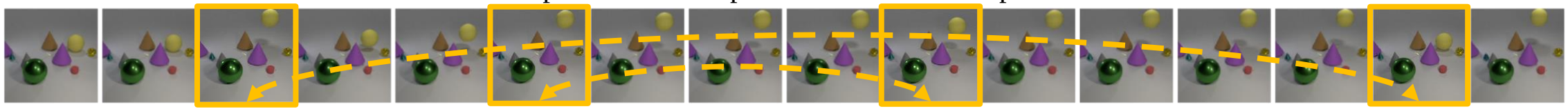
**Behavior of STREAM regarding noise affecting the "visual quality"**



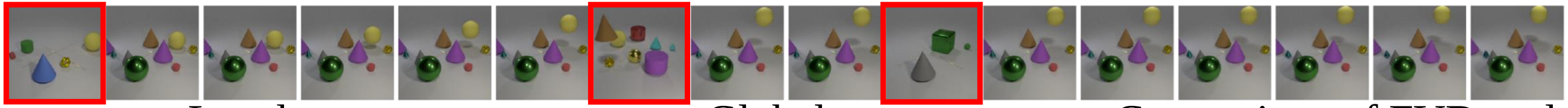
# Experiments & Results – toy data experiment

**Comparison of the behaviors of STREAM and FVD on "temporal degradations"**

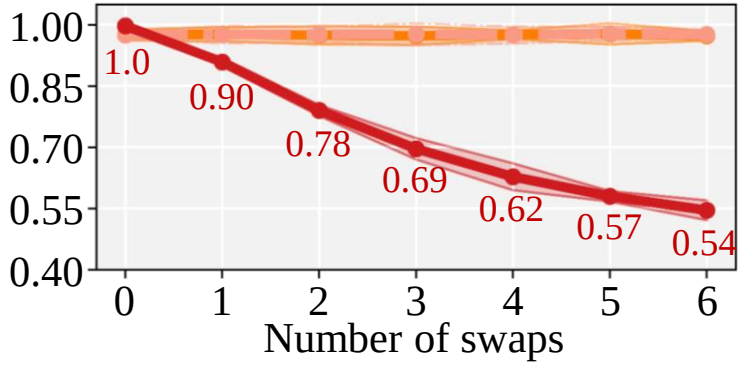
Example of local swap when number of swaps are 2



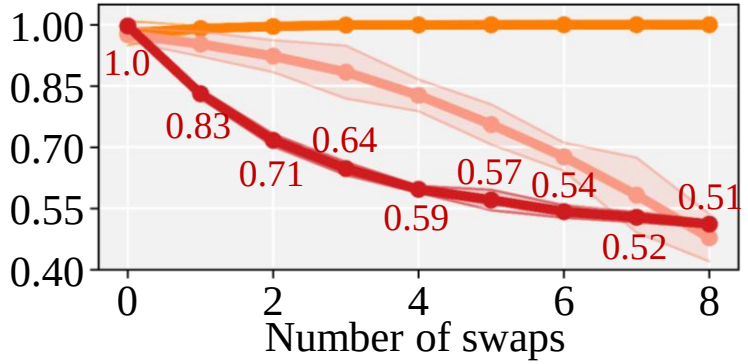
Example of global swap when number of swaps are 3



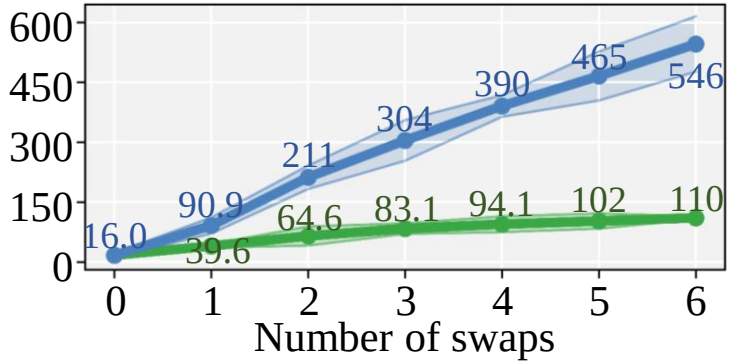
Local swap




Global swap




Comparison of FVD scale




# Experiments & Results – ranking experiment

**Comparison of unconditional video generative models trained with 16 frame videos**

Table 1: Comparison and analysis of video generative models (unconditional). All the models are trained on the UCF-101 dataset generating 16 frame videos with $128 \times 128$ resolution. The numbers in parentheses next to evaluation scores represent the standard deviation of the scores, calculated through five repeated measurements.

| | VIS (↑) | FVD (↓) | STREAM-T (↑) | STREAM-F (↑) | STREAM-D (↑) |
|---|---|---|---|---|---|
| MoCoGAN | 16.64 (±0.09) | 1174.3 (±36.69) | 0.9683 (±0.001) | 0.1595 (±0.023) | 0.0000 (±0.000) |
| DIGAN | 24.32 (±0.19) | 763.64 (±28.82) | 0.9743 (±0.000) | 0.3101 (±0.011) | 0.0662 (±0.005) |
| TATS | 34.39 (±0.30) | 693.27 (±21.55) | 0.9832 (±0.000) | 0.9120 (±0.011) | 0.0850 (±0.005) |
| VideoGPT | 30.35 (±0.55) | 647.75 (±15.34) | 0.9782 (±0.000) | 0.7806 (±0.030) | 0.3272 (±0.005) |
| MeBT | 64.54 (±0.51) | 504.21 (±24.50) | 0.9616 (±0.001) | 0.7441 (±0.006) | 0.1852 (±0.019) |
| PVDM | 60.02 (±0.82) | 415.70 (±25.59) | 0.9843 (±0.002) | 0.6416 (±0.014) | 0.3112 (±0.005) |

# Experiments & Results – ranking experiment

**Comparison of unconditional video generative models trained with 16 frame videos**

Table 1: Comparison and analysis of video generative models (unconditional). All the models are trained on the UCF-101 dataset generating 16 frame videos with $128 \times 128$ resolution. The numbers in parentheses next to evaluation scores represent the standard deviation of the scores, calculated through five repeated measurements.

| | VIS (↑) | FVD (↓) | STREAM-T (↑) | STREAM-F (↑) | STREAM-D (↑) |
|---|---|---|---|---|---|
| MoCoGAN | 16.64 (±0.09) | 1174.3 (±36.69) | 0.9683 (±0.001) | 0.1595 (±0.023) | 0.0000 (±0.000) |
| DIGAN | 24.32 (±0.19) | 763.64 (±28.82) | 0.9743 (±0.000) | 0.3101 (±0.011) | 0.0662 (±0.005) |
| TATS | 34.39 (±0.30) | 693.27 (±21.55) | 0.9832 (±0.000) | 0.9120 (±0.011) | 0.0850 (±0.005) |
| VideoGPT | 30.35 (±0.55) | 647.75 (±15.34) | 0.9782 (±0.000) | 0.7806 (±0.030) | 0.3272 (±0.005) |
| MeBT | 64.54 (±0.51) | 504.21 (±24.50) | 0.9616 (±0.001) | 0.7441 (±0.006) | 0.1852 (±0.019) |
| PVDM | 60.02 (±0.82) | 415.70 (±25.59) | 0.9843 (±0.002) | 0.6416 (±0.014) | 0.3112 (±0.005) |

**Generative models for short video clips properly accounts for temporal naturalness**

# Experiments & Results – ranking experiment

**Comparison of unconditional video generative models trained with 16 frame videos**

Table 1: Comparison and analysis of video generative models (unconditional). All the models are trained on the UCF-101 dataset generating 16 frame videos with $128 \times 128$ resolution. The numbers in parentheses next to evaluation scores represent the standard deviation of the scores, calculated through five repeated measurements.

| | VIS (↑) | FVD (↓) | STREAM-T (↑) | STREAM-F (↑) | STREAM-D (↑) |
|---|---|---|---|---|---|
| MoCoGAN | 16.64 (±0.09) | 1174.3 (±36.69) | 0.9683 (±0.001) | 0.1595 (±0.023) | 0.0000 (±0.000) |
| DIGAN | 24.32 (±0.19) | 763.64 (±28.82) | 0.9743 (±0.000) | 0.3101 (±0.011) | 0.0662 (±0.005) |
| TATS | 34.39 (±0.30) | 693.27 (±21.55) | 0.9832 (±0.000) | 0.9120 (±0.011) | 0.0850 (±0.005) |
| VideoGPT | 30.35 (±0.55) | 647.75 (±15.34) | 0.9782 (±0.000) | 0.7806 (±0.030) | 0.3272 (±0.005) |
| MeBT | 64.54 (±0.51) | 504.21 (±24.50) | 0.9616 (±0.001) | 0.7441 (±0.006) | 0.1852 (±0.019) |
| PVDM | 60.02 (±0.82) | 415.70 (±25.59) | 0.9843 (±0.002) | 0.6416 (±0.014) | 0.3112 (±0.005) |

**Current generative models cannot generate diverse samples**

# Experiments & Results – ranking experiment

**Comparison of unconditional video generative models trained with 16 frame videos**

Table 1: Comparison and analysis of video generative models (unconditional). All the models are trained on the UCF-101 dataset generating 16 frame videos with $128 \times 128$ resolution. The numbers in parentheses next to evaluation scores represent the standard deviation of the scores, calculated through five repeated measurements.

| | VIS (↑) | FVD (↓) | STREAM-T (↑) | STREAM-F (↑) | STREAM-D (↑) |
|---|---|---|---|---|---|
| MoCoGAN | 16.64 (±0.09) | 1174.3 (±36.69) | 0.9683 (±0.001) | | |
| DIGAN | 24.32 (±0.19) | 763.64 (±28.82) | 0.9743 (±0.000) | | |
| TATS | 34.39 (±0.30) | 693.27 (±21.55) | 0.9832 (±0.000) | 0.9120 (±0.011) | 0.0850 (±0.005) |
| VideoGPT | 30.35 (±0.55) | 647.75 (±15.34) | 0.9782 (±0.000) | 0.7806 (±0.030) | 0.3272 (±0.005) |
| MeBT | 64.54 (±0.51) | 504.21 (±24.50) | 0.9616 (±0.001) | 0.7441 (±0.006) | 0.1852 (±0.019) |
| PVDM | 60.02 (±0.82) | 415.70 (±25.59) | 0.9843 (±0.002) | 0.6416 (±0.014) | 0.3112 (±0.005) |

**VIS: TATS > VideoGPT**
**FVD: TATS < VideoGPT**

**STREAM is capable to interpret the performance of generative models**

# Experiments & Results – ranking experiment

**Comparison of unconditional video generative models trained with 16 frame videos**

Table 3: Human evaluation results of video generative models listed in Table 1. Each spatial and temporal scrore denotes the sum of scores evaluated by 81 raters for each model, and the numbers in parentheses indicate the average scores for each model.

| Models | MoCoGAN-HD | DIGAN | TATS-base | VideoGPT | MeBT |
|---|---|---|---|---|---|
| Spatial score | 183 (2.25) | 237 (2.92) | 378 (4.66) | 304 (3.75) | 356 (4.39) |
| Temporal score | 122 (1.50) | 141 (1.74) | 161 (1.98) | 129 (1.59) | 145 (1.79) |

# Experiments & Results – ranking experiment

**Comparison of unconditional video generative models trained with 16 frame videos**

Table 3: Human evaluation results of video generative models listed in Table 1. Each spatial and temporal scrore denotes the sum of scores evaluated by 81 raters for each model, and the numbers in parentheses indicate the average scores for each model.

| Models | MoCoGAN-HD | DIGAN | TATS-base | VideoGPT | MeBT |
|---|---|---|---|---|---|
| Spatial s | | .9 | | | 56 (4.39) |
| Tempora | | .7 | | | 45 (1.79) |

**Spatial (STREAM): 0.9**
**Temporal (STREAM): 0.6**

**Spatial (FVD): 0.7**
**Temporal (FVD): 0.5**

# Experiments & Results – ranking experiment

**Comparison of unconditional video generative models trained with 128 frame videos**

Table 2: Comparison and analysis of video generative models which produce long video frames (unconditional). All the models are trained on UCF-101 dataset generating 128 frame videos with $128 \times 128$ resolution. sVIS and sFVD denotes the modified version of VIS and FVD measured for every 16 frames using a sliding window. See Appendix A.8 and A.9 for the sample qualities.

| | sVIS (↑) | sFVD (↓) | STREAM-T (↑) | STREAM-F (↑) | STREAM-D (↑) |
|---|---|---|---|---|---|
| MoCoGAN | 11.8450 | 1454.1 | 0.3274 | 0.0615 | 0.0000 |
| DIGAN | 18.0075 | 1103.0 | 0.1327 | 0.1206 | 0.2656 |
| TATS | 40.3345 | 1008.0 | 0.0302 | 0.6284 | 0.2104 |
| MeBT | 33.9492 | 948.51 | 0.8265 | 0.6284 | 0.1601 |

# Experiments & Results – ranking experiment

**Comparison of unconditional video generative models trained with 128 frame videos**

Table 2: Comparison and analysis of video generative models which produce long video frames (unconditional). All the models are trained on UCF-101 dataset generating 128 frame videos with $128 \times 128$ resolution. sVIS and sFVD denotes the modified version of VIS and FVD measured for every 16 frames using a sliding window. See Appendix A.8 and A.9 for the sample qualities.

| | sVIS (↑) | sFVD (↓) | STREAM-T (↑) | STREAM-F (↑) | STREAM-D (↑) |
|---|---|---|---|---|---|
| MoCoGAN | 11.8450 | 1454.1 | 0.3274 | 0.0615 | 0.0000 |
| DIGAN | 18.0075 | 1103.0 | 0.1327 | 0.1206 | 0.2656 |
| TATS | 40.3345 | 1008.0 | 0.0302 | 0.6284 | 0.2104 |
| MeBT | 33.9492 | 948.51 | 0.8265 | 0.6284 | 0.1601 |

**Current generative models cannot properly account for the video temporal naturalness**

# Experiments & Results – ranking experiment

**Comparison of unconditional video generative models trained with 128 frame videos**

Table 2: Comparison and analysis of video generative models which produce long video frames (unconditional). All the models are trained on UCF-101 dataset generating 128 frame videos with $128 \times 128$ resolution. sVIS and sFVD denotes the modified version of VIS and FVD measured for every 16 frames using a sliding window. See Appendix A.8 and A.9 for the sample qualities.

| | sVIS (↑) | sFVD (↓) | STREAM-T (↑) | STREAM-F (↑) | STREAM-D (↑) |
|---|---|---|---|---|---|
| MoCoGAN | 11.8450 | 1454.1 | 0.3274 | 0.0615 | 0.0000 |
| DIGAN | 18.0075 | 1103.0 | 0.1327 | 0.1206 | 0.2656 |
| TATS | 40.3345 | 1008.0 | 0.0302 | 0.6284 | 0.2104 |
| MeBT | 33.9492 | 948.51 | 0.8265 | 0.6284 | 0.1601 |

**Current generative models cannot generate diverse samples**

# Thank you

Project Page

Quick Start!

```
pip install v-stream
```

Use our method by only pip command!