# Optimizing Immersive Video Streaming for Head-Mounted Virtual Reality

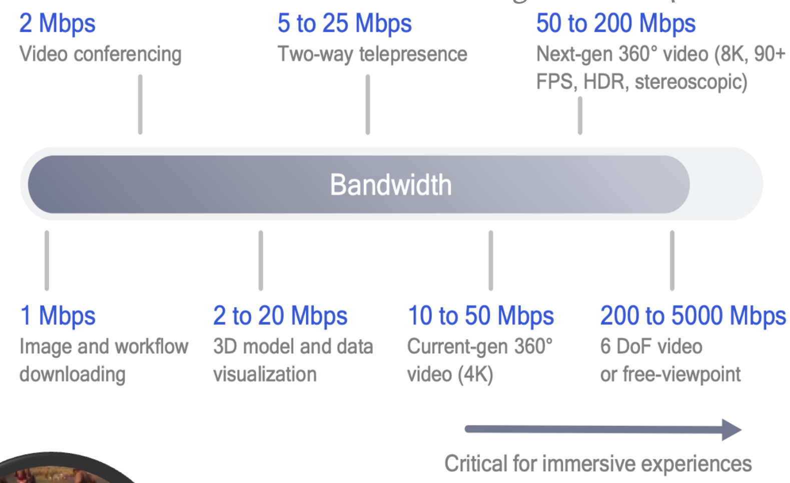

Ching-Ling Fan (ch.ling.fan@gmail.com)
*Supervised by Prof. Cheng-Hsin Hsu*

*Department of Computer Science, National Tsing Hua University, Taiwan*

# Immersive Videos (a.k.a. 360° Videos)

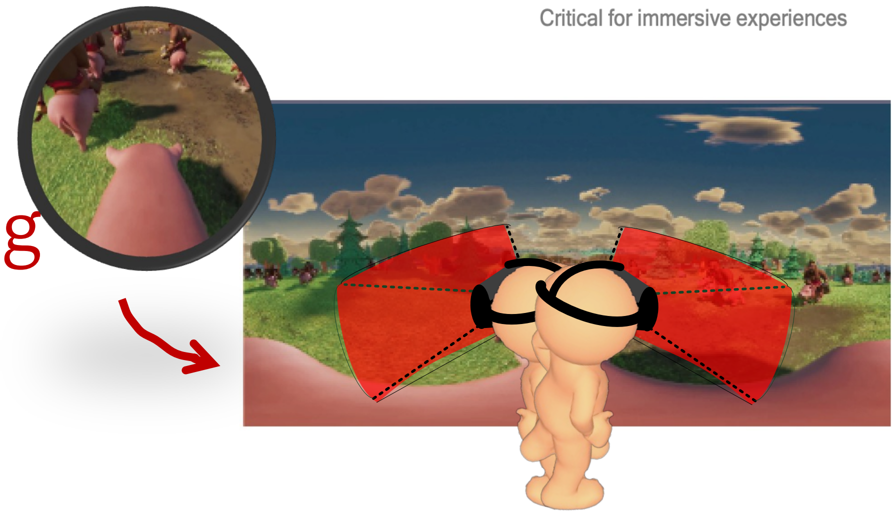
# Challenges of Streaming 360° Videos

*Image Source: Qualcomm*




- 360° videos contain wider view than conventional videos
  ⇒ extremely large file size
  (> 130 Mbps in HEVC for 4K viewport)

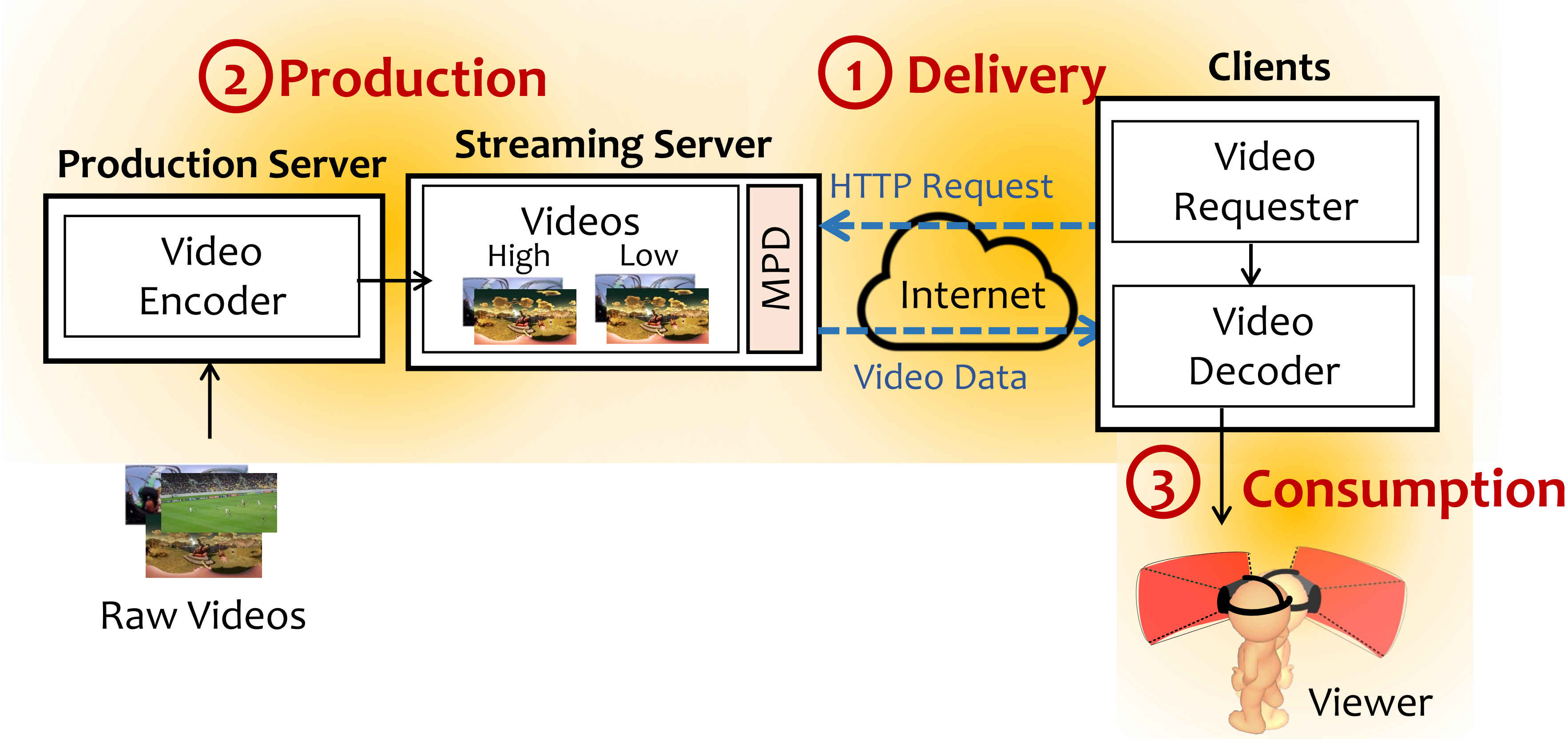
- Shape distortion and diverse user behavior
  ⇒ hard to capture QoE using existing quality metrics



***Insufficient bandwidth & complex and unknown QoE***

# 360° Video Streaming Platform

- Three crucial phases in 360° video streaming

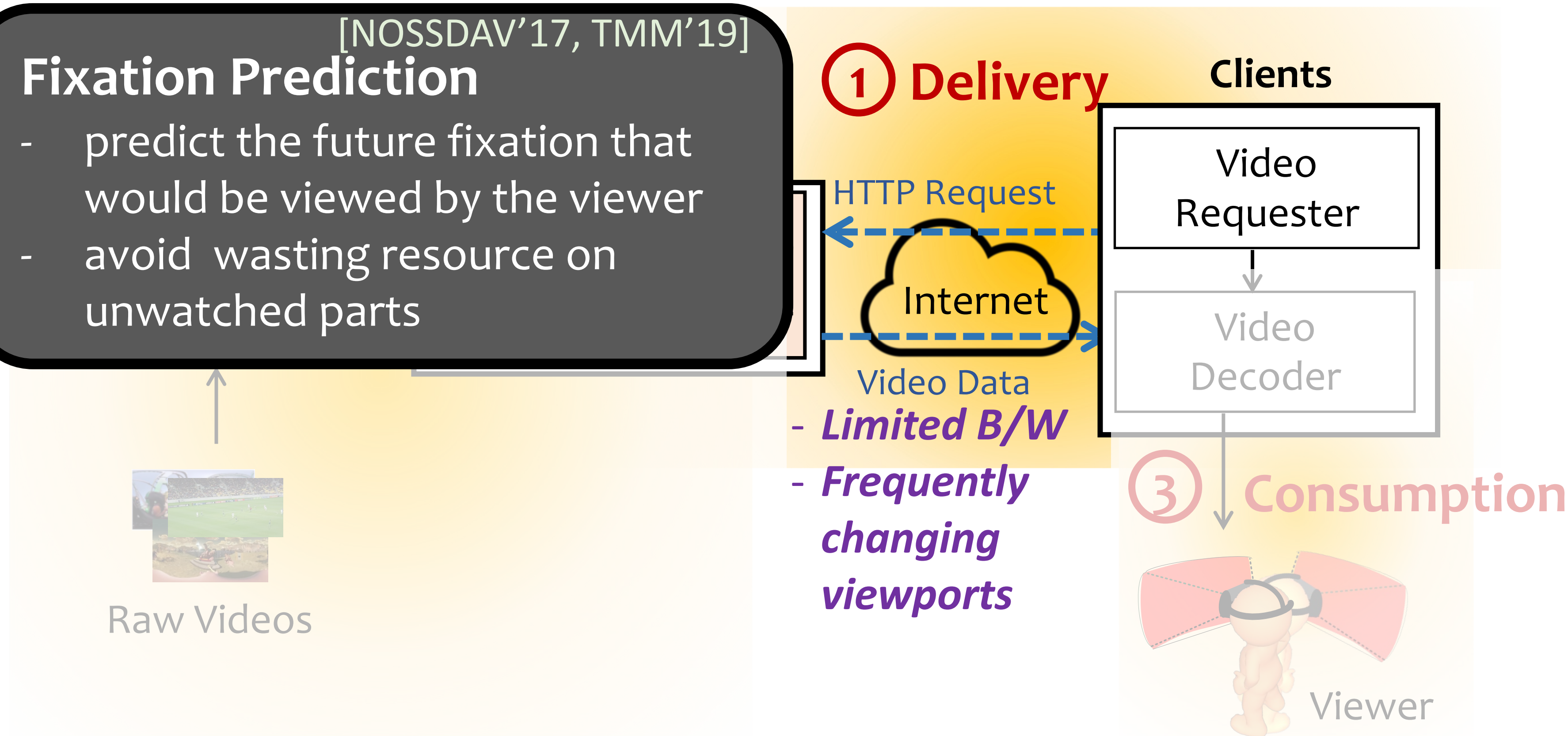


# 360° Video Streaming Platform

[NOSSDAV’17, TMM’19]

**Fixation Prediction**

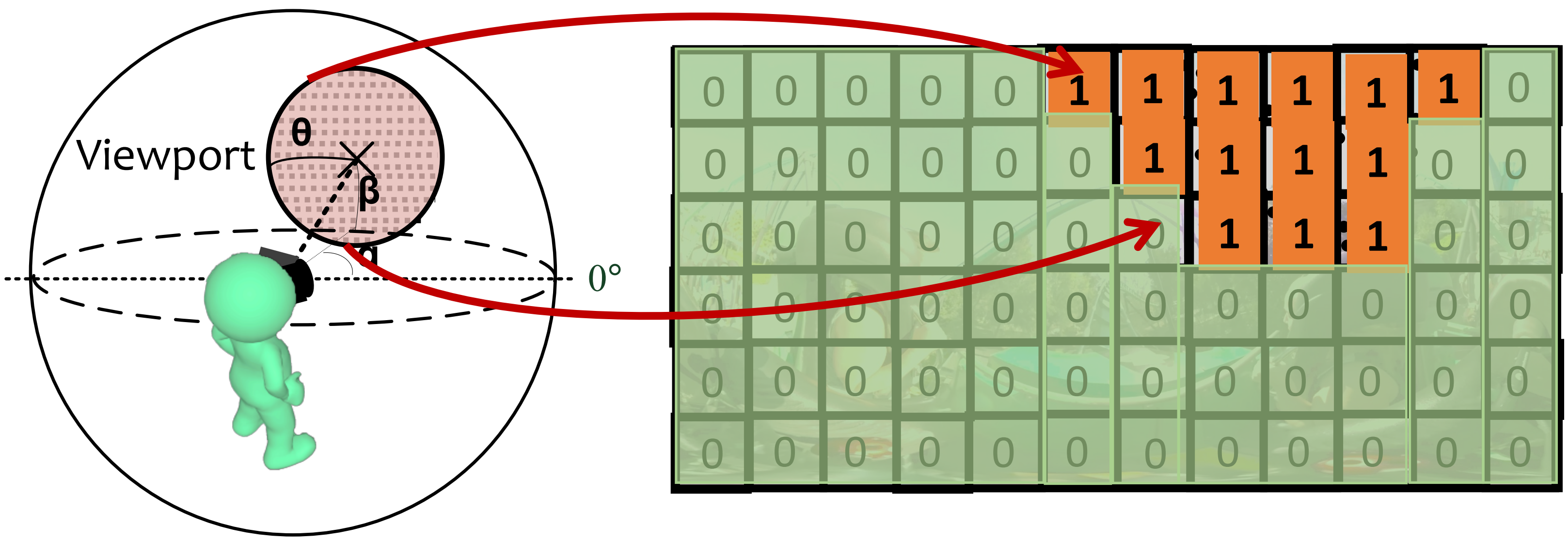
- predict the future fixation that would be viewed by the viewer
- avoid wasting resource on unwatched parts



# How to Save Bandwidth When Streaming 360 Videos?

- The HMD viewer only gets to see a small part of the whole 360° video (< **1/3** )

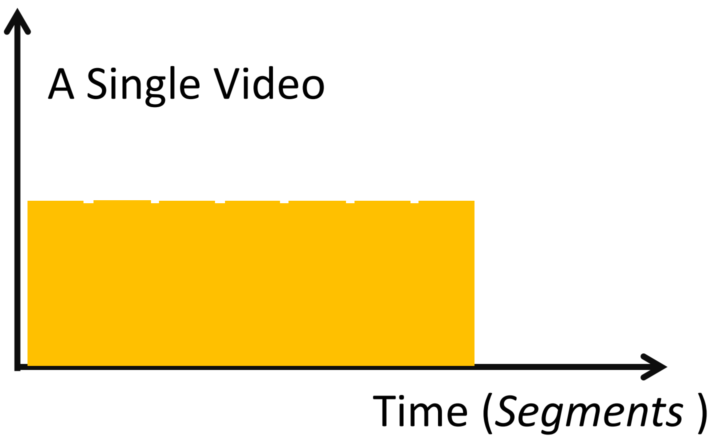
⇒ HEVC **Tiles**

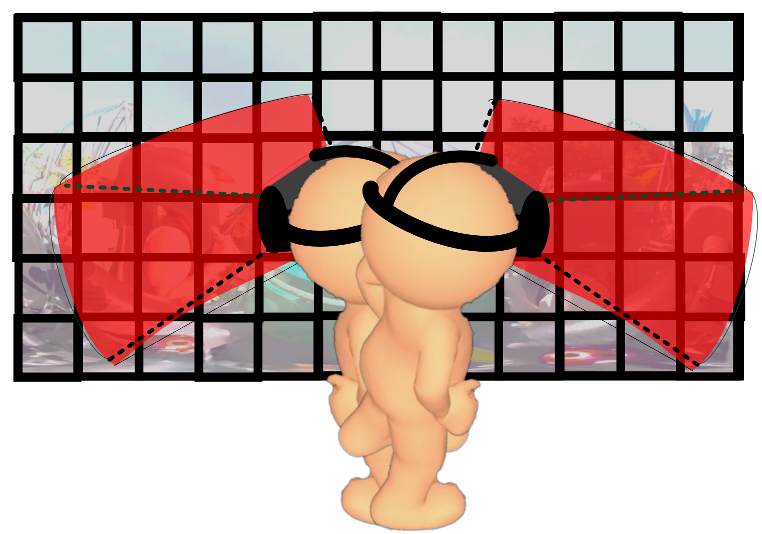


# Viewport-Adaptive Streaming

- *Tiling* with *MPEG* *DASH* *(Dynamic Adaptive Streaming over HTTP)*

***Temporal*** ***Spatial***



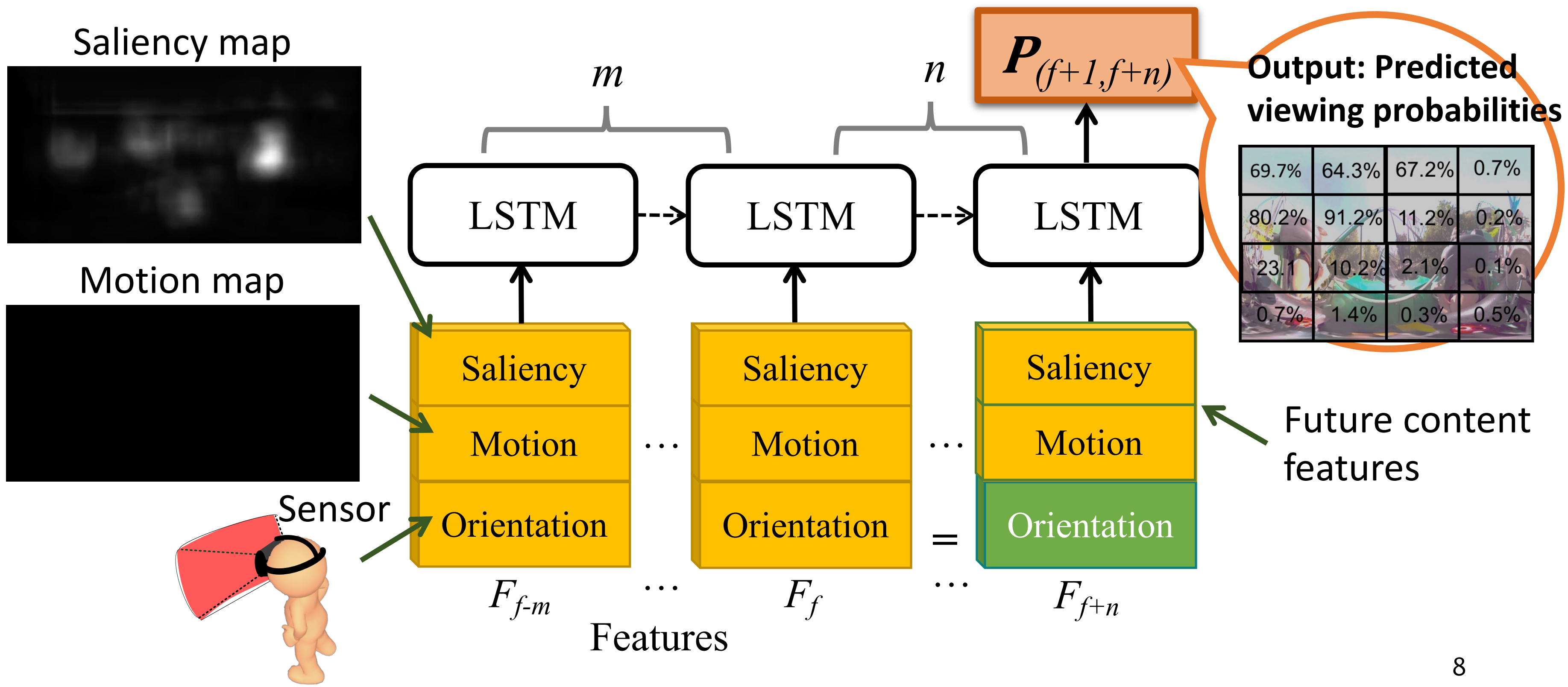



- Basic transmission unit: **Tiled-segments**

# Fixation Prediction Results

# LSTM-Based Neural Networks

- Future-aware network works the best
  - Sensor features: viewer's yaw, roll, and pitch
  - Content features: saliency maps and motion maps



# The Adopted Saliency Maps in the Content Features are Faulty

- Existing saliency detection networks are typically trained with photos taken by 2D cameras
- Existing codecs do not support spherical videos

→Distortion due to mapping spherical videos to other coordinate system

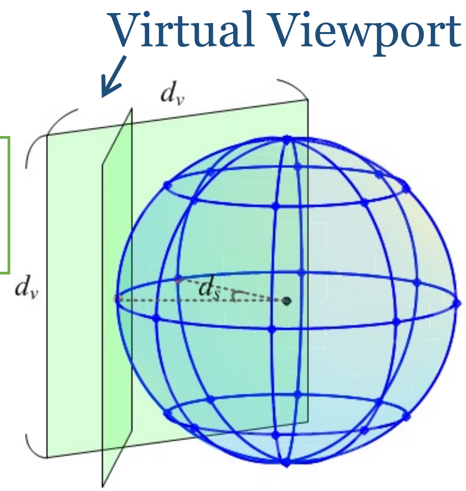
- E.g., shape distortion and ill segmentation

⇒ **We need a new model !**

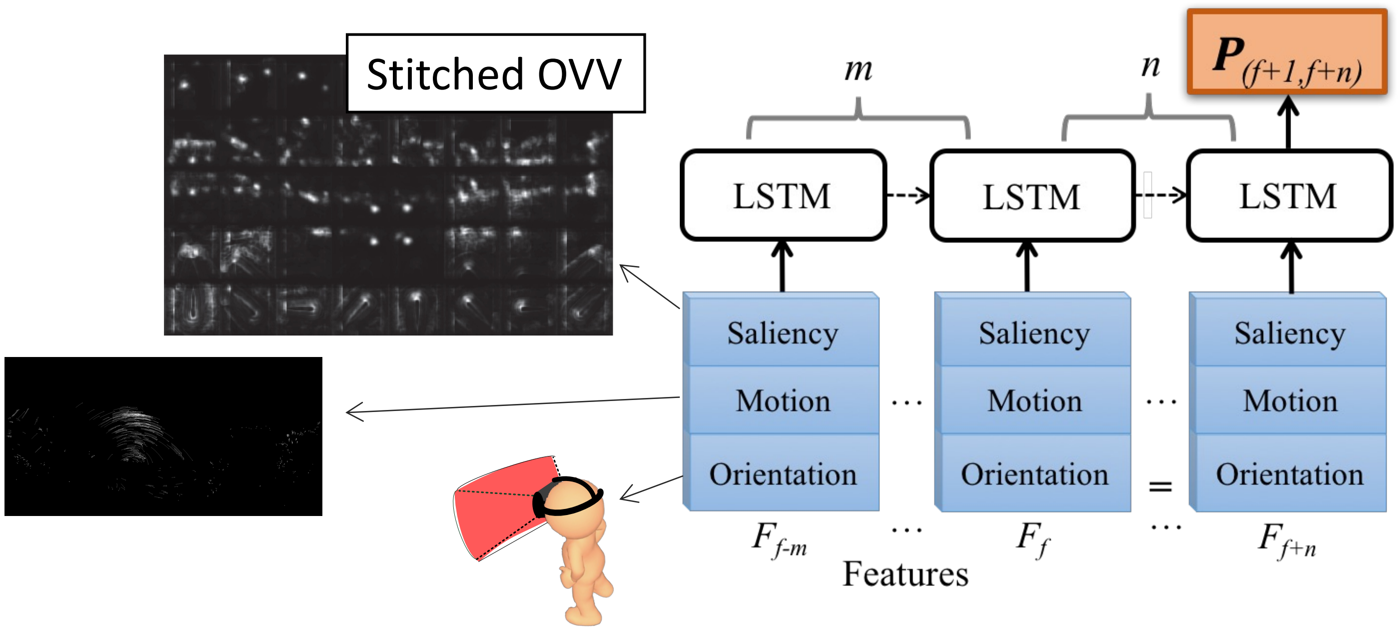
# Overlapping Virtual Viewport (OVV)

- OVV covering the whole sphere space
  - $d_v$: viewable degree
  - $d_s$: sampling degree

Example of $d_v = 90^\circ$ and $d_s=45^\circ$




⇒ free from **shape distortion** and **ill segmentation**



# Evaluations

10 videos (1800 frames) and 50 viewers = 900k samples

| Category | Videos |
|---|---|
| NI, fast-paced | Mega Coaster |
| | Roller Coaster |
| | Driving with |
| NI, slow-paced | Shark Shipwreck |
| | Perils Panel |
| | Kangaroo Island |
| | SFR Sport |
| CG, fast-paced | Hog Rider |
| | Pac-Man |
| | Chariot Race |

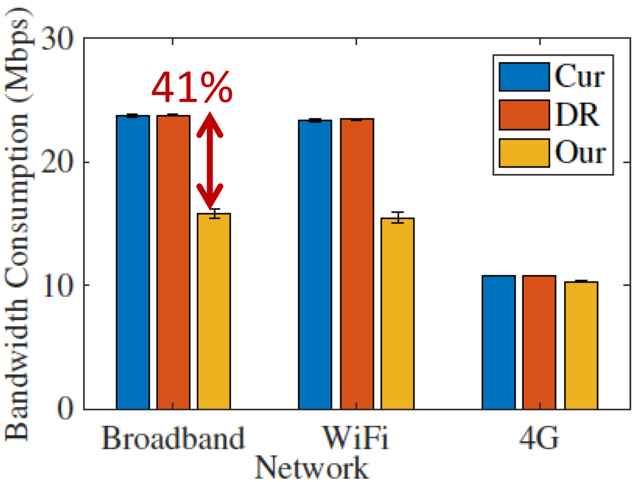
- Prediction
  - Higher accuracy and F-score
- Streaming in ns-3 simulator
  - Lower bandwidth consumption, lower rebuffering time, and comparable video quality

41% Bandwidth Saving

< -1 dB V-PSNR

- Small-scale user study
  - Lower MOS score by < 0.1 (out of 5) while saving 41% of bandwidth compared to the current practice

| Prediction Algorithm | | Accuracy | F-Score |
|---|---|---|---|
| Our | | 81.8% | 63.1% |
| CUB360 [1] | $K$=0 | 73.1% | 31.0% |
| | $K$=2 | 73.0% | 53.4% |
| | $K$=5 | 73.0% | 54.3% |
| | $K$=10 | 72.2% | 54.6% |




[1] Y. Ban, L. Xie, Z. Xu, X. Zhang, Z. Guo, and Y. Wang, "Cub360: Exploiting cross-users behaviors for viewport prediction in 360 video adaptive streaming," in Proc. of IEEE International Conference on Multimedia and Expo (ICME'18), 2018, pp. 1–6.

# State-of-the-Art Prediction Algos

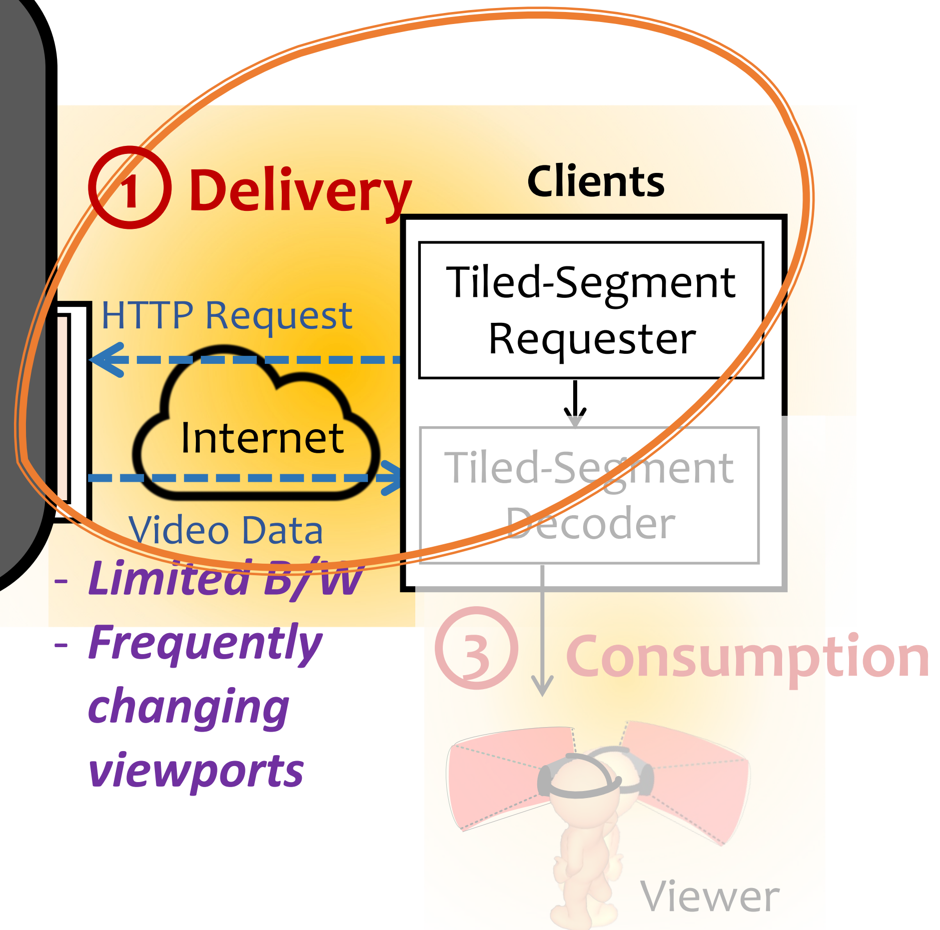
| Approach | Classification | Literature |
|---|---|---|
| LSTM | None | Fan et al. 2017, Fan et al. 2019, Nguyen et al. 2018, Xu et al. 2018, Hou et al. 2019, Hou et al. 2020 |
| CNN + LSTM | None | Xu et al. 2018, Chen et al. 2020, Feng et al. 2020, Cheng et al. 2018 |
| Spherical CNN | None | Zhang et al. 2018, Wu et al. 2020 |
| Others | None | Bai et al. 2017, Qian et al. 2018, Xu et al. 2018, Vielhaben et al. 2019, Xu et al. 2018 |
| Others, e.g, SVM, LR, RL | Video content, viewer's behavior, or per video | Feng et al. 2019, Nasrabadi et al. 2020, Ban et al. 2018, Xie et al. 2018 |

# Tiled 360° Video Streaming Platform

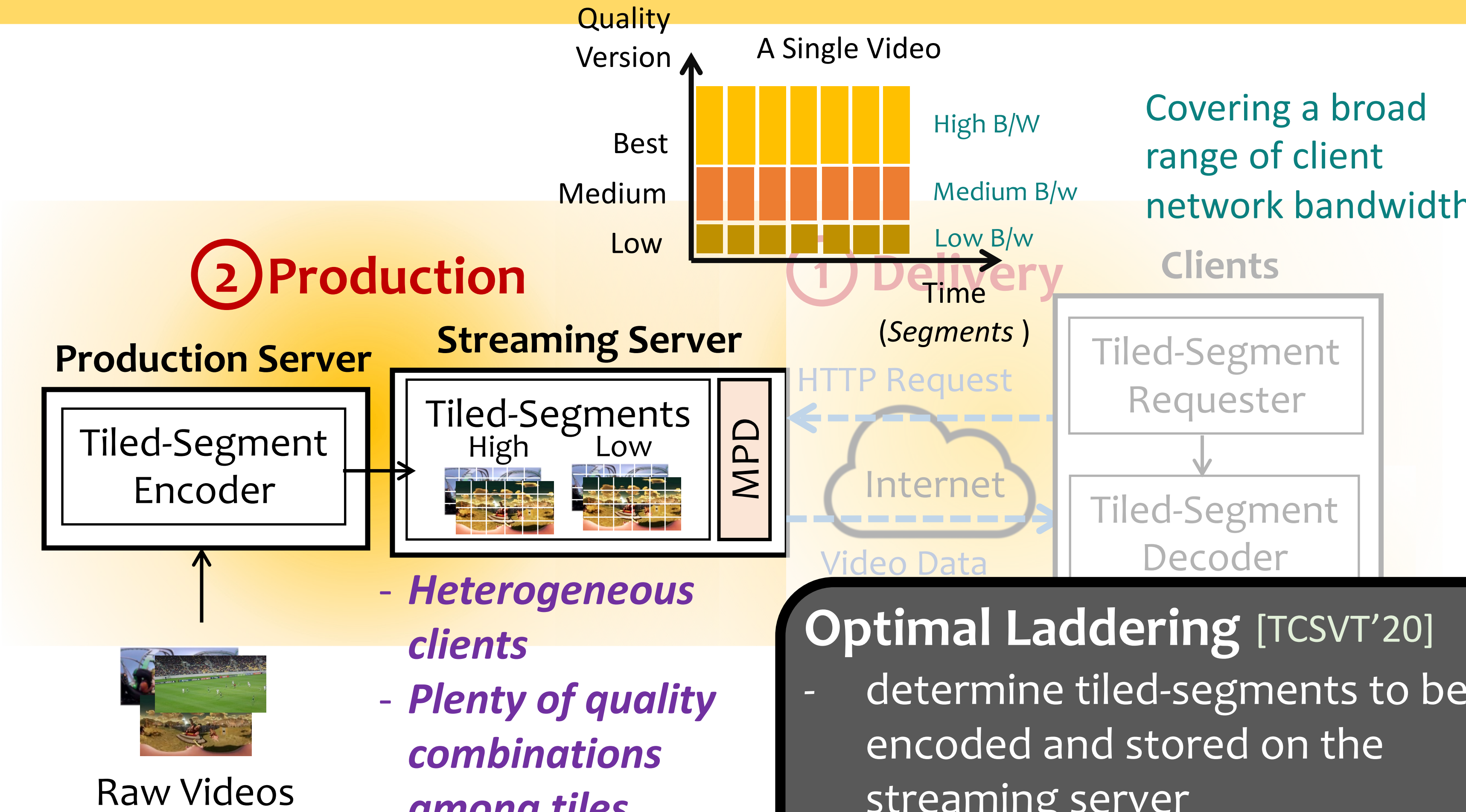
[NOSSDAV'17, TMM'19]

**Fixation Prediction**

- predict the future tiled-segments that would be viewed by the viewer
- **leverage LSTM with sensor and content features**
- ***leads to comparable video quality while saving up to 41% of bandwidth***

# Tiled 360° Video Streaming Platform



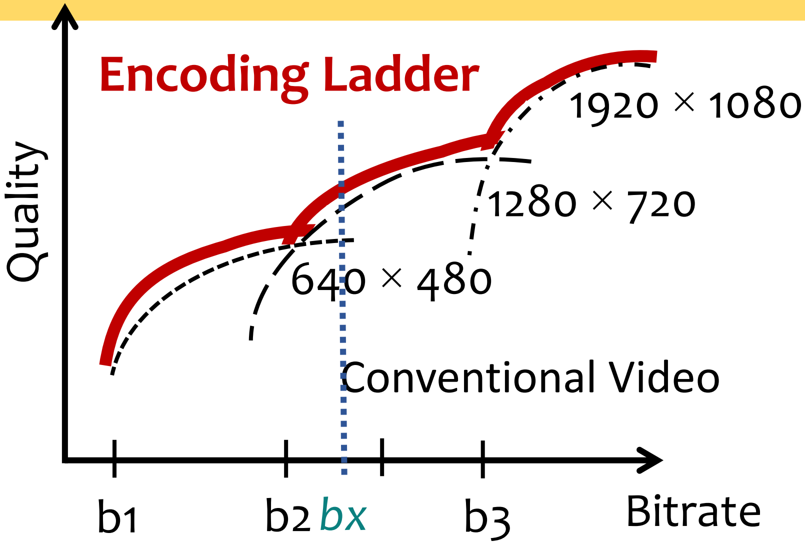

- ***Heterogeneous clients***
- ***Plenty of quality combinations among tiles***
- ***Limited space***

**Optimal Laddering** [TCSVT'20]

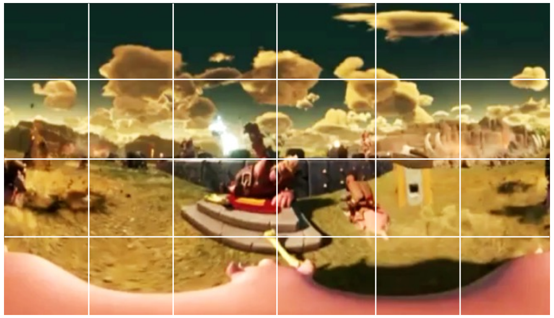
- determine tiled-segments to be encoded and stored on the streaming server
- maximize overall viewing quality

# Optimal Laddering Problem

- Determine the optimal encoding ladder to cover a broad range of clients
- Challenges for tiled 360° videos
    - Different tiles have different characteristics and lead to huge amount of quality version combinations
    - Storage space is limited




Clients with b/w at *bx* request the video in 1280 × 720 resolution

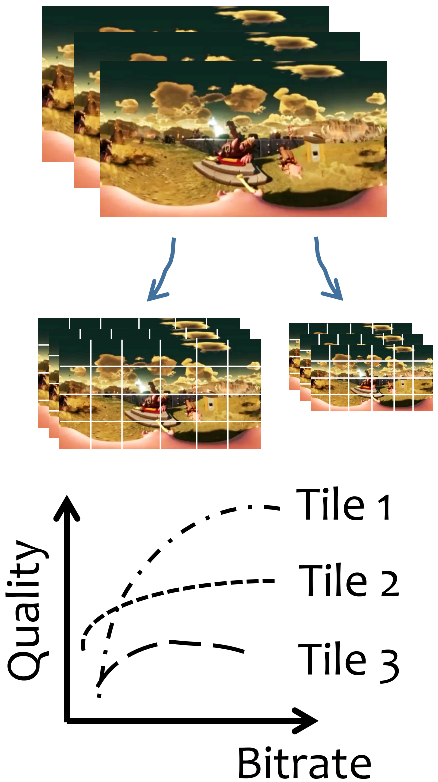


# Problem Statement

Bandwidth/Videos

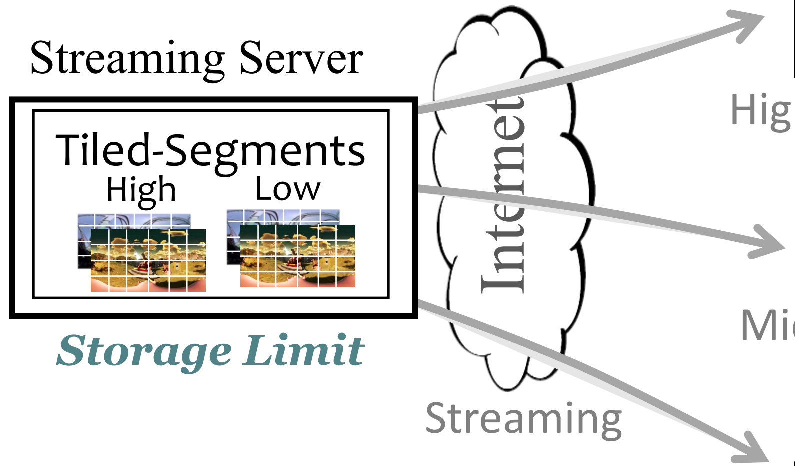
***Client Distribution***

**Goal: Maximize the overall viewing quality of clients**

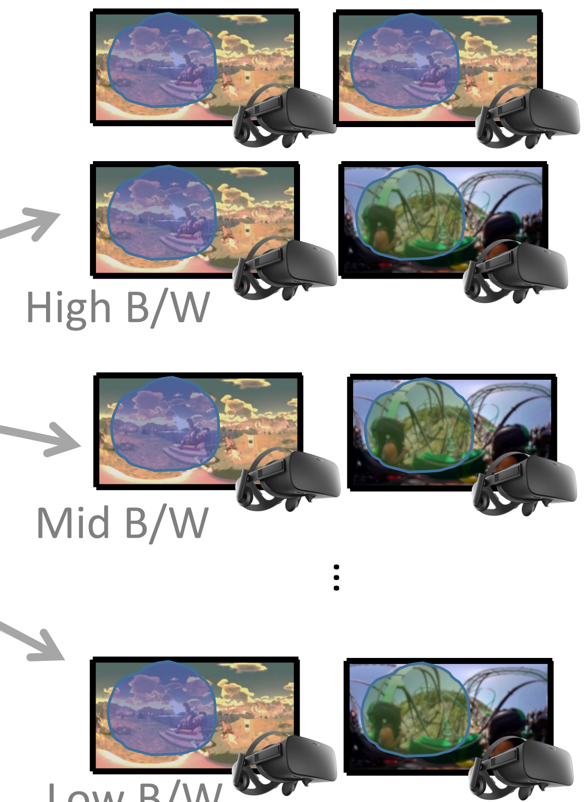



***Video Models***

Tile Complexity




***Storage Limit***




**Which version of tiled-segment of each video should be stored on the server?**

***Viewing Probability***

Tile Importance

# Problem Formulation

distortion model

$$\min \sum_{c=1}^{C} \sum_{\phi \in \Phi} f_{\phi,c} p_{\phi} a_{\phi} \sum_{q=1}^{Q} d_{\phi}(q) \underline{x_{\phi,c,q}}$$

Minimize the overall client distortion

$$st: \sum_{n=1}^{N} \sum_{q=1}^{Q} r_{\phi}(q) \underline{x_{\phi,c,q}} \le b_c$$

bitrate model

The bitrate of the tiled-segment streamed to each class is bounded by the available bandwidth

$$\sum_{\phi \in \Phi} \sum_{q=1}^{Q} r_{\phi}(q) \underline{y_{\phi,q}} \le S;$$

The required size for storing tiled-segments is bounded by the storage limit

$$\underline{x_{\phi,c,q} \le y_{\phi,q}}$$

Only the tiled-segments stored on the server can be selected to be streamed to clients

$$\sum_{q=1}^{Q} \underline{x_{\phi,c,q}} = 1$$

Only one version of tiled-segment is selected for each class

$$\boxed{\begin{array}{l} x_{\phi,c,q} \in \{0,1\} \\ y_{\phi,q} \in \{0,1\} \end{array}}$$

$c \in [1,C], q \in [1,Q], \phi \in \mathbf{\Phi};$

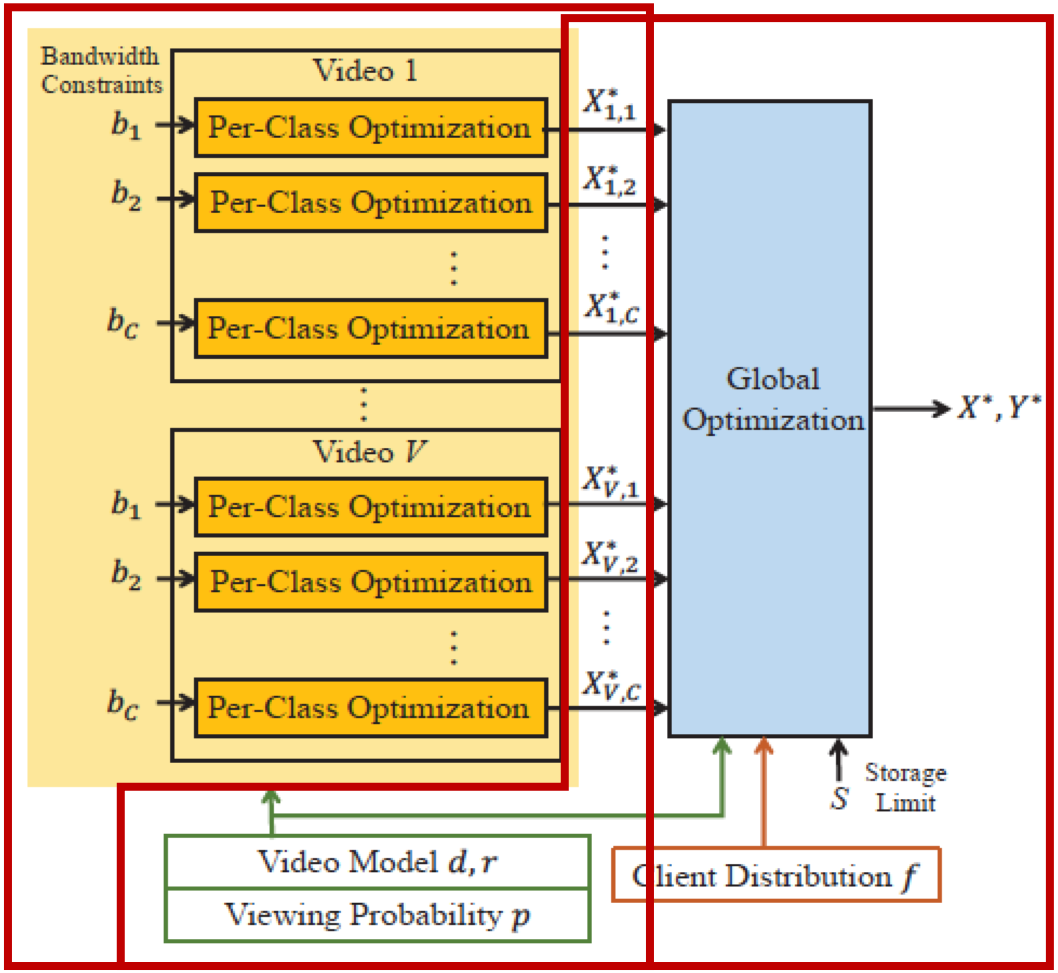
$q \in [1,Q], \phi \in \mathbf{\Phi}.$

$\phi = (v,t,n)$

$\mathbf{\Phi} = \{(v,t,n) | v \in [1,V], t \in [1,T], n \in [1,N]\}$

# Decompose the Problem (Divide-and-Conquer)

- ***Per-class optimization***: minimize the distortion under the *bandwidth constraint* for each class
- ***Global optimization***: minimize the overall distortion under the *storage limit*



# Sample Formulation: Per-Class Optimization

$$\min \sum_{t=1}^{T} \sum_{n=1}^{N} p_{v,t,n} a_n \sum_{q=1}^{Q} d_{v,t,n}(q) x_{v,t,n,c,q}$$

$$st: \sum_{n=1}^{N} \sum_{q=1}^{Q} r_{v,t,n}(q) x_{v,t,n,c,q} \leq b_c$$

$$\sum_{q=1}^{Q} x_{v,t,n,c,q} = 1$$

$$x_{v,t,n,c,q} = \{0, 1\}$$

Minimize the viewing distortion of class

The bitrate is bounded by the available bandwidth

- Lagrangian-Based Algorithm (PC-LBA)
  - leverages the **convexity** of the video models
- Greedy-Based Algorithm (PC-GBA)
  - runs more efficiently

# LBA to Solve the Subproblem

Convex Optimization

- Leverage the Lagrangian Multiplier to transform the constrained problem into an **unconstrained problem**

Objective $\min \sum_{n=1}^{N} d_{v,t,n}(\underline{\kappa_{v,t,n}}) p_{v,t,n} a_n$

Decision Variable QP

Constraint $st\colon \sum_{n=1}^{N} r_{v,t,n}(\kappa_{v,t,n}) \le b_n$

Lagrangian Multiplier

**Unconstrained problem**

$$\min \ L(\mathbf{K_{v,t,c}}, \mu) = \sum_{n=1}^{N} d_{v,t,n}(\kappa_{v,t,n,c} + \underline{\mu}(\sum_{n=1}^{N} r_{v,t,n}(\kappa_{v,t,n,c}) - b_c)$$

Objective Constraint

**QP**

$$\longrightarrow \kappa_{v,t,n,c} = \frac{1-\beta^d_{v,t,n}}{\beta^r_{v,t,n}} W\left(\frac{\beta^r_{v,t,n}}{1-\beta^d_{v,t,n}} \mathrm{e}^{\frac{-\ln \frac{\mu \alpha^r_{v,t,n} \beta^r_{v,t,n}}{-\alpha^d_{v,t,n} \beta^d_{v,t,n} p_{v,t,n} a_n}}{1-\beta^d_{v,t,n}}}\right)$$

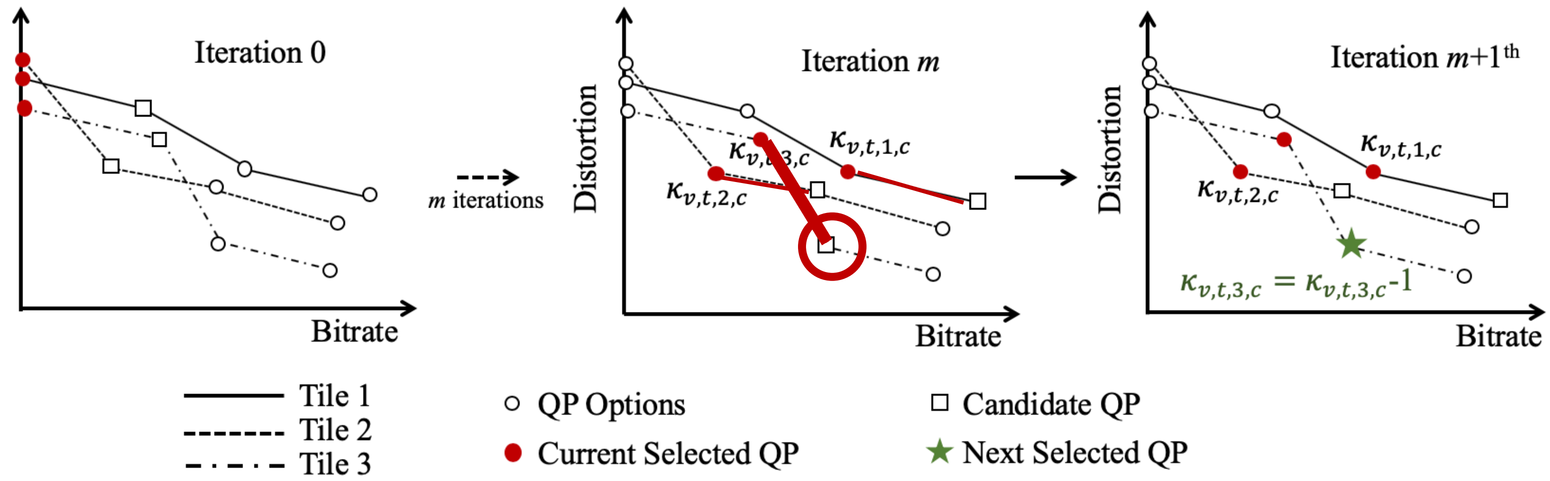
# Greedy-based: PC-GBA

- Iteratively allocate more bitrate to the tile with the highest coding efficiency by reducing its QP
  - until there is no remaining bandwidth or all tiles are coded at the smallest QP

Weighted distortion reduction

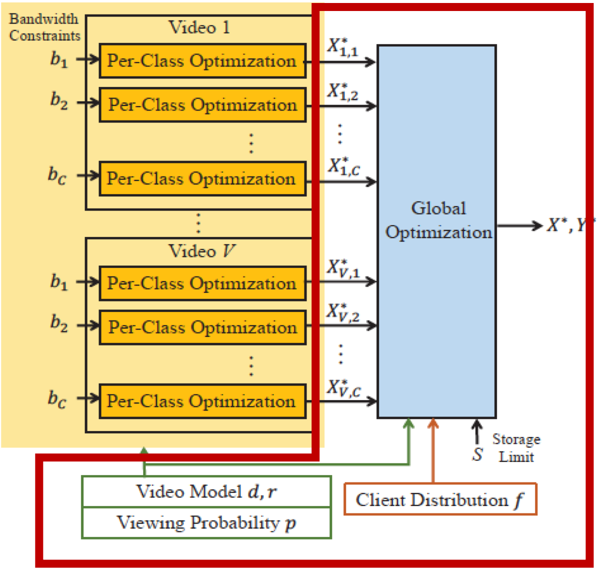
$$\boxed{\theta_{\phi,c} = \frac{(d_\phi(\kappa_{\phi,c}-1) - d_\phi(\kappa_{\phi,c}))p_\phi a_\phi}{r_\phi(\kappa_{\phi,c}-1) - r_\phi(\kappa_{\phi,c})}}$$

Bitrate increment



# Global Optimization

- Greedily adjust the per-class solutions $\mathcal{X}^*_{v,c}$ to minimize the expected distortion while meeting both the client bandwidth constraints and *overall server storage limit*
  - iteratively select the tiled-segment with the minimum $\epsilon_{\phi,q}$




Weighted distortion gain

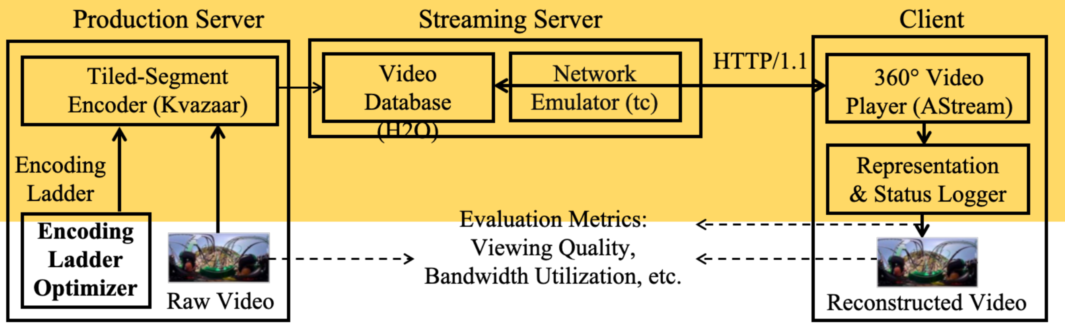
step size

$$\epsilon_{\phi,q} = \frac{\sum_{v=1}^{V}\sum_{c=1}^{C} f_{v,c}\cdot[d_\phi(q+\delta)-d_\phi(q)]p_\phi a_\phi x_{\phi,c,q}}{[r_\phi(q)-r_\phi(q+\delta)(1-y_{\phi,q+\delta})]y_{\phi,q}}$$

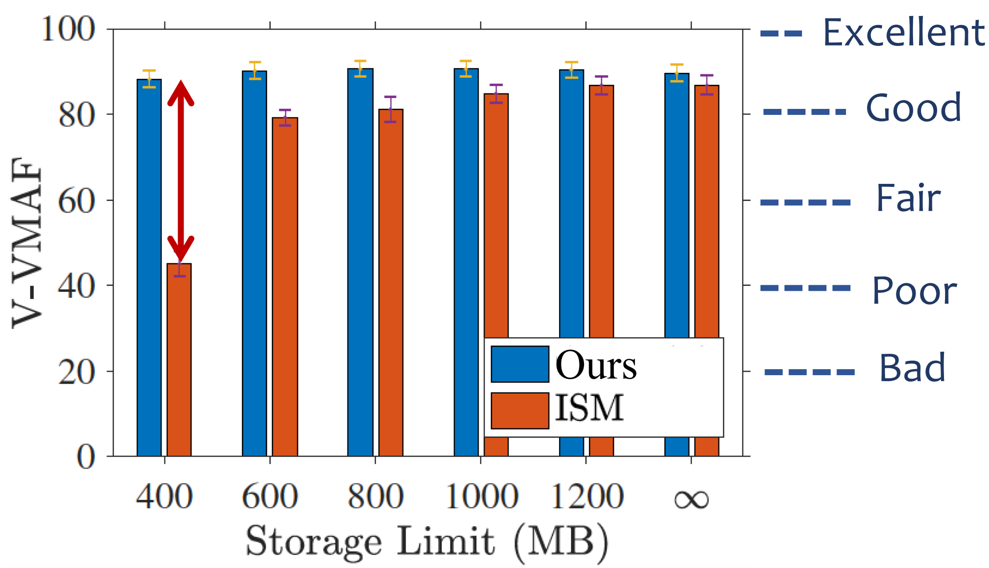
Reduced storage size on server if the QP value of tiled-segment increases

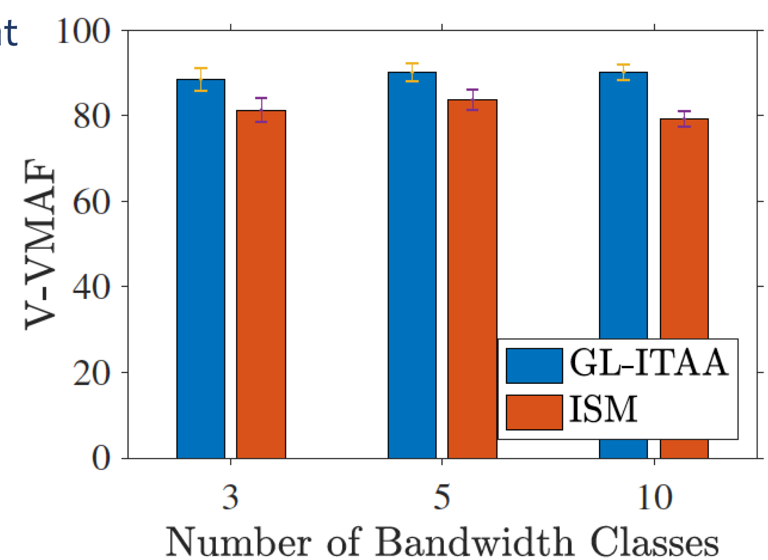
already selected to be stored on the server or not

# Sample Results




- User's bandwidth follows the distribution in Cisco's report [5]
- An ABR for 360 videos [6] is employed during streaming

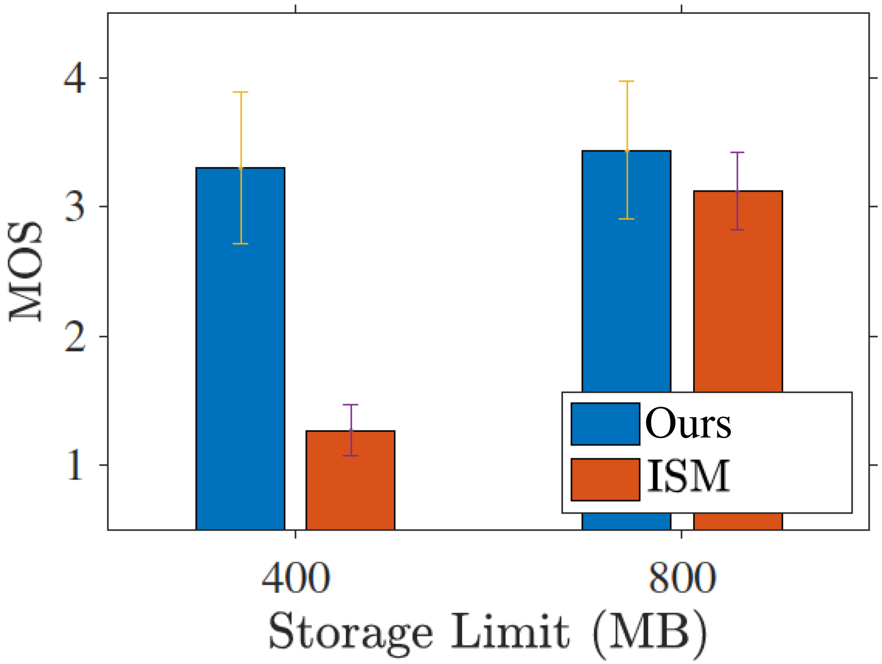



Our solution **outperforms ISM by up to 43.14 in V-VMAF** and
**has good scalability under both storage limits and bandwidth classes**

# User Study Evalutation

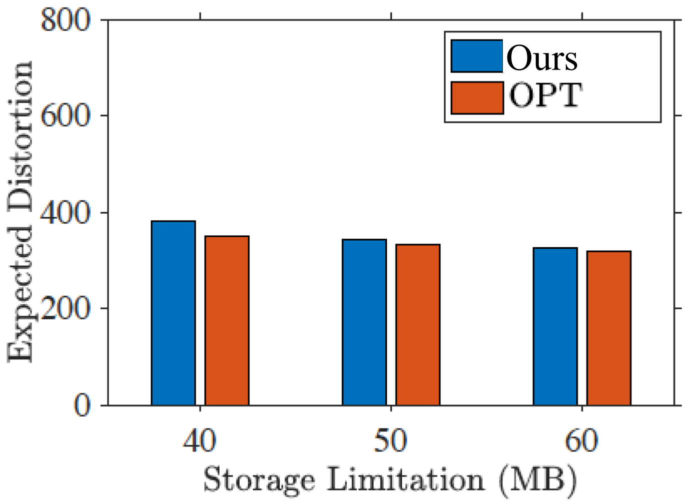
- 12 subjects watch the 12 viewport videos from a random user trace (6 video × 2 storage limits)
- MOS [1,5]

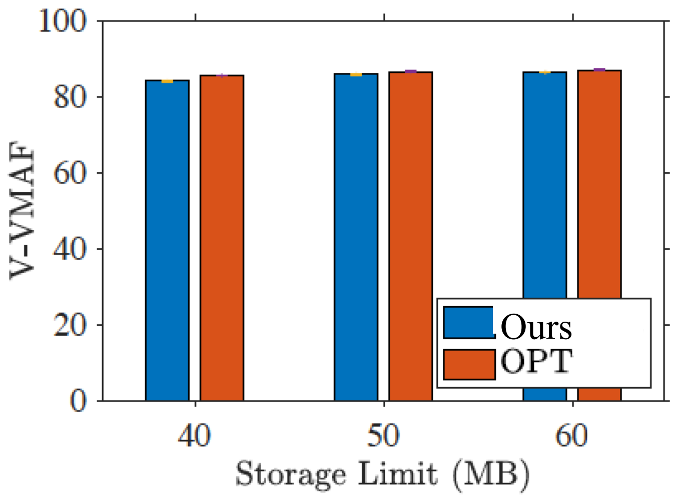



Our solution **outperforms ISM** and **has good scalability under different storage limits**

# Comparison with the Optimal Solution

- OPT directly solves the ILP problem using CPLEX
- Reduced problem size:
  $C = 3, T = 15$, and $S = \{40, 50, 60\}$ MB




Our solution achieves **very close expected distortion** and **actual viewing quality (V-VMAF)** to OPT

**Run at least 8.5 times faster than OPT**

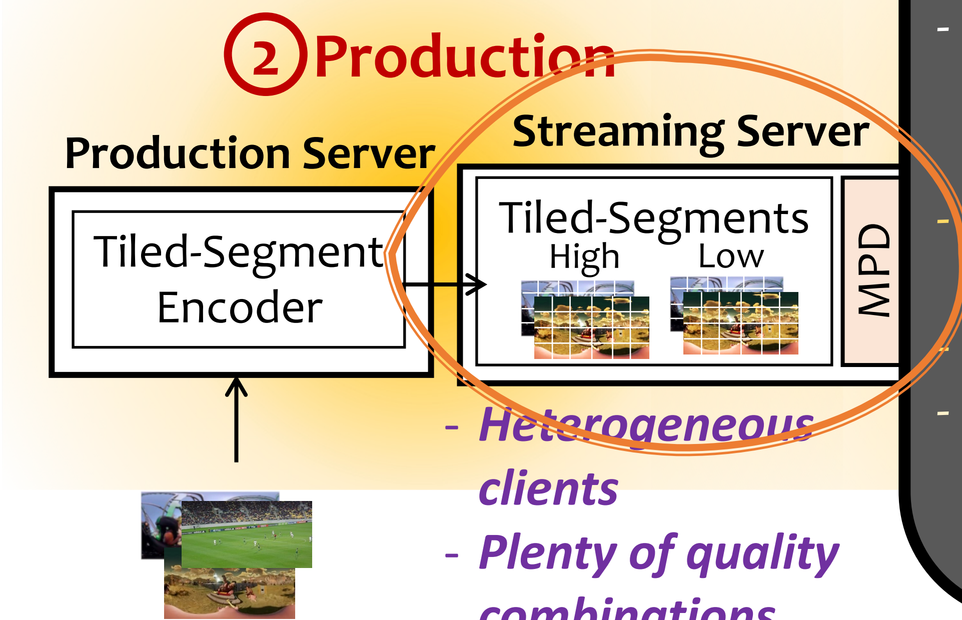
# Fairness Among Client Classes

- Max-min fairness:
  maximize the minimum allocated resource for any clients
- Objective: $\min_{1\le c\le C, 1\le v\le V} \max D_{v,c}$

- Jain's fairness index:
  $$J(f_1, f_2, \cdots, f_N) = \frac{(\sum_{n=1}^{N} f_n)^2}{N \sum_{n=1}^{N} f_n^2} = \frac{1}{1+\widehat{\nu_f}^2}$$
- Objective:
  $$\max \frac{(\sum_{v=1}^{V} \sum_{c=1}^{C} D_{v,c})^2}{V \sum_{v=1}^{V} C \sum_{c=1}^{C} D_{v,c}^2} = \max \frac{1}{1+\widehat{\nu_D}^2}$$

- *The revised solution:*
  - *Per-class optimization: minimize the distortion of each class, which is restricted by $b_c$*
  - *Global optimization: iteratively increases the QP of the tiled segment having the lowest $\epsilon_{v^*,t,n,c^*,q}$, where $(v^*, c^*) = arg \min_{v\in[1,V], c\in[1,C]} D_{v,c}$*

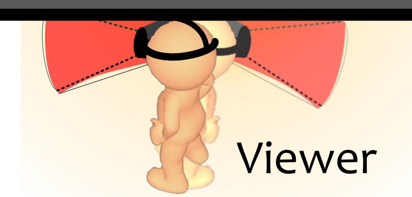
# Tiled 360° Video Streaming Platform



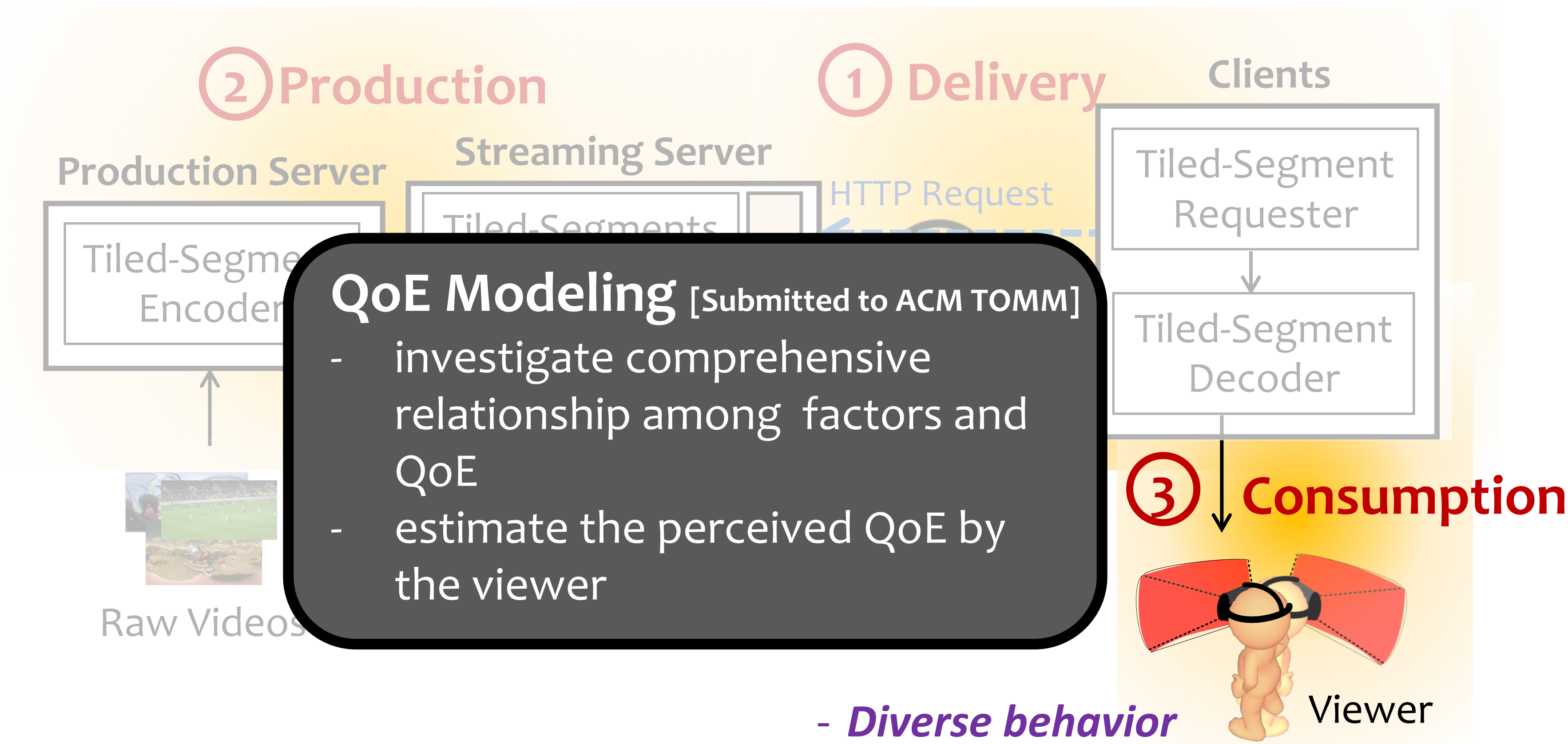

- *Heterogeneous clients*
- *Plenty of quality combinations among tiles*
- *Limited space*

**Optimal Laddering** [TCSVT'20]

- determine tiled-segments to be stored on the streaming server to maximize overall viewing quality
- **problem decomposition with divide-and-conquer mathematical optimization**
- *leads to higher viewing quality and better scalability under different storage limits*



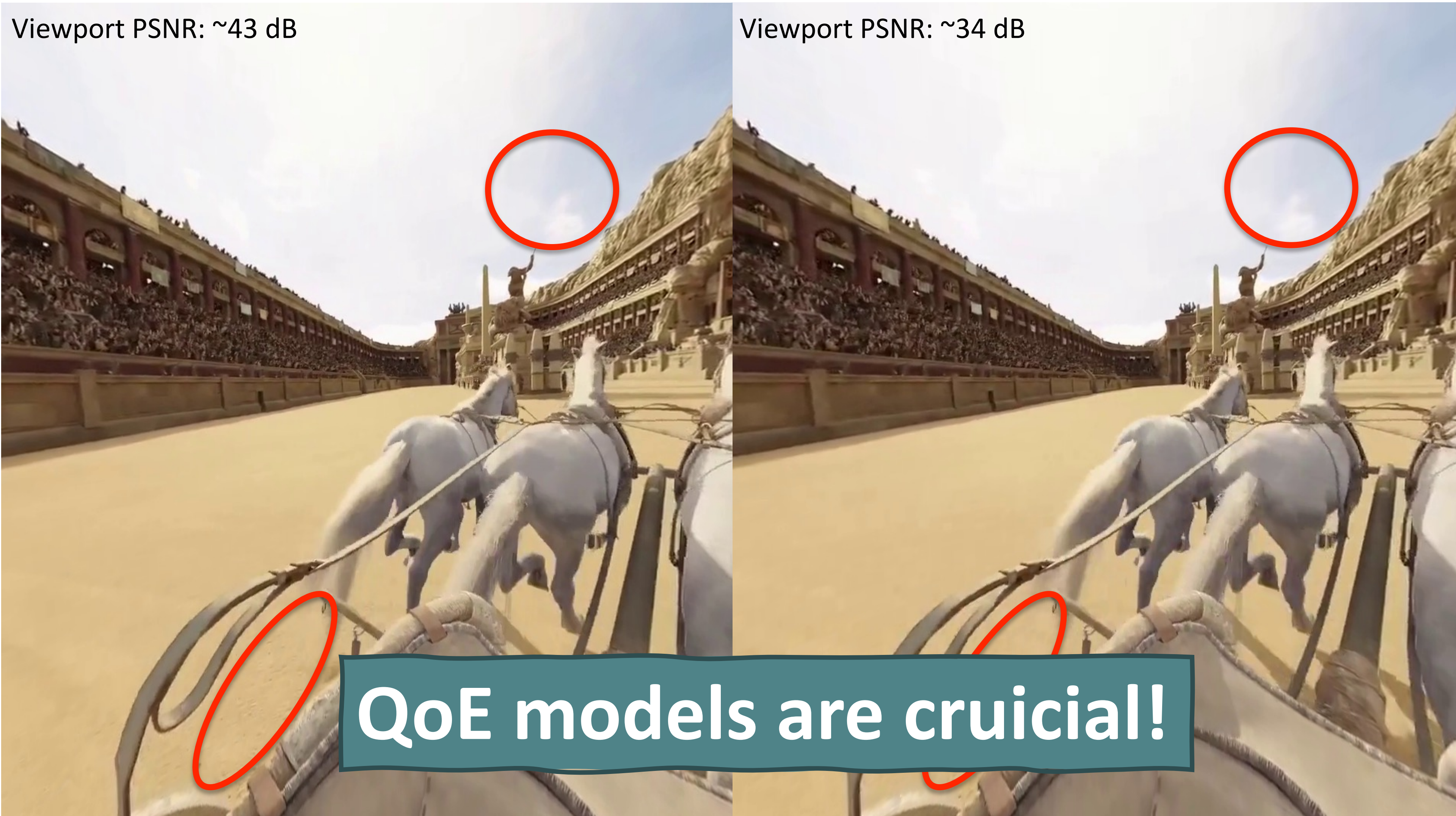
# Tiled 360° Video Streaming Platform




**QoE Modeling** **[Submitted to ACM TOMM]**
- investigate comprehensive relationship among factors and QoE
- estimate the perceived QoE by the viewer

- ***Diverse behavior***
- ***Complex and unknown QoE***

# Existing Quality Metrics Failed to Reflect Real User Experience



# QoE is Affected by Plenty of Factors

**The Composition of QoE**

***Overall QoE***

$\boldsymbol{OQ}$

......

***QoE Features***

$IQ \quad FG \quad IM \quad CS \quad AT$

- Perceived image quality, perceived cybersickness level, etc.

......

***QoE Factors***

$I_1 \quad I_2$

......

- Content factors: encoding bitrates, video types, etc.
- Human factors: gender, historical motion sickness level, etc.
- Context factors: environments, moving speeds, etc.
- System factors: video players, devices, etc.

**Definition**

Comprehensive user experience

Nameable perceived user experience aspects

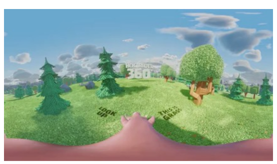
Primitive and measurable metrics

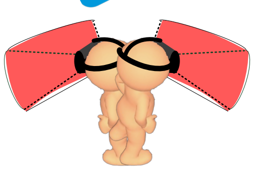
# QoE Features and Factors

- QoE Features

| | Feature | Question | Lowest Score (1) | Highest Score (9) |
|---|---|---|---|---|
| Overall QoE | - | How would you rate the overall quality? | Bad | Excellent |
| Image Quality | IQ | How would you rate the image quality? | Bad | Excellent |
| Fragmentation | FG | How would you rate the fragmentation level? | None | Severe |
| Immersion | IM | How would you rate the immersion level? | Bad | Excellent |
| Cybersickness | CS | How would you rate the perceived cybersickness level? | None | Severe |
| Attractiveness | AT | How would you rate the attractiveness level? | Not Attractive | Attractive |

- QoE Factors

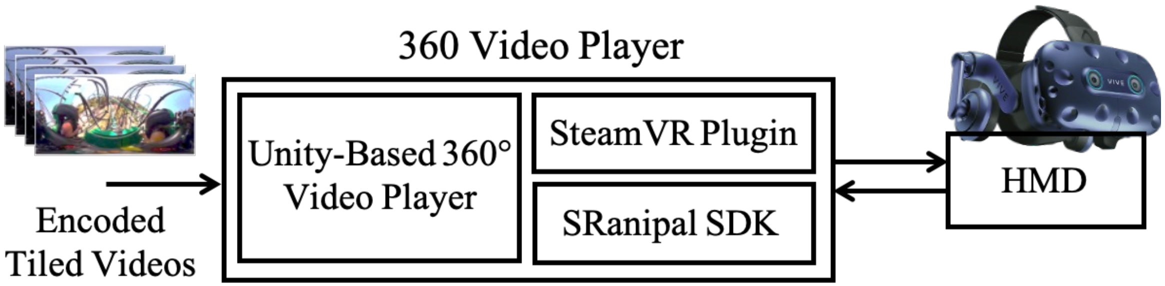




  - ***Content factors***: bitrate, complexity, motion, video quality, video quality variance
  - ***Human factors***: gender, historical sickness, avg. head/gaze rotation speed
  - ***Context factors***: head/gaze rotation speed, viewport complexity, viewport motion, viewport quality, viewport quality variance

# Testbed and Test Videos

- Unity-based testbed with eye-tracking feature




- Test videos
  - 6 raw videos from JVET, ERP to EAC, 3840×1920, 20 seconds
  - 12x8 tiles, bitrates: 1, 3, 6, 9, 12, 15 Mbps

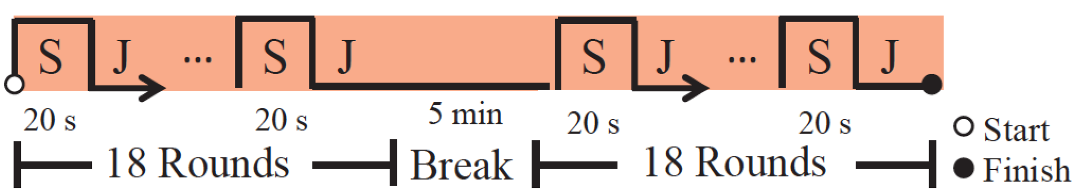
| Category | Video | Resolution | Frame Rate |
|---|---|---|---|
| Fixed Camera | SkateboardTrick | 8192x4096 | 60 fps |
| | Harbor | 8192x4096 | 30 fps |
| | PoleVault | 3840x1920 | 30 fps |
| Moving Camera | Landing | 6144x3072 | 30 fps |
| | Balboa | 6144x3072 | 30 fps |
| | BranCastle | 6144x3072 | 30 fps |

# Subjects and Procedure

- 24 Subjects

| | |
|---|---|
| **Gender** | Male: 58%, Female: 42% |
| **Age** | Range: [19,30], Standard Deviation: 2.78 |
| **HMD Experience** | Never: 4%, Seldom: 79%, Medium: 17% |
| **Vision Correction** | Glasses: 13%, Contacts: 75%, None: 12% |
| **Education** | High School: 37%, Bachelor: 42%, Master: 21% |

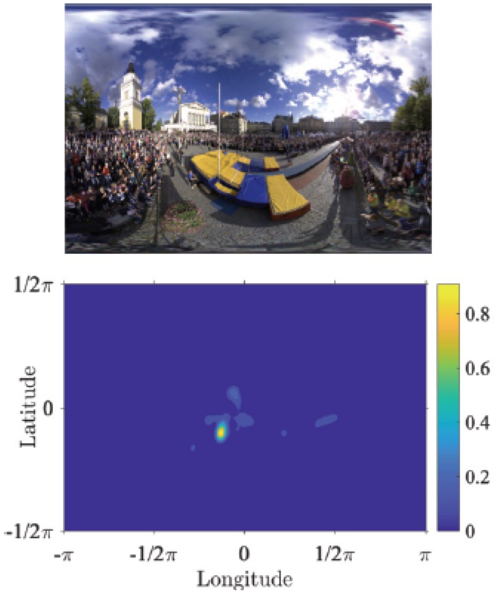
- Procedure follows ITU-T 910
  - Absolute Category Rating (ACR)
  - Score: [1,9]

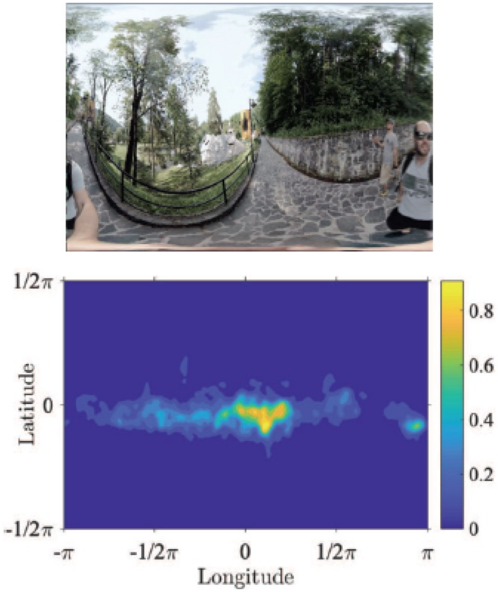


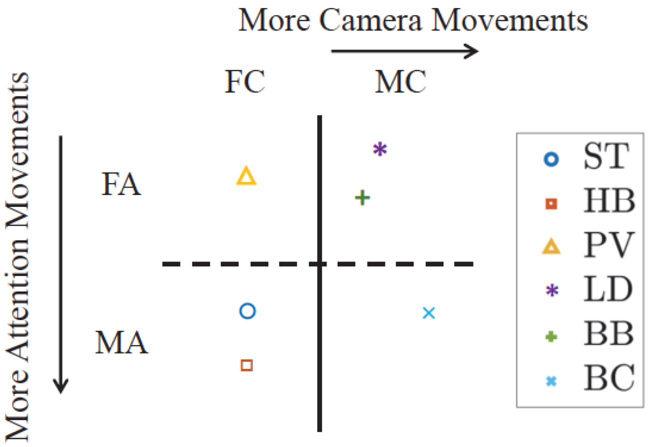

[1] Jukka Hakkinen, Tero Vuori, and M Paakka. 2002. Postural stability and sickness symptoms after HMD use. In IEEE International Conference on Systems, Man and Cybernetics, Vol. 1. 147–152.

# Analysis

- Different videos drive different viewing behaviors



# QoE Modeling

- Overall QoE, IQ, FG, IM, CS
  - Mean Opinion Score (MOS) and Individual Score (IS)
- Dataset: 70% training set (5-fold validation)
- Metrics: Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SROCC)
- Regressors

| *Regressor* | *Parameters* | | | *Training Set* | | *Validation Set* | |
|---|---|---|---|---|---|---|---|
| | | | | *PLCC* | *SROCC* | *PLCC* | *SROCC* |
| *Linear* | - | | | **0.9925** | **0.9823** | **0.9518** | **0.9175** |
| *Random Forest* | *Max No. Features* | *No Estimators* | *Max Depth* | 0.9686 | 0.9501 | 0.9215 | 0.8541 |
| | *auto* | *200* | *8* | | | | |
| *Gradient Boosting* | *Max No. Features* | *No Estimators* | *Learning Rate* | 0.9934 | 0.9761 | 0.9451 | 0.8962 |
| | *sqrt* | *100* | *0.01* | | | | |
| *Support Vector* | *Max Iterations* | *C* | *$\epsilon$* | 0.9880 | 0.9730 | 0.9350 | 0.9021 |
| | *20* | *10* | *0.05* | | | | |

# MOS Modeling

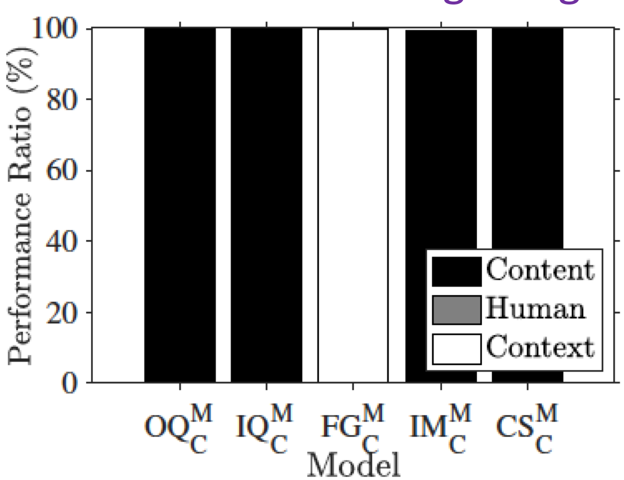
- Our derived models model well on the overall QoE and QoE features using all factors (***content, human, and context***)

| Model | OQ | IQ | FG | IM | CS |
|---|---|---|---|---|---|
| PLCC | 0.988 | 0.989 | 0.980 | 0.944 | 0.908 |
| SROCC | 0.971 | 0.977 | 0.975 | 0.889 | 0.902 |

**PLCC > 0.90**
**SROCC > 0.88**

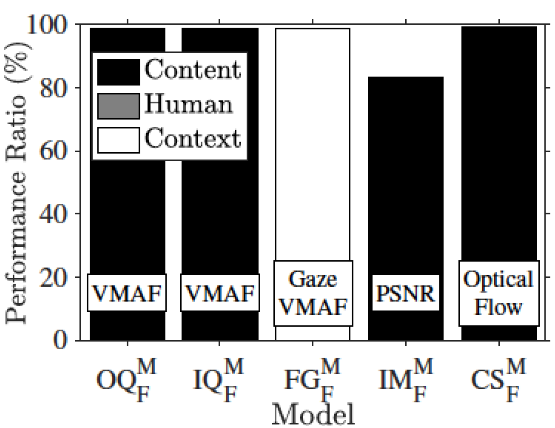
*Performance ratio: Normalize to the model using all factors*

with the dominating category




Content dominates the category: > 98% performance ratio

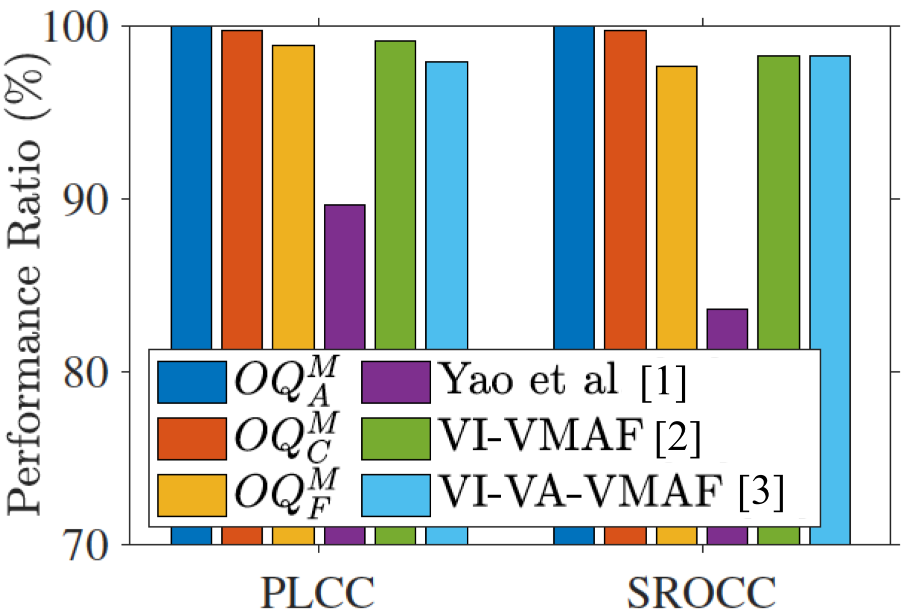
with the dominating factor




- (Gaze) VMAF dominates the factors for OQ, IQ, and FG
- Optical flow dominates the factors for CS

# Compared to the State-of-The-Art

| Model | QoE Factor | | | Overall QoE | QoE Feature | | | | Model Type | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Content | Human | Context | | IQ | FG | IM | CS | MOS | IS |
| **Ours** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Yao et al. [1] | ✓ | | | ✓ | | | | | ✓ | |
| VI-VMAF [2] | ✓ | | | ✓ | | | | | ✓ | |
| VI-VA-VMAF [3] | ✓ | | ✓ | ✓ | | | | | ✓ | |




- $OQ_A^M$ and $OQ_C^M$ outperform other state-of-the-art QoE models
- VI-VMAF outperforms $OQ_F^M$


[1] S. Yao et al. Towards Quality-of-Experience Models for Watching 360° Videos in Head-Mounted Virtual Reality. In Proc. of QoMEX'19.
[2] S. Croci et al. Voronoi-Based Objective Quality Metrics for Omnidirectional Video. In Proc. of QoMEX'19.
[3] S. Croci et al.. Visual attention-aware quality estimation framework for omnidirectional video using spherical Voronoi diagram. Springer Quality and User Experience 5, 1 (2020).

# IS Modeling

- IS modeling leads to slightly inferior results compared to MOS modeling
  - Heterogeneous characteristics and behaviors among different subjects

**PLCC , SROCC > 0.70**

| Model | OQ | IQ | FG | IM | CS |
|---|---|---|---|---|---|
| PLCC | 0.915 | 0.896 | 0.883 | 0.801 | **0.579** |
| SROCC | 0.868 | 0.847 | 0.868 | 0.725 | **0.594** |

**CS needs more human factors**

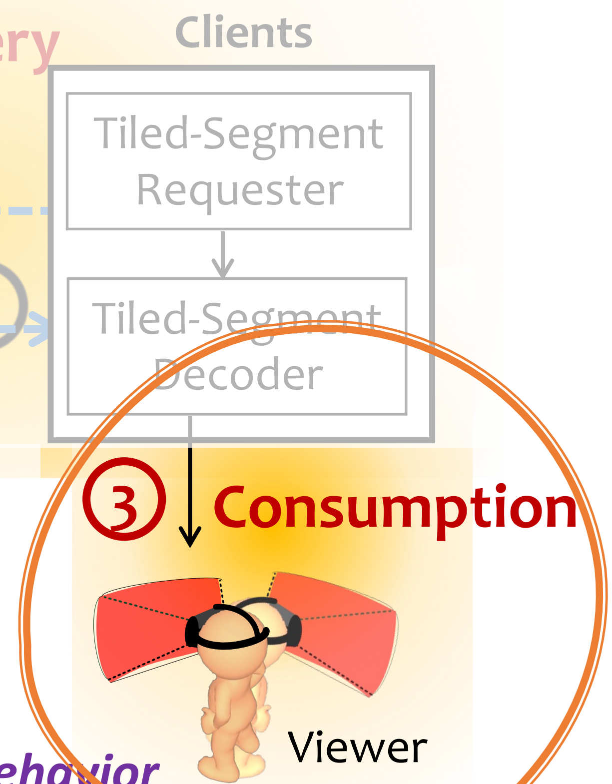
- Observations are similar to MOS modeling
  - Content dominates the factor category except for FG
  - achieve > 97% performance ratio for the overall QoE and most QoE features
  - IM cannot be well modeled by a single factor

# Tiled 360° Video Streaming Platform

**QoE Modeling**

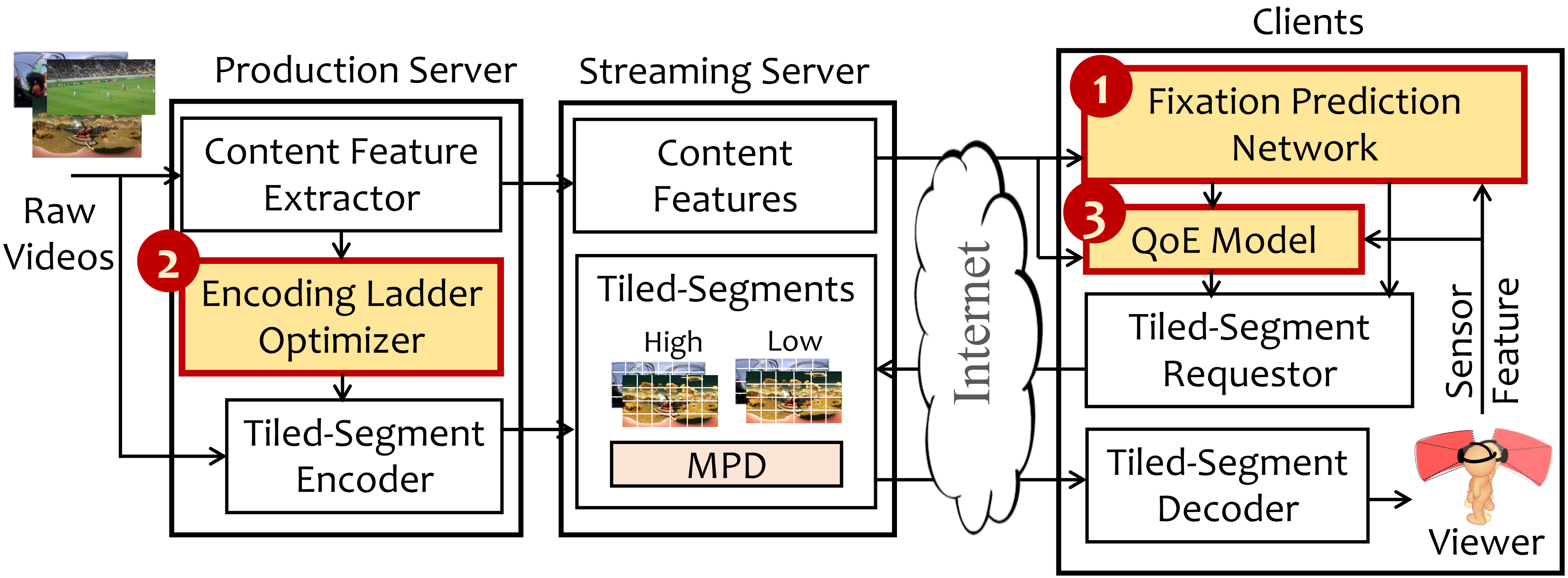
- Estimate the perceived QoE by the viewer
- We derived models for both **MOS** and **IS**
- We identify the **dominating factor categories** and **factors**
- ***Several observations are made for future improvements***

Raw Videos



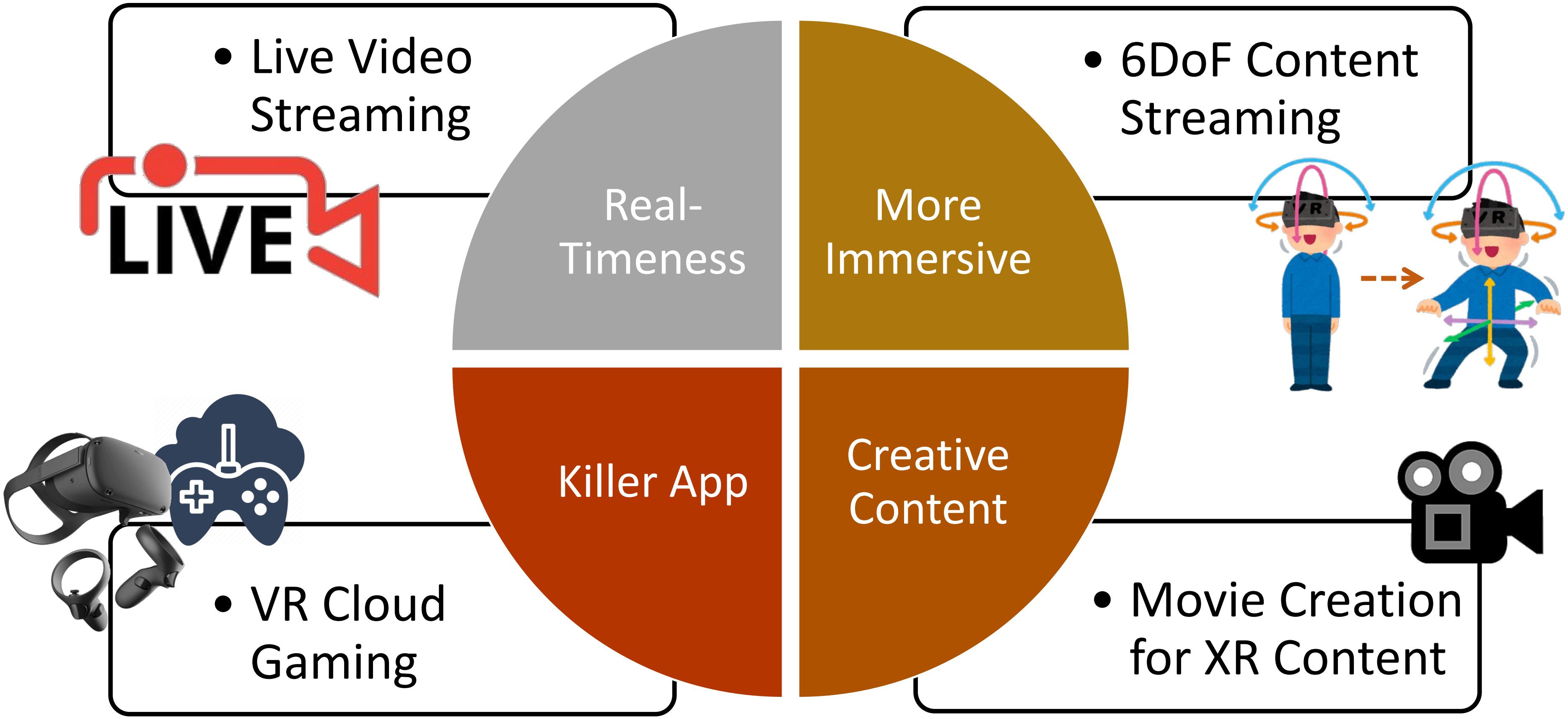

- ***Diverse behavior***
- ***Complex and unknown QoE***

# Optimized 360° Video Streaming Platform




***QoE-Driven Optimized 360° Video Streaming Platform***

# Future Research Directions

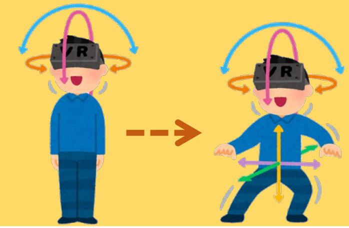


# Real-Timeness: Live Video Streaming

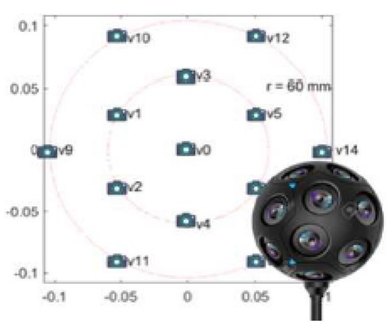
- Applying our proposed solution
  - Optimal laddering: per-class optimization
- Challenges: dependence of *content features*
- Possible solutions:
  - Speed up content feature generation, e.g., real-time saliency detection [1]
  - Eliminating the dependence of content features, e.g., video prediction network [2]

[1] H. Zhou, X. Xie, J. Lai, Z. Chen, and L. Yang. Interactive two-stream decoder for accurate and fast saliency detection. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'20), June 2020.
[2] O. Shouno. Photo-realistic video prediction on natural videos of largely changing frames. arXiv preprint arXiv:2003.08635, 2020.

# More Immersive: 6DoF Content Streaming



- Challenges
  - Even larger data size
  - More complex computation
  - Unknown QoE




RGB-D

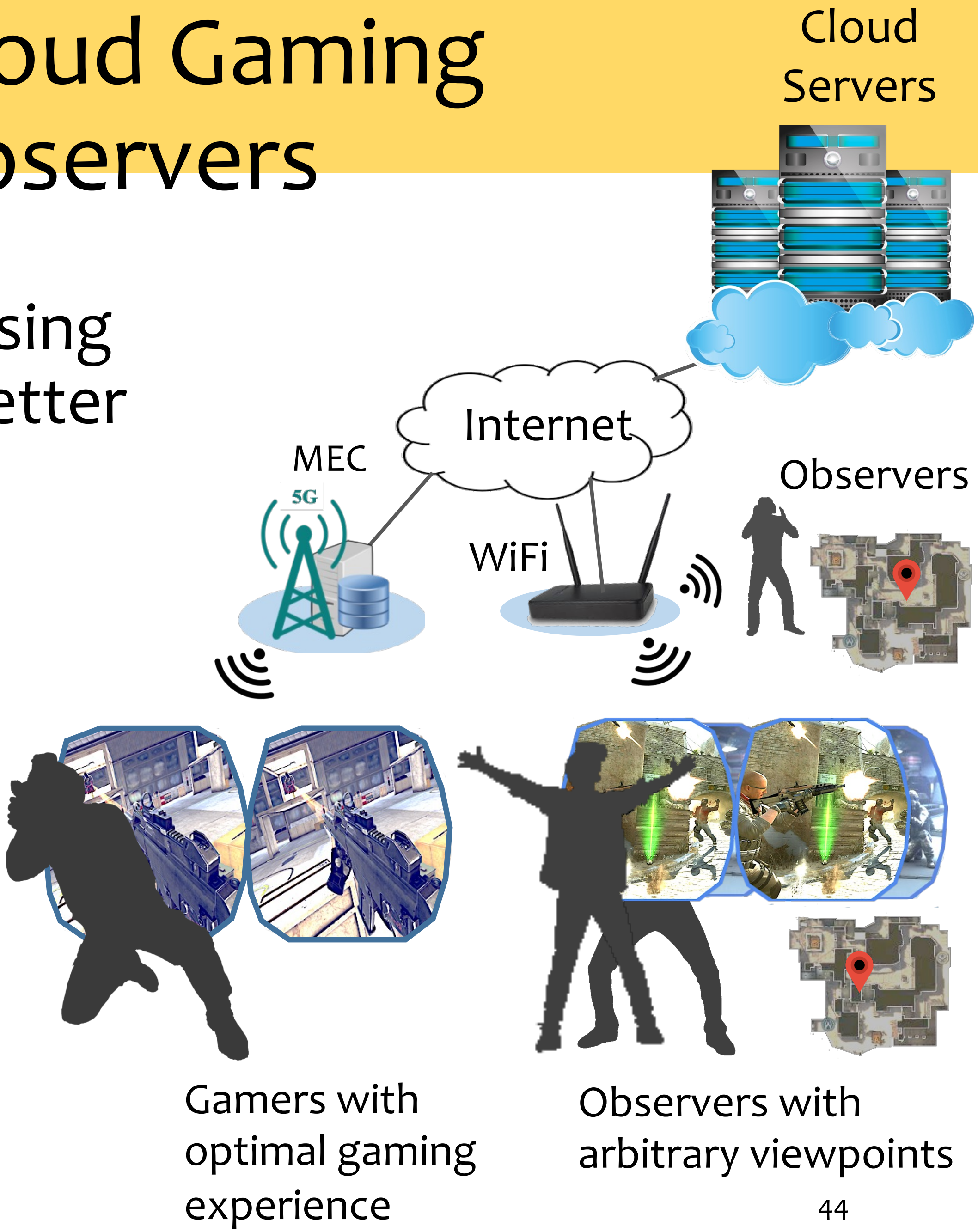
Light-Field

Mesh

Point Cloud

# Killer App: VR Cloud Gaming with Multiple Observers

- Viewport *prediction* using *in-game context* for better bitrate allocation
- *QoE*-optimized 6*DoF streaming*
- *Cross-layer optimized* for global *resource allocation*



# Creative Content: Movie Creation for XR Content

- Challenges
  - the richness of the story are difficult to express
  - any scene transitions can ruin the audience's immersion
  - the comfort needs to be improved

- Possible solutions:
  - factors investigation for gaze attraction and sickness elimination, e.g., motion, glance, and transition effects

  ⇒ scene presentation and transition recommendation


# Thank You

Ching-Ling Fan (ch.ling.fan@gmail.com)

# Backup Slides

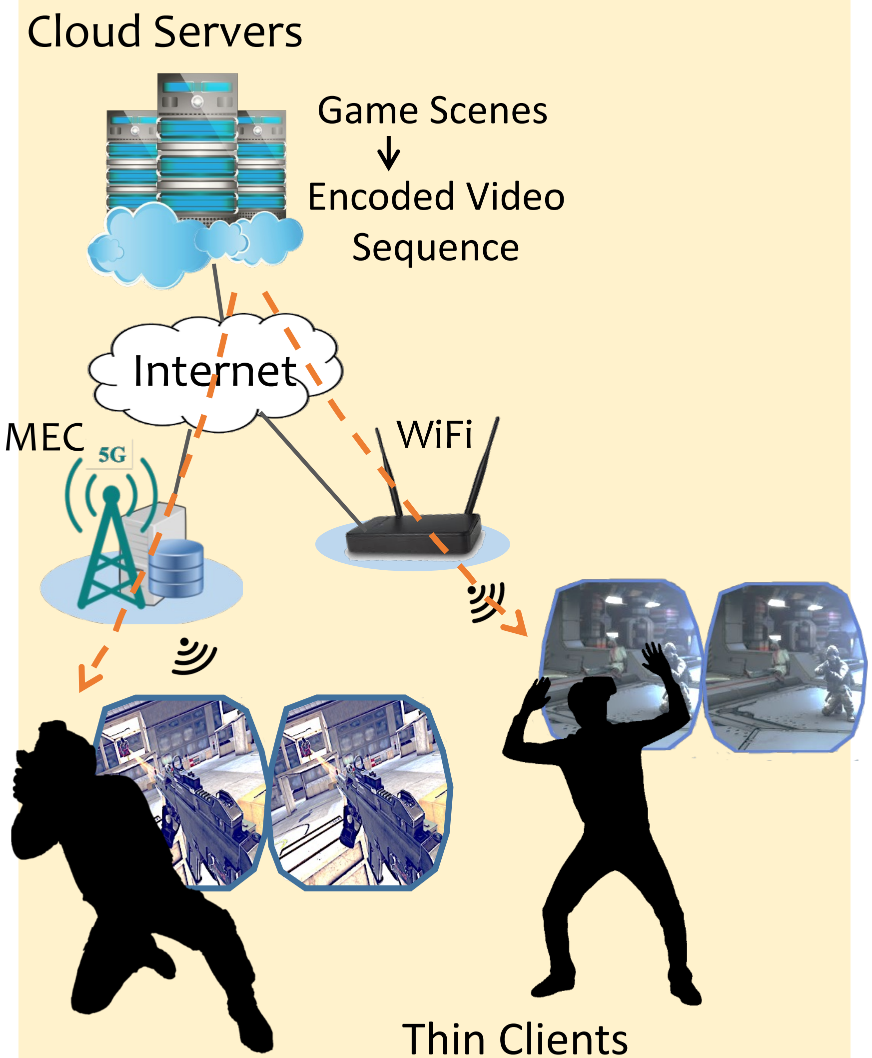
# State-of-the-Art Prediction Algorithms

| Literature | Approach | Classification | Considered Features | Output |
|---|---|---|---|---|
| Fan et al. [55, 57] | LSTM | No | Historical sensor data, saliency maps, and motion maps of frames | Future tile viewing probabilities |
| Nguyen et al. [142] | LSTM | No | Saliency maps and historical orientation maps of frames | Future saliency maps |
| Bai et al. [13] | Neural Network | No | Historical orientation | Future orientation |
| Xu et al. [221] | LSTM | No | Historical orientation | Future orientation |
| Qian et al. [167] | Regressor | No | Historical orientation | Future orientation |
| Xu et al. [223] | Regressor | No | Historical orientation | Future orientation |
| Zhang et al. [230] | Spherical CNN | No | Spherical video frames | Future saliency maps |
| Xu et al. [222] | CNN+LSTM | No | Historical viewer fixation trajectories, video frames | Future gaze trajectory |
| Hou et al. [77] | LSTM | No | Historical orientation | Future orientation |
| Hou et al. [75] | LSTM | No | Historical viewed tiles | Future viewed tiles |
| Wu et al. [214] | Spherical CNN | No | Video frames, viewport, and motion | Future viewport |

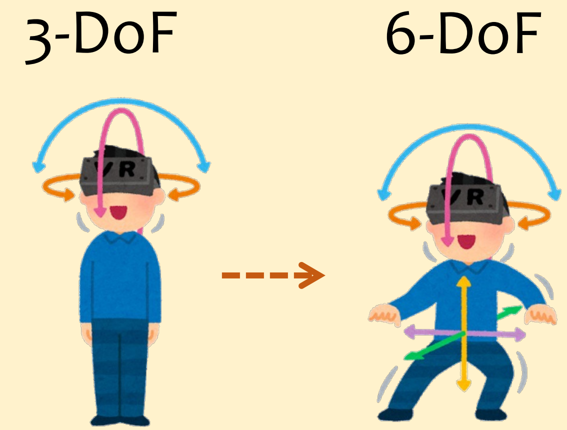
# State-of-the-Art Prediction Algorithms

| Literature | Approach | Classification | Considered Features | Output |
|---|---|---|---|---|
| Chen et al. [30] | CNN+LSTM | No | Video frames and historical orientation | Future orientation |
| Feng et al. [59] | CNN+LSTM | No | Video segment and historical orientation | Future orientation |
| Vielhaben et al. [203] | Regressor | No | Historical orientation | Future orientation |
| Cheng et al. [31] | CNN+Convolutional LSTM | No | Faces of cubic frames | Future saliency maps |
| Xu et al. [220] | Reinforcement Learning | No | Historical viewer orientation and video frames | Future head-moving directions |
| Feng et al. [60] | Bayes prediction | Clustered by video content and viewer behavior | Viewer orientation and video frames | Future tile viewing probabilities |
| Nasrabadi et al. [137] | Extrapolation | Clustered by viewer behavior | Historical and other's orientation | Future orientation |
| Ban et al. [12] | KNN | Per video | Historical and other's orientation | Future tile viewing probabilities |
| Xie et al. [217] | SVM | Per video | Historical orientation | Viewing behavior class |

# Cloud VR Gaming



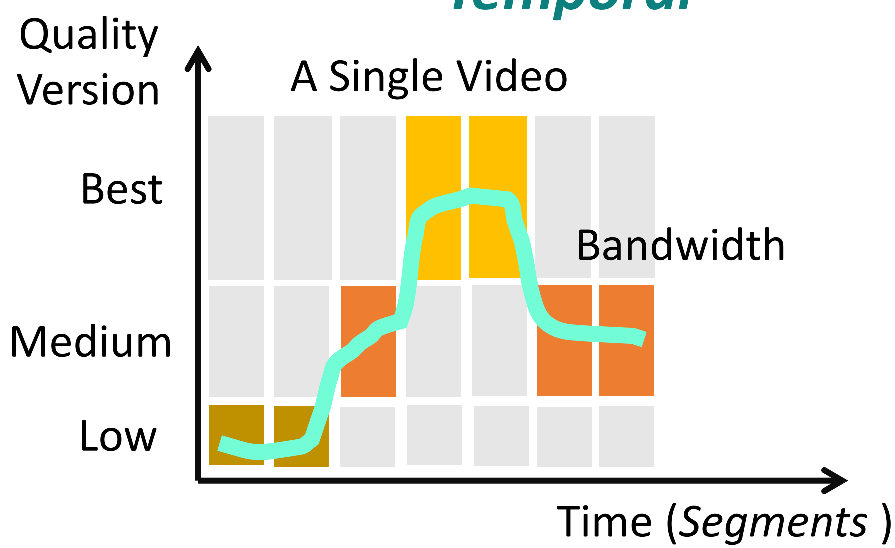

# 6-DoF Streaming

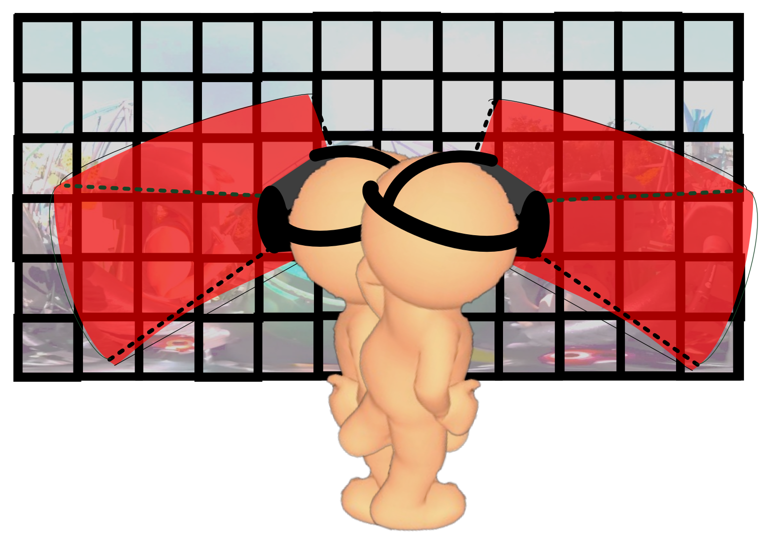



# Viewport-Adaptive Streaming

- *Tiling* *with MPEG* *DASH* *(Dynamic Adaptive Streaming over HTTP)*

***Temporal*** ***Spatial***

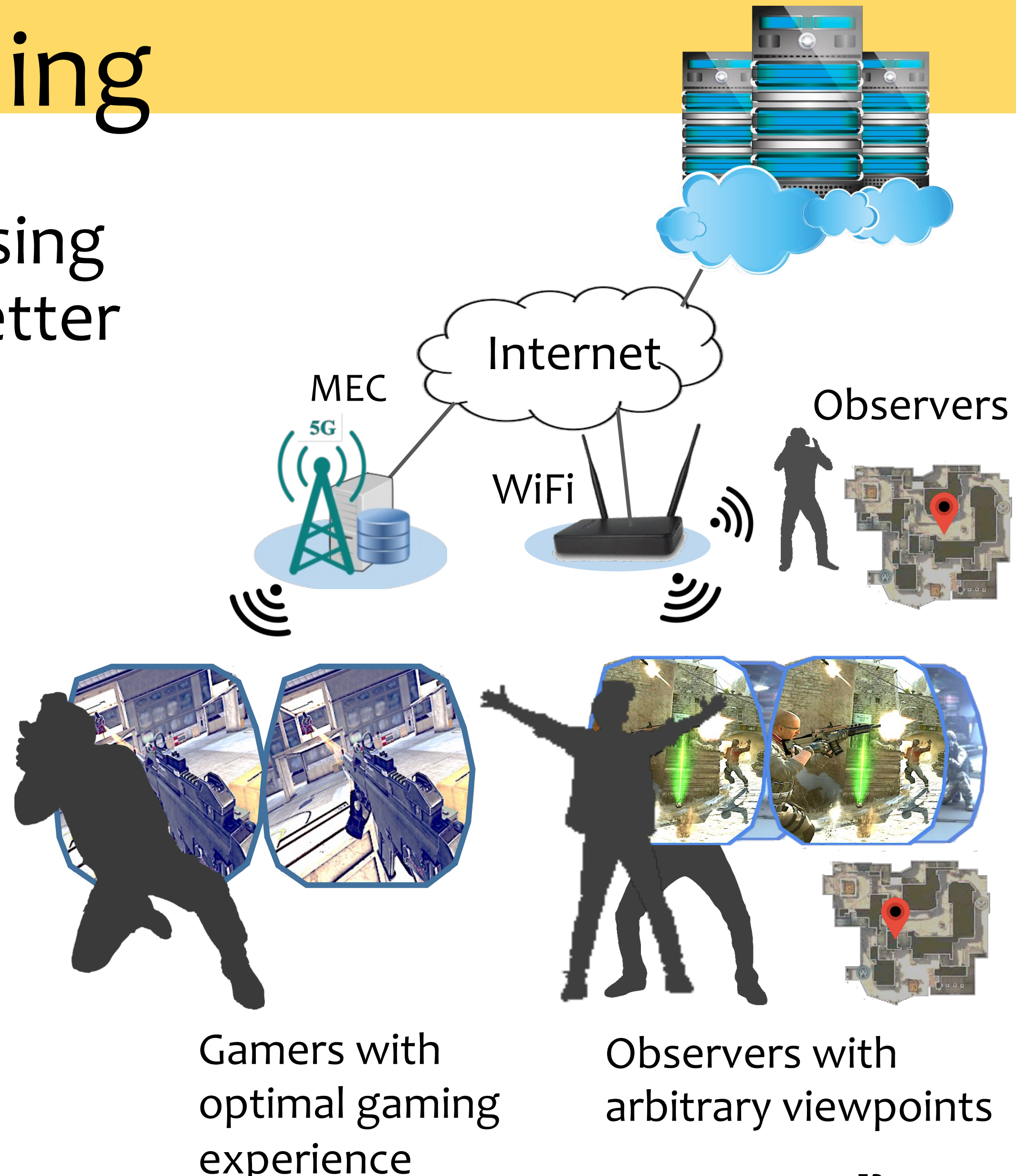


- Basic transmission unit: **Tiled-segments**

# Sample Application: Cloud VR Gaming

- Viewport *prediction* using *in-game context* for better bitrate allocation
- *QoE*-optimized 6DoF *streaming*
- *Cross-layer optimized* for global *resource allocation*



# A Small-Scale User Study

- Play the viewport videos to 7 subjects and collect the MOS scores (1-5)
- Our fixation prediction network achieves similar MOS scores while saves 41% bandwidth on average

Missing Ratio < 10%

| Trace | MOS | | | Bandwidth (Mbps) | | |
|---|---|---|---|---|---|---|
| | Cur | DR | Our | Cur | DR | Our |
| *Roller Coaster* | 3.14 | 2.86 | 2.86 | 24.35 | 24.33 | 15.32 |
| *Hog Rider* | 3.43 | 3.43 | 3.43 | 24.18 | 24.21 | 13.32 |
| *SFR Sport* | 3.14 | 3.00 | 3.29 | 24.19 | 24.25 | 13.71 |
| *Average* | *3.24* | *3.10* | *3.20* | *24.24* | *24.26* | *14.12* |

**-0.04 ~ 0.1 MOS score** **-41% bandwidth**

# Lagrangian-based: PC-LBA

- Both distortion and bitrate models are convex

$$d_{v,t,n}(q) = \alpha^d_{v,t,n} q^{\beta^d_{v,t,n}} + \gamma^d_{v,t,n}$$

$$r_{v,t,n}(q) = \alpha^r_{v,t,n} e^{\beta^r_{v,t,n}}$$

$$P'(v,t,c) = \min \sum_{n=1}^{N} d_{v,t,n}(\kappa_{v,t,n,c}) p_{v,t,n} a_n$$

$$st: \sum_{n=1}^{N} r_{v,t,n}(\kappa_{v,t,n,c}) \le b_c;$$

$$\Downarrow \qquad \underline{\kappa_{v,t,n,c} \in [\kappa_{min}, \kappa_{max}].}$$

- Transform the discrete decision variables $x_{v,t,n,c,q}$ into continuous decision variables $\kappa_{v,t,n,c}$ (QP)

$$\underline{\min\ L(\mathbf{K_{v,t,c}}, \mu) = \sum_{n=1}^{N} d_{v,t,n}(\kappa_{v,t,n,c}) p_{v,t,n} a_n + \mu\left(\sum_{n=1}^{N} r_{v,t,n}(\kappa_{v,t,n,c}) - b_c\right)}$$

Unconstrained problem

$$\longrightarrow \quad g(\mu) = \inf_{\mathbf{K_{v,t,c}}} (\mathbf{K_{v,t,c}}, \mu) = \inf_{\mathbf{K_{v,t,c}}} \left(\sum_{n=1}^{N} d_{v,t,n}(\kappa_{v,t,n,c}) p_{v,t,n} a_n + \mu\left(\sum_{n=1}^{N} r_{v,t,n}(\kappa_{v,t,n,c}) - b_c\right)\right)$$

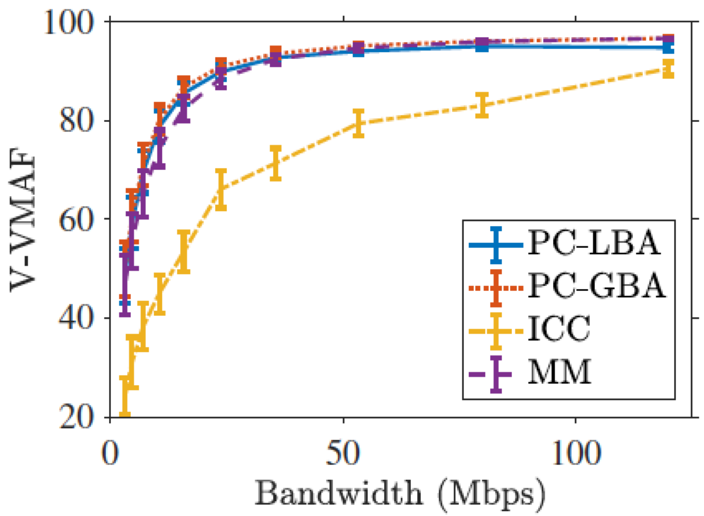
$$\longrightarrow \quad \frac{\partial L}{\partial \kappa_{v,t,n,c}} = \left(\alpha^d_{v,t,n} \beta^d_{v,t,n} \kappa_{v,t,n,c}^{\beta^d_{v,t,n}-1}\right) p_{v,t,n} a_n + \mu \alpha^r_{v,t,n} \beta^r_{v,t,n} \mathrm{e}^{\beta^r_{v,t,n} \kappa_{v,t,n,c}} = 0$$

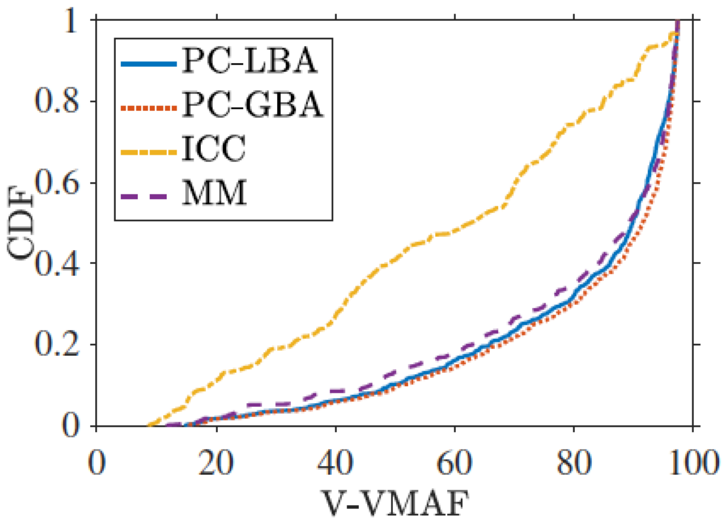
**QP**

$$\longrightarrow \quad \kappa_{v,t,n,c} = \frac{1-\beta^d_{v,t,n}}{\beta^r_{v,t,n}} W\left(\frac{\beta^r_{v,t,n}}{1-\beta^d_{v,t,n}} \mathrm{e}^{\frac{-\ln \frac{\mu \alpha^r_{v,t,n} \beta^r_{v,t,n}}{-\alpha^d_{v,t,n} \beta^d_{v,t,n} p_{v,t,n} a_n}}{1-\beta^d_{v,t,n}}}\right)$$

# Sample Results: Per-Class Optimization

- 10 bandwidth classes: 3.12 -- 119.87 Mbps
- (10 users , 6 videos) in each bandwidth classes

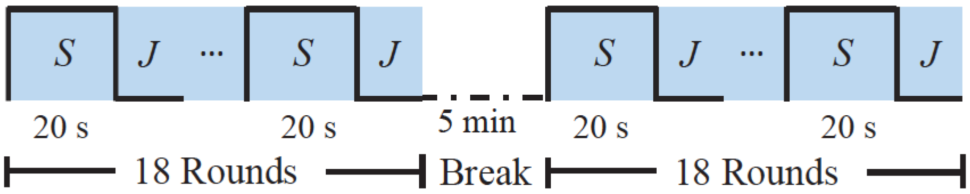



**Our solution outperforms others by up to 52.17 and 26.35 in V-VMAF**

# Procedure

- ITU-T 910, Absolute Category Rating (ACR)
  - Random order
  - 36 rounds
  - Scores: [1,9]

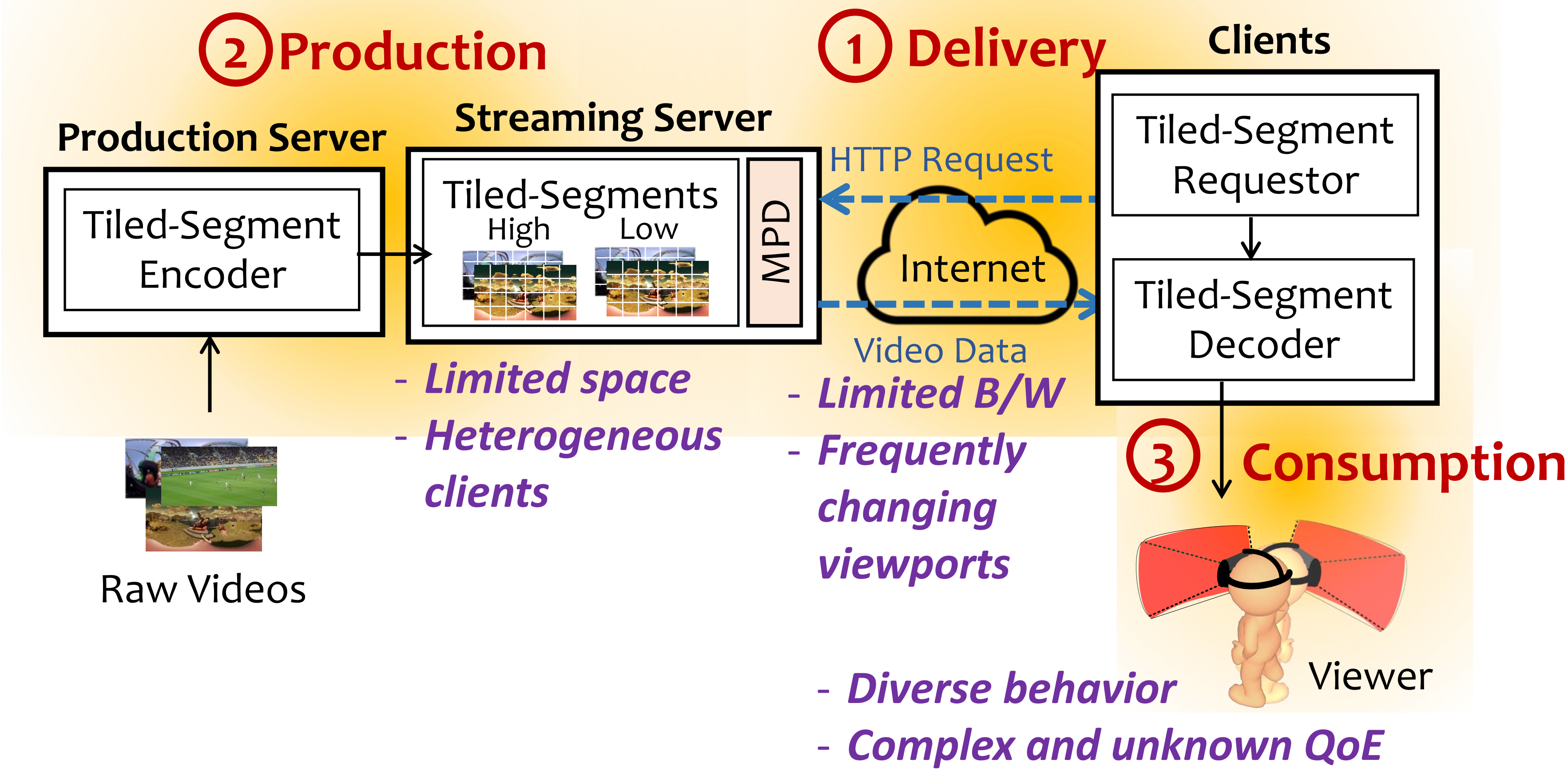



- Questionnaire

| Feature | Question | Lowest Score (1) | Highest Score (9) |
|---|---|---|---|
| - | How would you rate the overall quality? | Bad | Excellent |
| IQ | How would you rate the image quality? | Bad | Excellent |
| FG | How would you rate the fragmentation level? | None | Severe |
| IM | How would you rate the immersion level? | Bad | Excellent |
| CS | How would you rate the perceived cybersickness level? | None | Severe |
| AT | How would you rate the attractiveness level? | Not Attractive | Attractive |

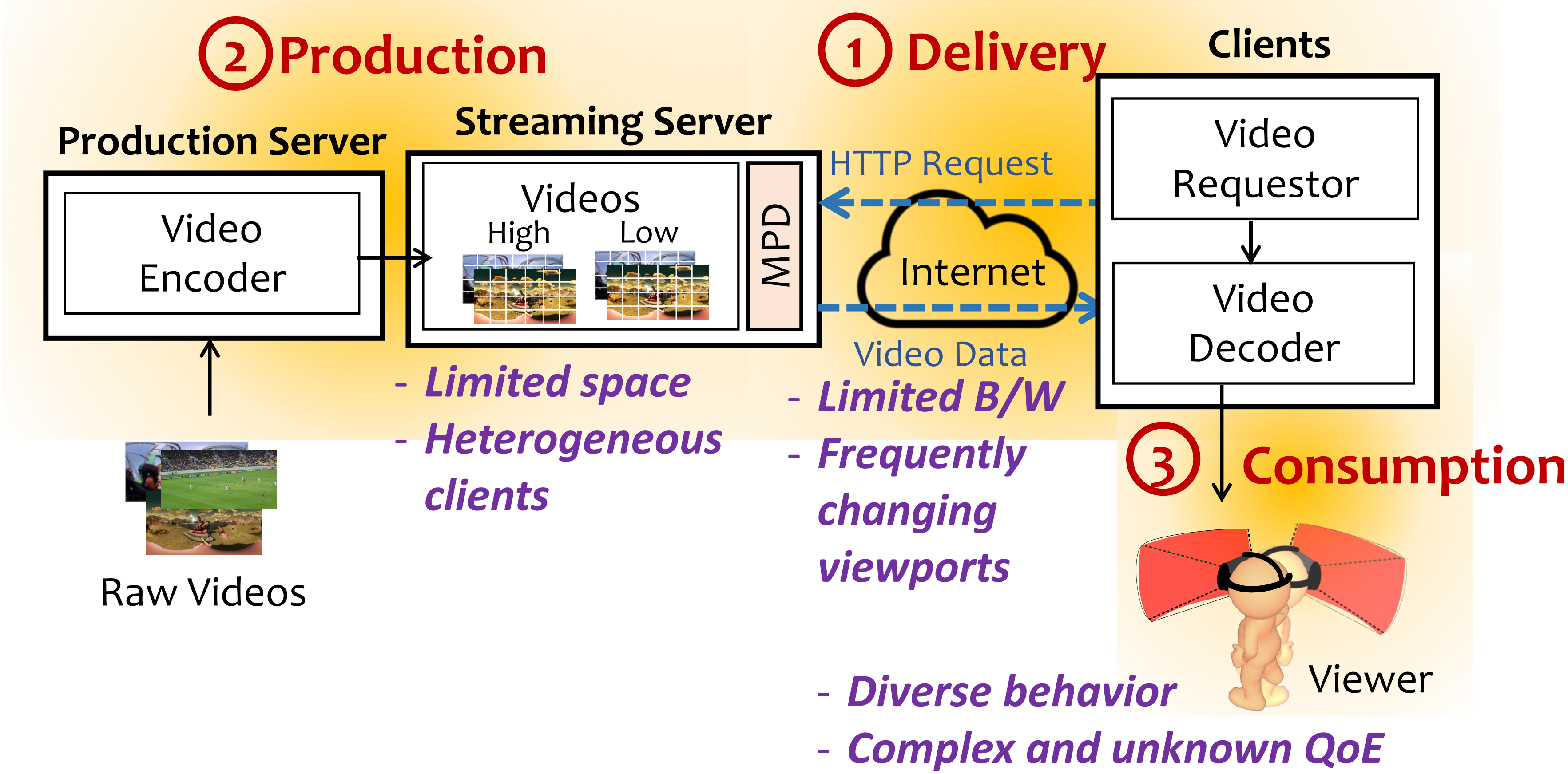
# Tiled 360° Video Streaming Platform

- Three crucial phases in tiled 360° video streaming



# 360° Video Streaming Platform

- Three crucial phases in 360° video streaming



# Tiled 360° Video Streaming Platform

- Three crucial phases in tiled 360° video streaming