

AHG 16: Scalable Representation and Layered Reconstruction for Generative Face Video Compression

Bolin Chen, Yan Ye, Jie Chen, Ru-Ling Liao, Shanzhi Yin and Shiqi Wang

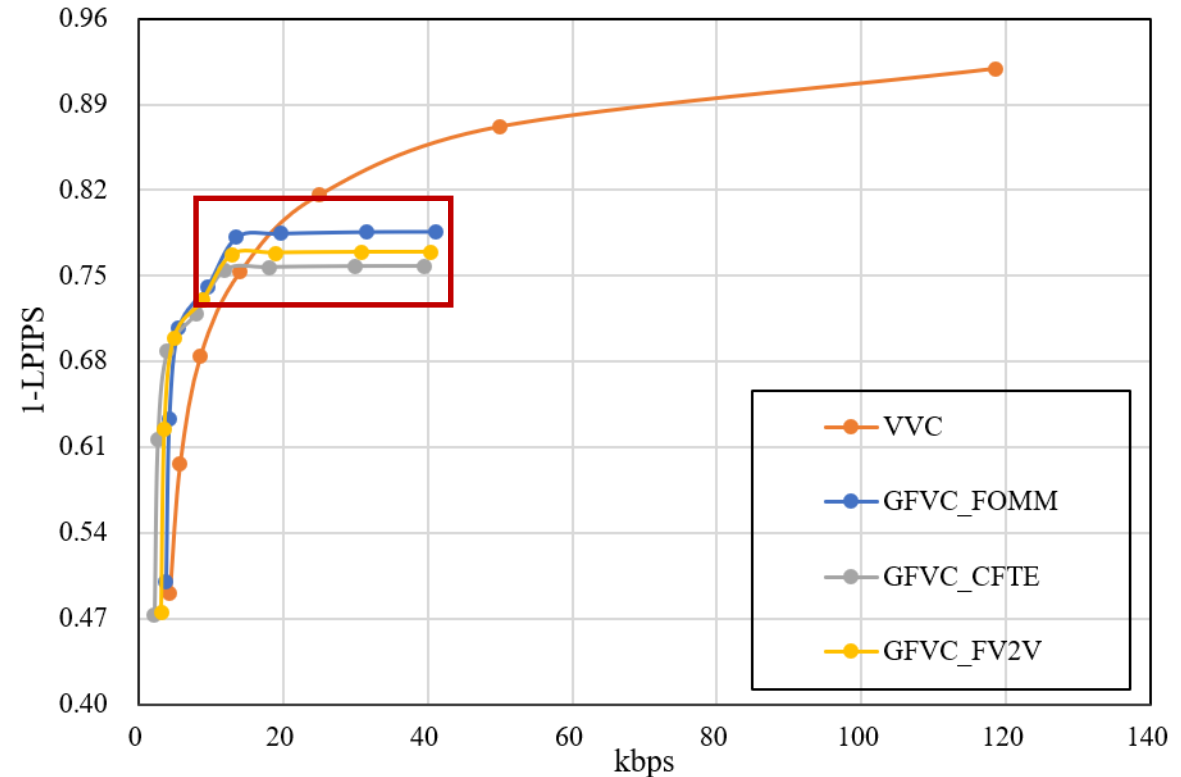
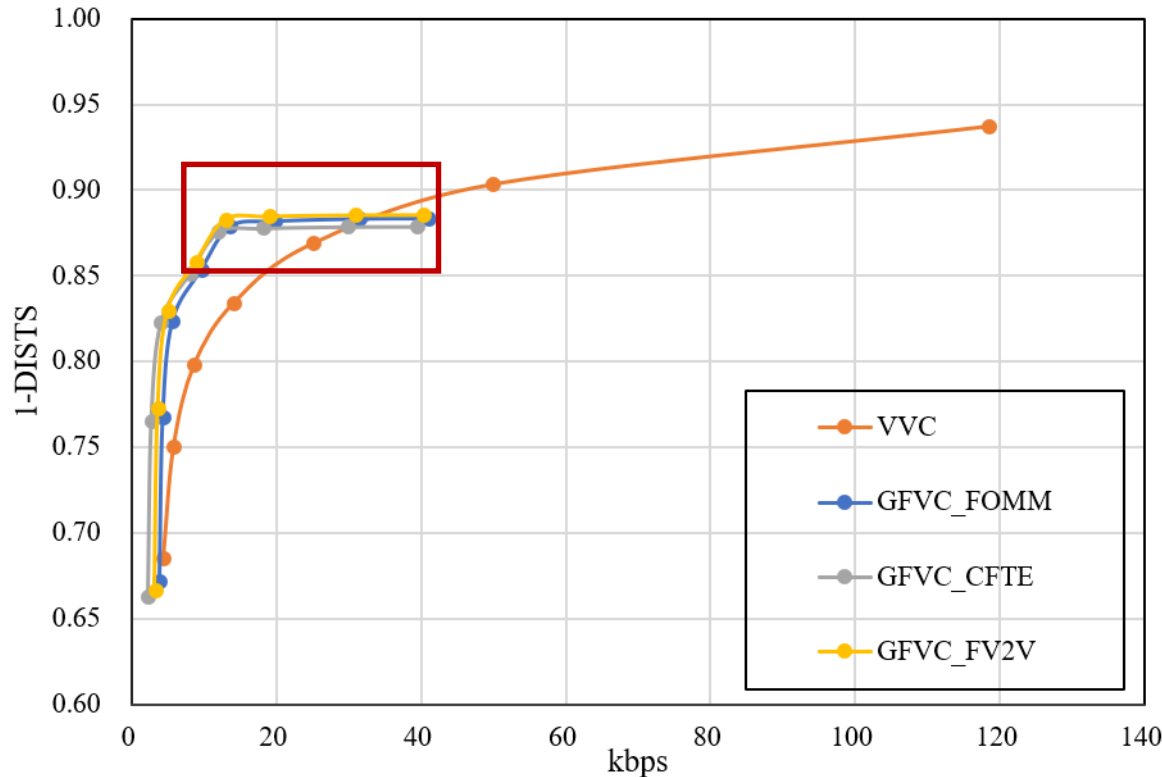
Alibaba Group and City University of Hong Kong

April. 2024

Proposal Review & Contribution

- ❑ **Review:** At the 33th JVET meeting, **a unified software package as well as test conditions** for generative face video coding (GFVC) were established for **JVET AHG16 experimentation**.
- ❑ **Problem:** It has been observed that the current GFVC software can only achieve promising face video compression in ultra-low bitrate ranges, and **cannot provide a more comprehensive coverage for mid-to-high bitrate ranges**.
- ❑ **Contribution:** we propose a Scalable Representation and Layered Reconstruction (SRLR) architecture to **extend the bitrate and quality range** that GFVC methods in AHG16 software can cover. Experimental results show that the proposed SRLR architecture can extend the quality and bitrate ranges for all three GFVC methods supported in the AHG16 software for all quality metrics in the GFVC test conditions.

Problem Statements- Limited Bitrate Coverage



- ❑ VVC: VTM 22.0 LDB QPs=22/27/32/37/42/47/52 || GFVC: base picture QPs =2/7/12/17/22/32/42/52
- ❑ Allocating more coding bits to base pictures can only improve perceptual quality within **a limited bitrate range** for generative face video coding

Problem Statements- Rec Artifacts

Original Video



Generated Occlusions



Low Face Fidelity

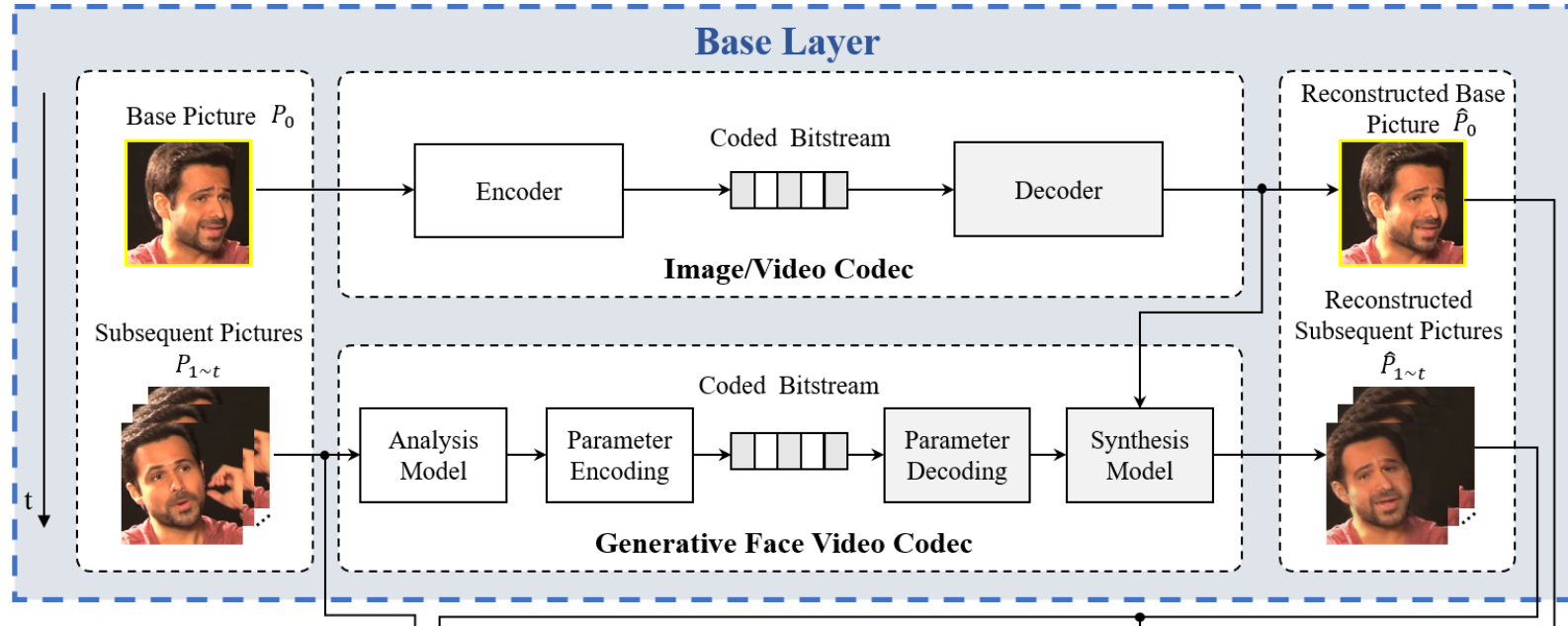


Poor Local Motion



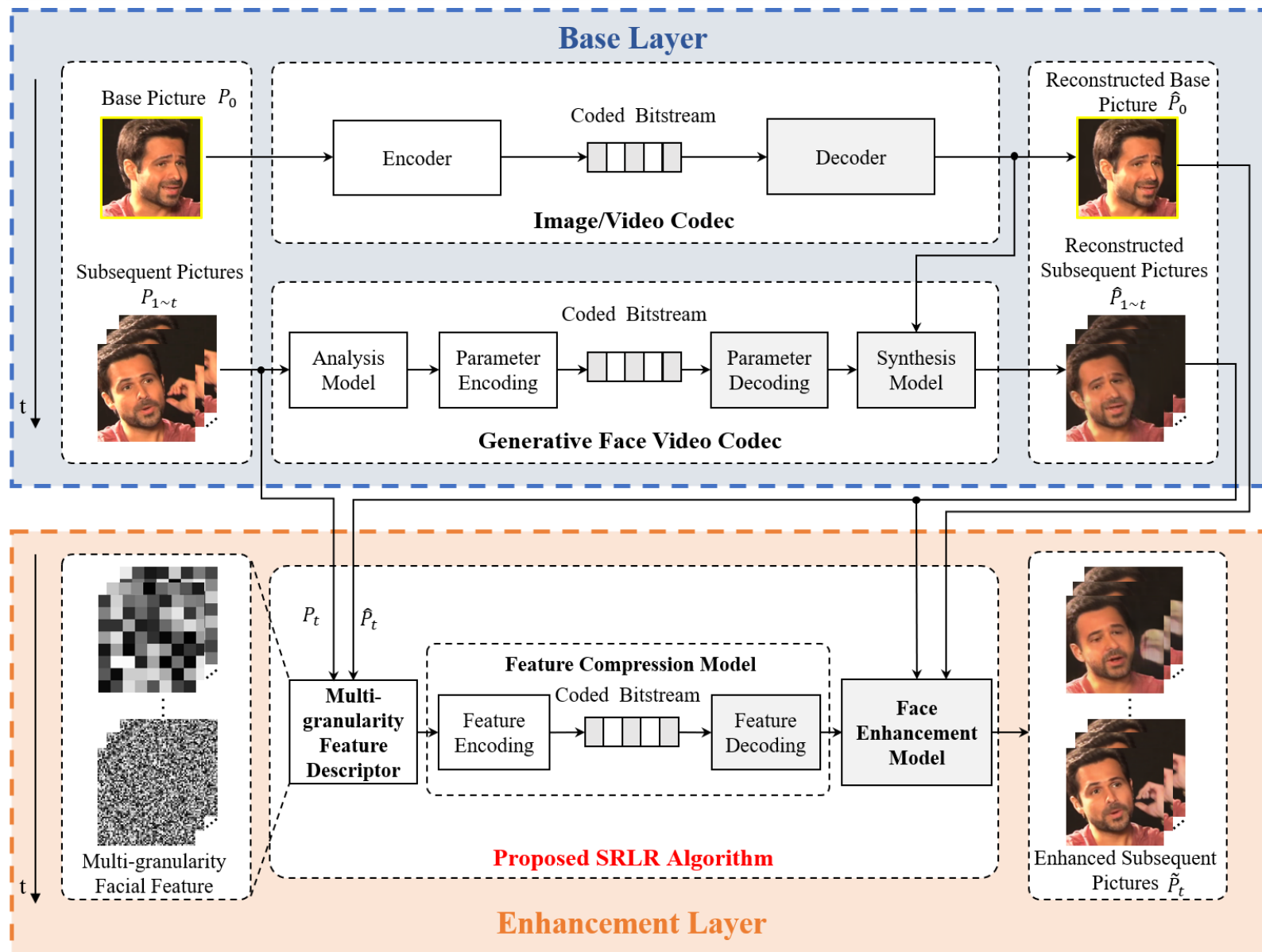
- ❑ The GFVC robustness in **complex motion and long-term dependency** scenes cannot always be guaranteed.
- ❑ GFVC is prone to failure cases with annoying artifacts, missing details and temporal inconsistency.

Algorithm Overview



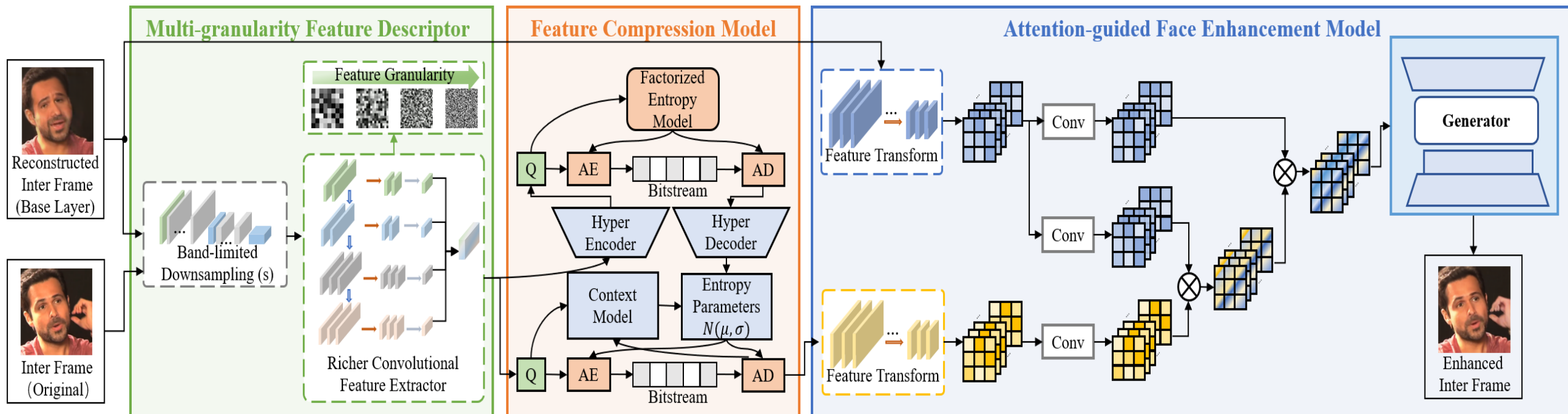
- The base layer can be one of any existing **GFVC algorithms in AHG16 software**, which can achieve ultra-low bitrate face video communication with compact representations like 2d keypoints, 3d keypoints and etc.

Algorithm Overview



- ❑ The base layer can be one of any existing **GFVC algorithms in AHG16 software**, which can achieve ultra-low bitrate face video communication with compact representations like 2d keypoints, 3d keypoints and etc.
- ❑ The enhancement layer is our proposed SRLR-based algorithm that is capable of providing **different-granularity feature space** with enriched facial motion information and support high-quality face video communication when bandwidth permitting.

Algorithm Details



- ❑ **Multi-granularity feature descriptor:** can hierarchically compensate for motion estimation errors by delivering multi-granularity feature representation with the size of 8*8, 16*16, 24*24 and 32*32.
- ❑ **Learning-based feature compression model:** can improve the compression efficiency of multi-granularity feature residuals via the joint autoregressive and hierarchical priors-based entropy model.
- ❑ **Attention-guided face enhancement model:** can reconstruct more reliable and robust face video.

Proposed Testing Conditions

□ Herein, we follow the test conditions and software reference configurations for GFVC testing specified in **JVET-AG2035** for **low bitrate range**. It should be noted that since the proposed SRLR-based algorithm is designed for **high bitrate coverage**, we propose to modify test settings as follows,

- **Low bitrate ranges (JVET-AG2035):**

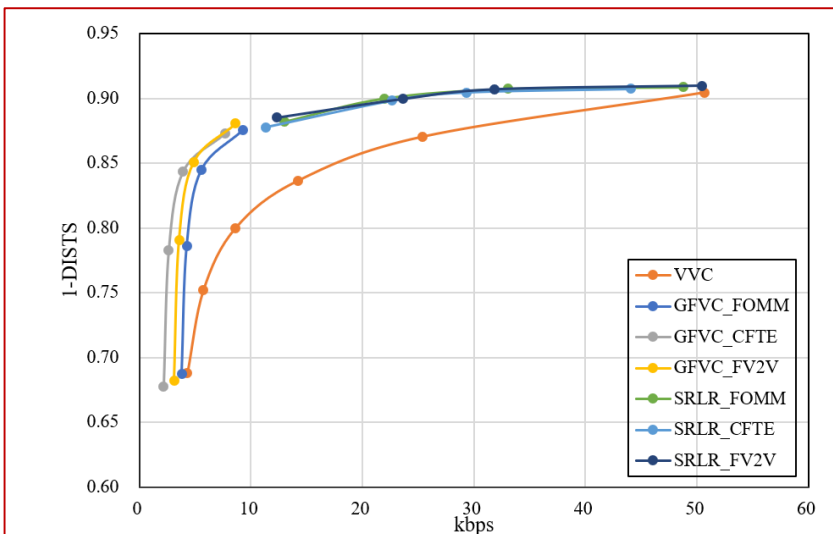
- ✓ VVC: VTM 22.2 LDB , QPs=37/42/47/52
- ✓ GFVC: Base picture QPs= 22/32/42/52

- **High bitrate ranges (Newly-proposed):**

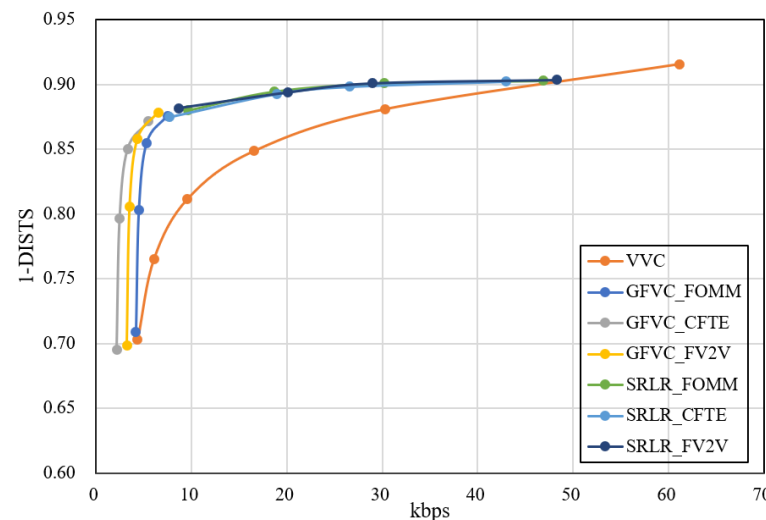
- ✓ VVC: VTM 22.2 LDB , QPs= **27/32/37/42**.
- ✓ SRLR-based algorithm: the base picture QP of base-layer GFVC algorithm is **17**. Enhancement layer includes 4 different feature granularities with the feature size of 8*8, 16*16, 24*24 and 32*32.

RD Performance Results (Perceptual-level)

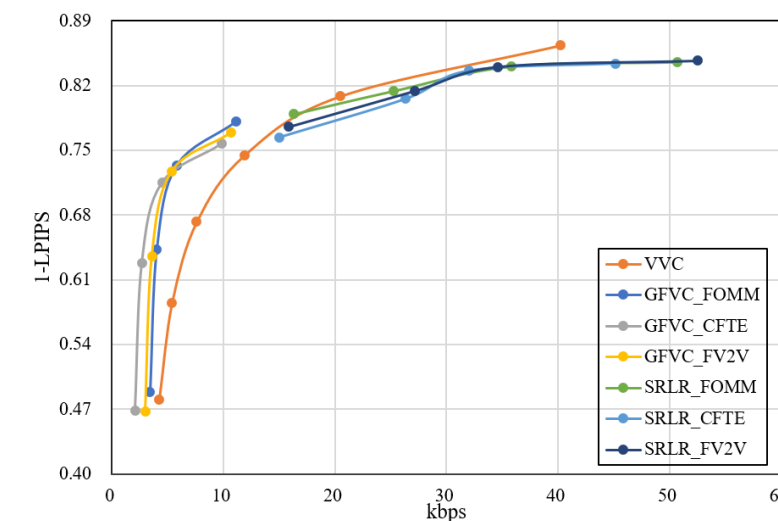
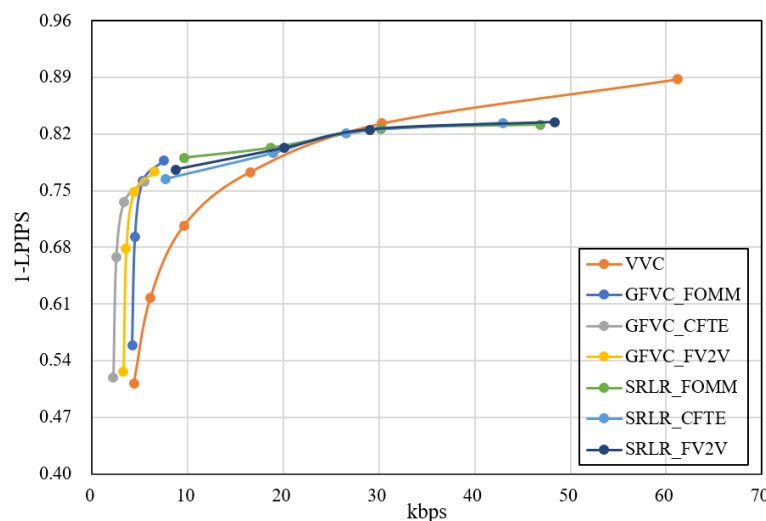
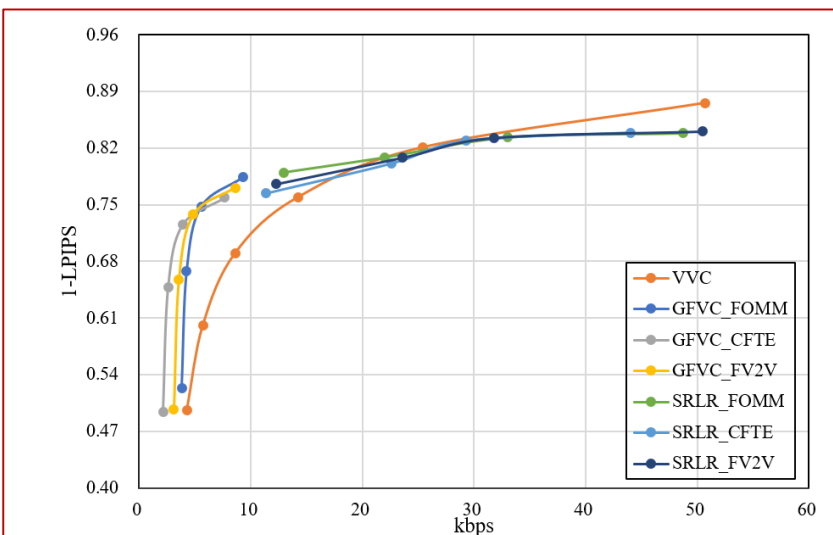
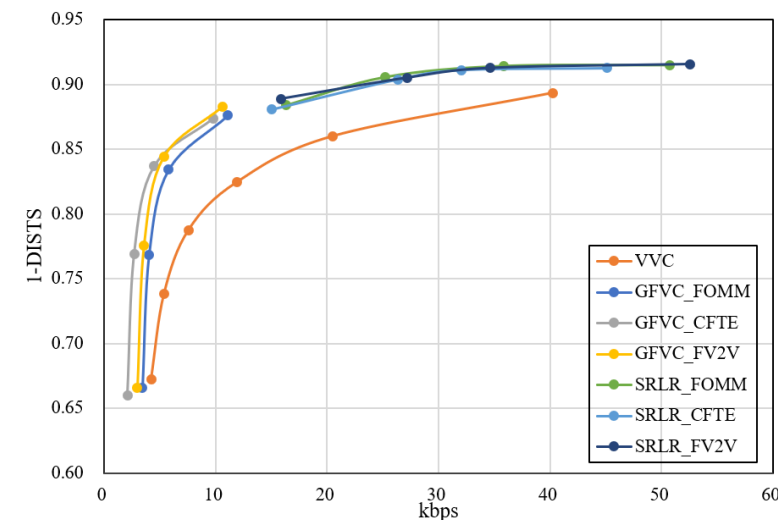
Overall



Class A (VoxCeleb)

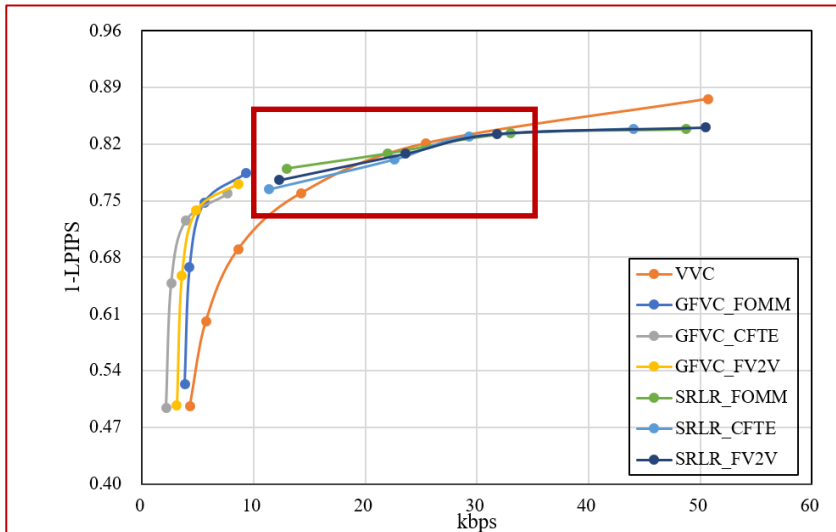
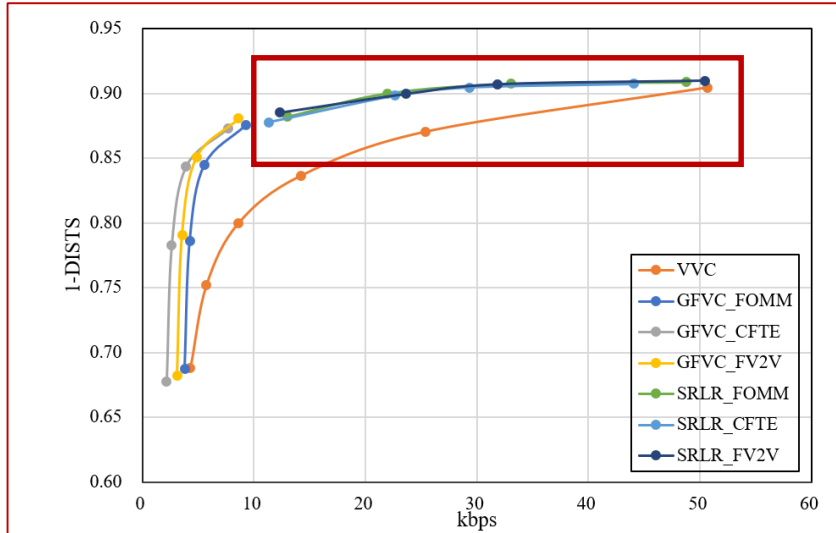


Class B (CFVQA)



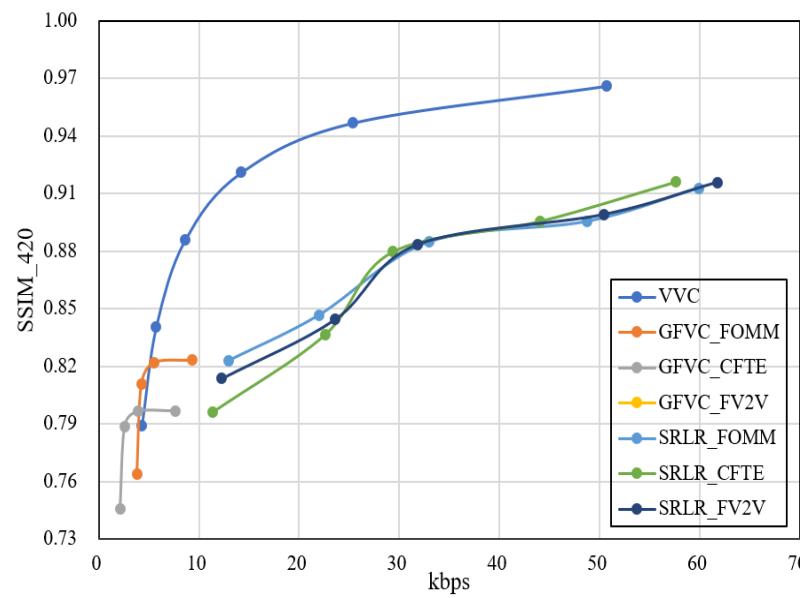
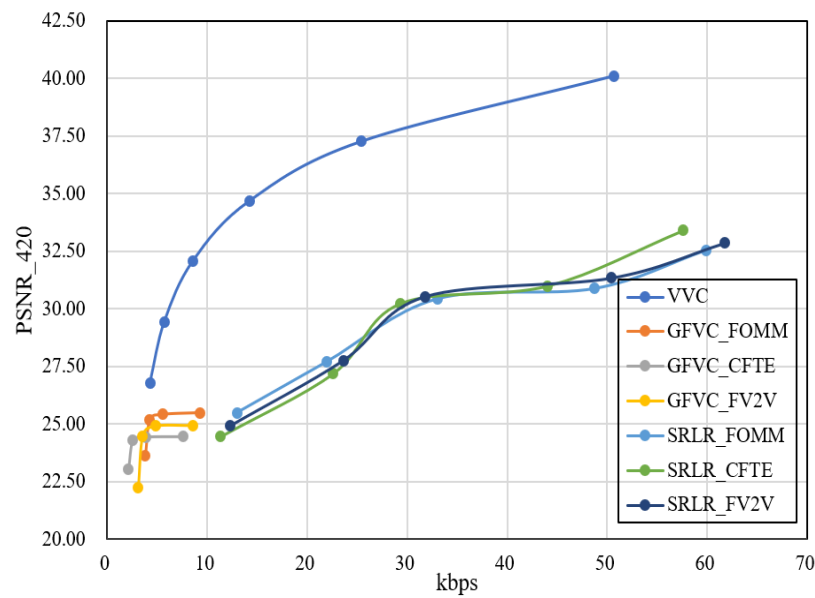
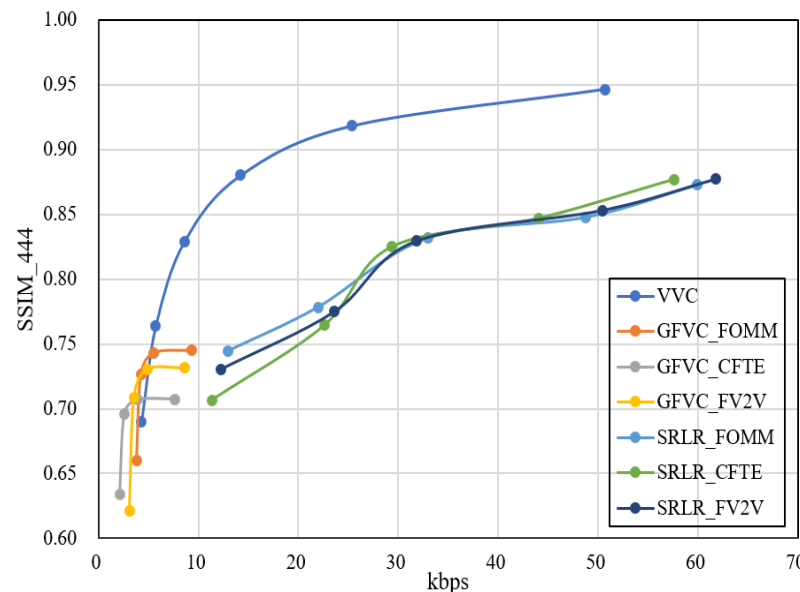
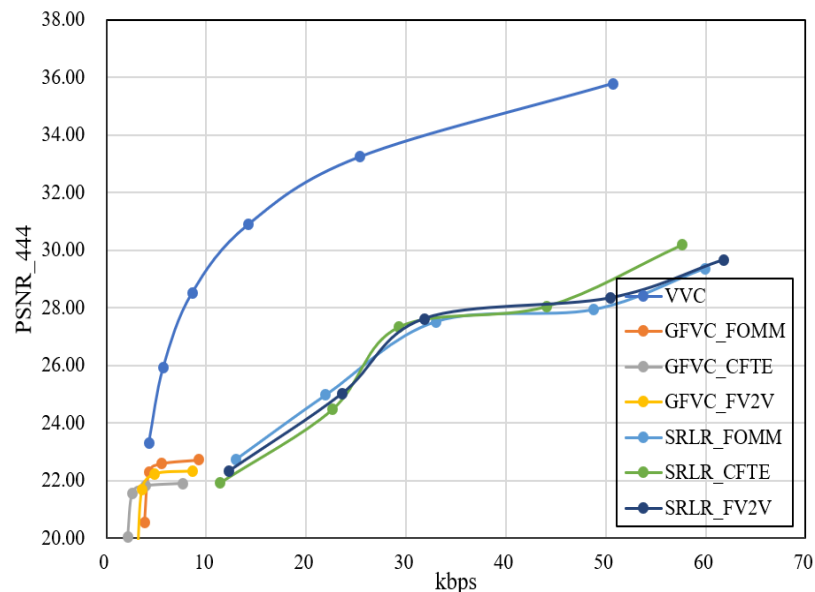
RD Performance Results (Perceptual-level)

Overall



- ❑ In terms of DISTS, the proposed method clearly have better rate distortion performance compared with VVC at relatively **high bitrate ranges** (10~50 kbps).
- ❑ As for the LPIPS, the proposed algorithm also can achieve performance gains at **middle bitrate ranges** (10~30 kbps) compared with VVC.
- ❑ It may be attributed that these testing models are **optimized by DISTS loss function** rather than LPIPS loss function, thereby the testing results are more inclined towards DISTS measurement.

RD Performance Results (Pixel-level)



- The proposed method can **improve the fidelity of signal reconstruction**
- Although the overall RD performance **is still not as good as VVC**, this is expected as all generative methods do not optimize for sample-level fidelity

BD-rate Savings

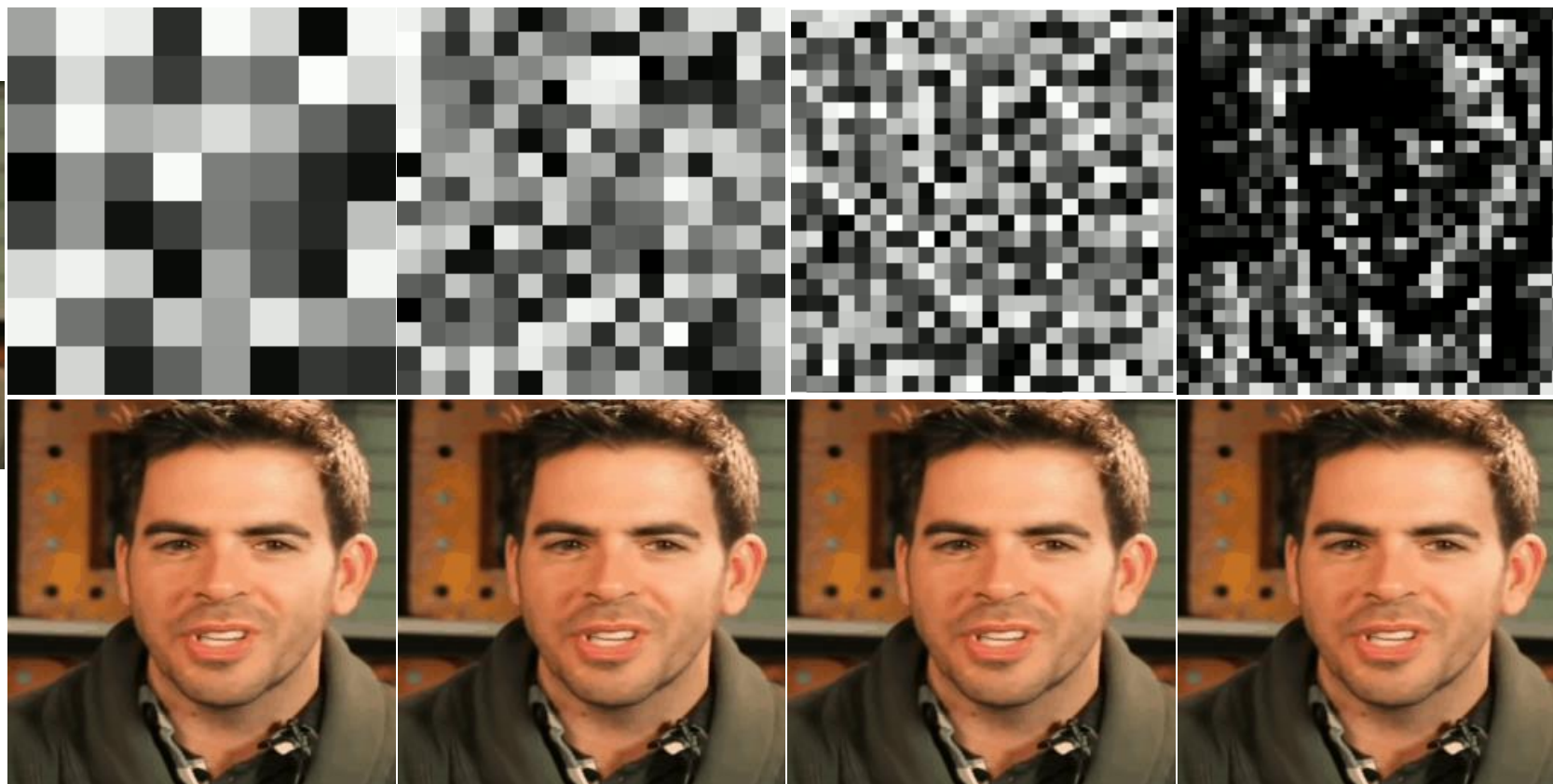
Dataset		SRLR-FOMM		SRLR-CFTE		SRLR-FV2V	
		Rate-DISTS	Rate-LPIPS	Rate-DISTS	Rate-LPIPS	Rate-DISTS	Rate-LPIPS
Average	Class A	-46.06%	7.37%	-47.78%	-1.41%	-42.86%	1.16%
	Class B	-19.77%	23.40%	-15.07%	31.01%	-11.43%	26.40%
	Overall	-32.91%	15.38%	-31.43%	14.80%	-27.14%	13.78%

- ❑ In comparison with the latest VVC codec at high bitrate ranges, the proposed SRLR-based algorithm can achieve performance gains for Rate-DISTS.
- ❑ As for Rate-LPIPS, nearly half of 30 sequences have performance gains, but on average, there are no gains.

Quality Improvement Comparisons

Feature granularity

8*8 16*16 24*24 32*32

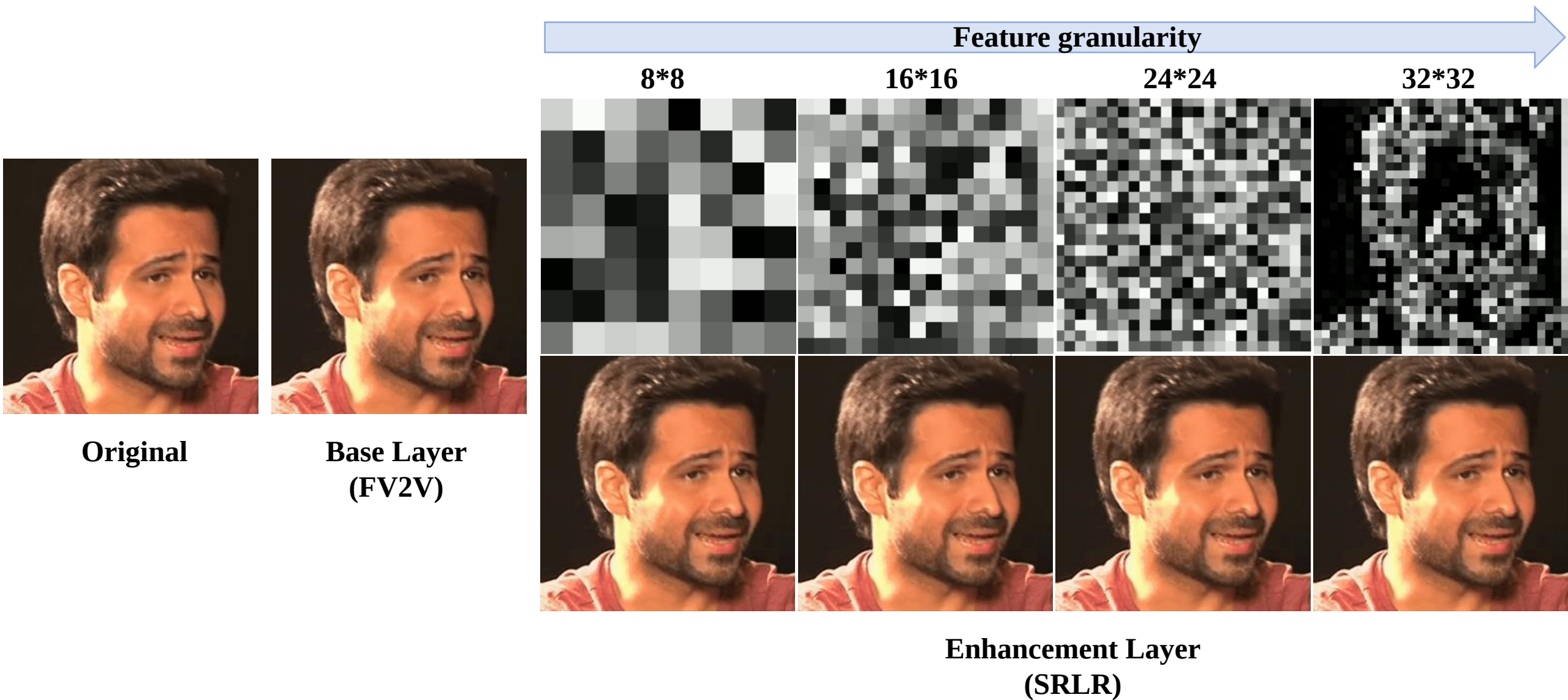


Original

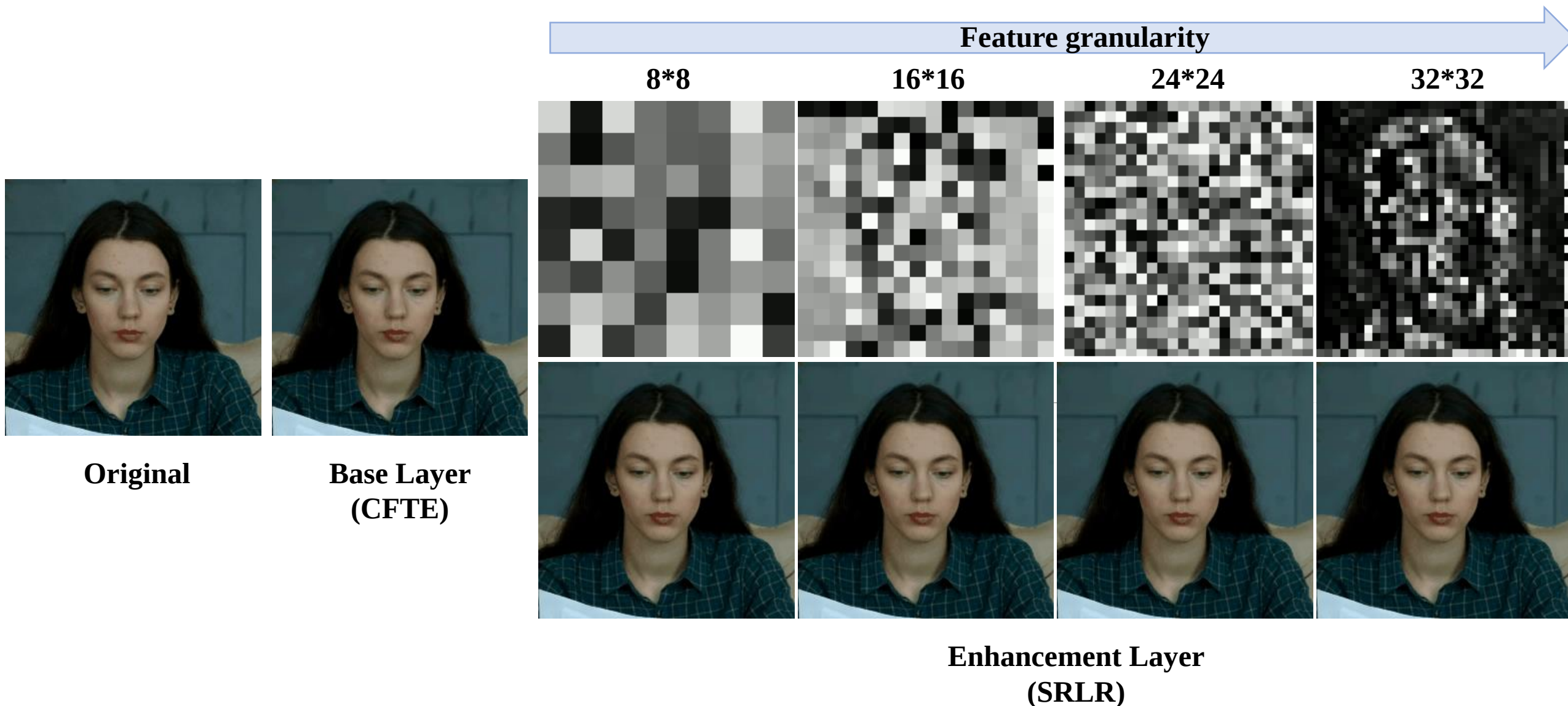
Base Layer
(FOMM)

Enhancement Layer
(SRLR)

Quality Improvement Comparisons



Quality Improvement Comparisons



Subjective Quality Comparisons

VVC
@ 32.33k
DISTS=0.095



SRLR_FOMM
@ 23.61k
DISTS= 0.089



Original

SRLR_CFTE
@ 24.87k
DISTS= 0.091



SRLR_FV2V
@ 25.67k
DISTS= 0.088



Subjective Quality Comparisons

VVC
@ 24.19k
DISTS=0.118



Original

SRLR_CFTE
@ 24.80k
DISTS=0.074



SRLR_FOMM
@ 29.01k
DISTS=0.077



SRLR_FV2V
@ 27.47k
DISTS=0.077



- ❑ This contribution proposes a scalable representation and layered reconstruction (SRLR) algorithm for all existing GFVC approaches, which can provide **a wider coverage of bitrate ranges** for face video compression.
- ❑ Further, visual examples are shown to verify that the proposed SRLR architecture can **reduce some annoying visual artifacts** caused by GFVC methods and improve the benefit of subjective quality.
- ❑ **It is suggested to consider the proposed SRLR architecture in AHG16 software.**



Q & A