Supplementary Material for
A neural video codec with spatial rate-distortion control

Noor Fathima, Jens Petersen, Guillaume Sautière, Auke Wiggers, Reza Pourreza
Qualcomm AI Research[‡]
{mohamedg, jpeterse, gsautie, auke, pourreza}@qti.qualcomm.com

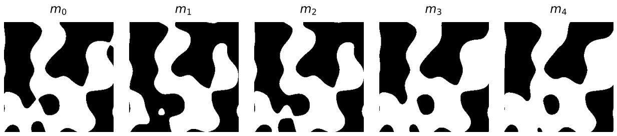


Figure 1: 5 frames from a synthetic binary ROI mask sequence m.

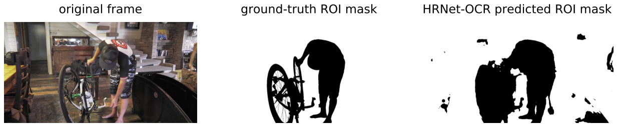


Figure 2: Original frame (left), ground-truth mask (middle) and HRNet-OCR predicted mask (right) for frame 0 of "bike-packing" sequence. IoU of 64.9%

A. Datasets

We evaluate on a common benchmark for video segmentation, the DAVIS dataset [8] (val subset), due to the high quality of its semantic annotations. It contains 90 short and diverse video sequences with varying resolutions. Every video is human labeled with pixel-wise annotations of instances of interest, which we convert to a binary foreground-background mask. As the original videos are compressed already, we downsample the frames to 720p using Pillow [4] in order to reduce the effect of compression artefacts. Additionally, we evaluate our models on common video compression benchmark datasets, specifically the 1920 × 1080 UVG dataset [5] and 1280 × 720 HEVC class-E2 sequences (teleconference) under common test conditions [2]. UVG videos are in full-HD resolution 1920x1080 at high framerates of up to 120 fps, and HEVC class-E is available in 1280x720 resolution at 60 fps.

For qualitative evaluation, our models are also evaluated on videos from Pexels.com (with permissive license and no objectionable content, https://www.pexels.com/license/). Pexels videos are in 4K resolution at 25 or 60 frames per second. Frames are downsampled to 720p to remove any compression artifacts.

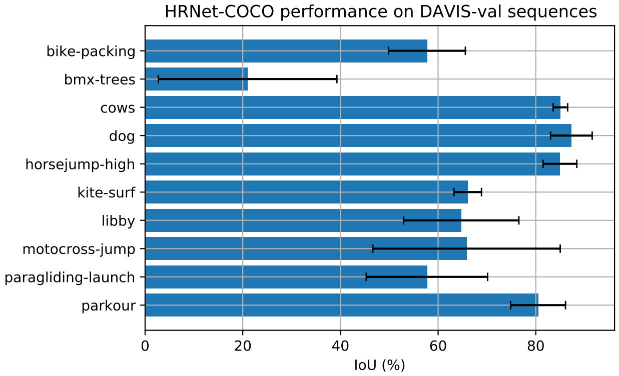


Figure 3: Performance of open-source model HRNet-OCR on some DAVIS-val sequences.

As semantic annotations are not publicly available for UVG and HEVC-E2 datasets, we extract ROI masks using an open source implementation of the segmentation network HRNet [10]. Note that for UVG, all sequences but "HoneyBee" contain semantic classes that HRNet can segment. We therefore leave out the Honeybee sequence from ROI evaluations, and refer to the resulting dataset of frames and semantic annotations as UVG-ROI.

B. ROI masks

This section provides further details on the generation of the synthetic ROI masks used for training. We investigate the influence of quality of masks at test time by evaluating DAVIS-val dataset using annotated masks and automatically generated masks.

B.1. Synthetic ROIs for training

Due to the scarcity of large annotated high-quality video datasets, we generate artificial ROI masks at training time.

[‡]Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

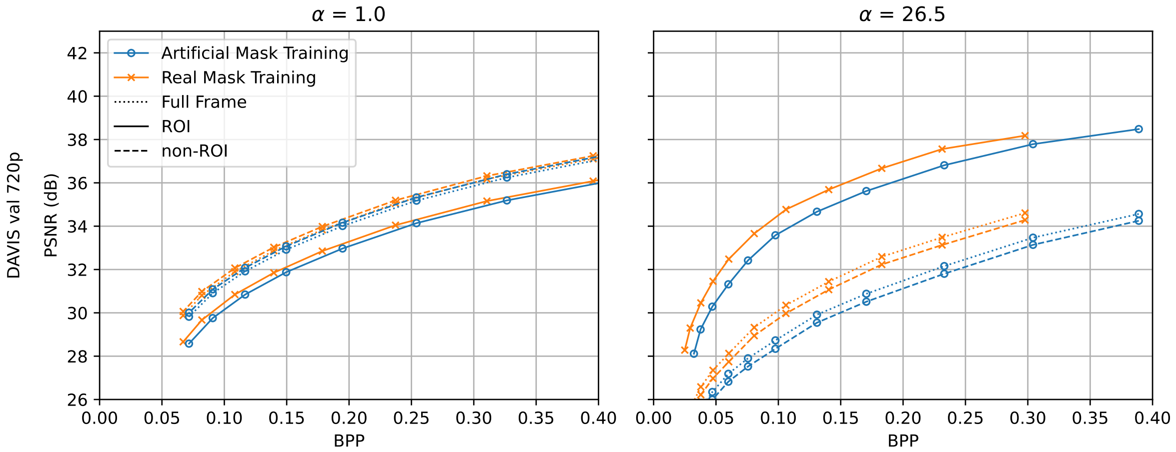


Figure 4: Performance on DAVIS val 720p, when training on DAVIS train with real (orange) vs. artificial (blue) masks.

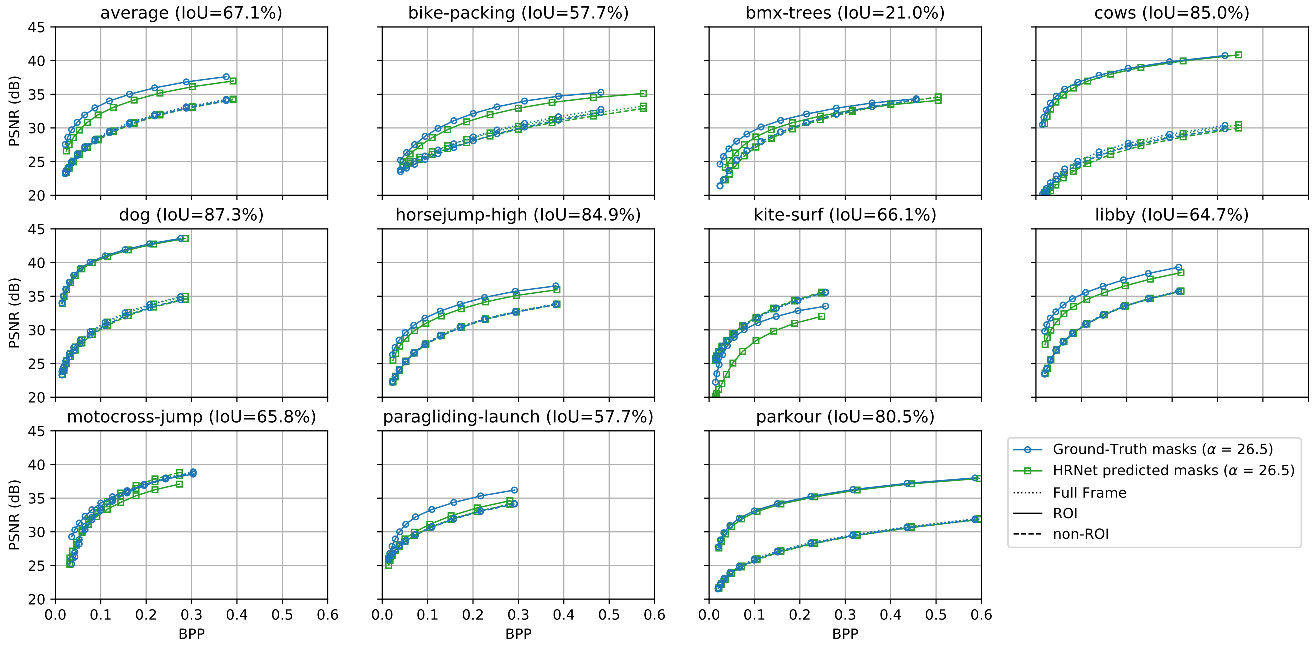


Figure 5: Performance of our model using either ground-truth (blue) or HRNet-OCR predicted (green) masks on individual sequences of DAVIS val.

To generate temporally consistent ROI masks, we use Perlin noise [6] following [7]. Perlin noise is a procedural semi-random noise process originally developed for use in computer graphics to synthesize rich and realistic looking textures. We use the perlin-numpy open-source package to generate the temporal masks with default settings, *i.e.*, noise periods of 1, 4 and 4 for time, height and width respectively. We apply a threshold at 0 to binarize them. We show an example of such mask in Fig. 1.

It should be noted that it is still preferable to use real annotations where available, the main benefit of synthetic masks is that much larger unannotated datasets can be used, which in turn leads to better performance and generalization (see Fig. 6 of the main body). In Fig. 4 we show the difference of training with real vs. artificial masks on the same (small) dataset DAVIS train, evaluated on DAVIS val. Here the real masks lead to better performance. Even for $alpha = 1$, where ideally the two should be identical, we see a small difference, meaning the model has not learned to completely ignore the mask input in this setting.

B.2. Predicted masks for evaluation

We use the official open-source pre-trained HRNet-OCR [10] trained on COCO-Stuff [3] to extract masks for the UVG [5] and HEVC-E2 [2] sequences, specifically model checkpoint `hrnet_ocr_cocostuff_3965_torch04.pth`. We use the class "human" for HEVC-E2, and the classes "human", "boat", "dog" and "horse" for UVG sequences. Finally, we convert the output of HRNet to a binary ROI mask by classifying as foreground all pixels from the class labels defined above. As outlined in Sec. 4.1, "HoneyBee" does not contain any classes from the COCO dataset. Consequently we exclude it from our evaluation, and call the resulting dataset of frames and ROI masks UVG-ROI.

To get a sense of the quality of ROI masks generated with HRNet-OCR, we used it to extract masks for 10 DAVIS val sequences for which there are both (1) one-to-one correspondence between the instance annotated in the ground-truth and the COCO classes and (2) no ambiguity on the instance when there are more than one object with the same class in the video. This allowed us to compute intersection-over-union (IoU) of HRNet-OCR masks with respect to DAVIS-val ground-truth, as shown in Fig. 3. In Fig. 2 we show as an example the ground-truth and predicted masks for the first frame of the "bike-packing" sequence.

Additionally, we evaluated our model on these sequences using the predicted masks as input instead of the ground-truth annotations. In Fig. 5 we compared R-D performances for each sequence between evaluating our model with (a) ground-truth or (b) HRNet-OCR predicted masks. Note that in both cases ROI and non-ROI PSNR metrics are computed using the ground-truth masks. We observe that when IoU is reasonably high ($> 80\%$), rate-distortion performances are quite similar, yet they tend to degrade quickly when below 70% except for a few sequences like "bike-packing" or "bmx-trees". It would seem our model suffers worse rate-distortion when the ROI is both small and inconsistent like in "kite-surf" and "paragliding-launch".

Finally, we train one model on DAVIS with ground-truth annotations, and one model on Vimeo90k with synthetic masks. We evaluate both models on DAVIS-val, using either ground-truth masks, or the masks predicted using HRNet-OCR. We show BD-rate (with respect to SSF) in Fig. 6 in main text. For both models, we observe that for small α values, performance is similar for ground-truth masks and predicted masks. However, at high α, non-ROI BD-rate degrades quicker when using predicted masks.

In conclusion, our model is moderately sensitive to "noisy" masks, especially for small α values, which is a likely operating point for most realistic use cases. Still, this result confirms that extracting high quality ROI masks at test time is crucial to produce reconstructions with high fidelity.

Table 1: BD rate gain with respect to SSF, lower is better.

Codec	BD rate gain
MR+MD SSF	+13.5%
ELF-VC	-25.0%
H.265	+63.7%

C. Additional qualitative examples

C.1. Comparison to single-rate SSF

In Figures 6, 7 we give two additional examples for a comparison with a single-rate SSF model, similar to Fig. 5. For the comparison we start with the SSF model at a given β (we trained models at 8 different β's), then tune β for our model with $\alpha = 1.0$ to match the BPP of the reference. The resulting performance for our model is slightly worse, which is expected for a multi-rate, multi-distortion approach. We then take the baseline model at a different β and tune α for our model (keeping β from the first step fixed) to again match the BPP of the baseline. We find that our model substantially outperforms the reference in ROI PSNR, at a moderate cost in non-ROI PSNR.

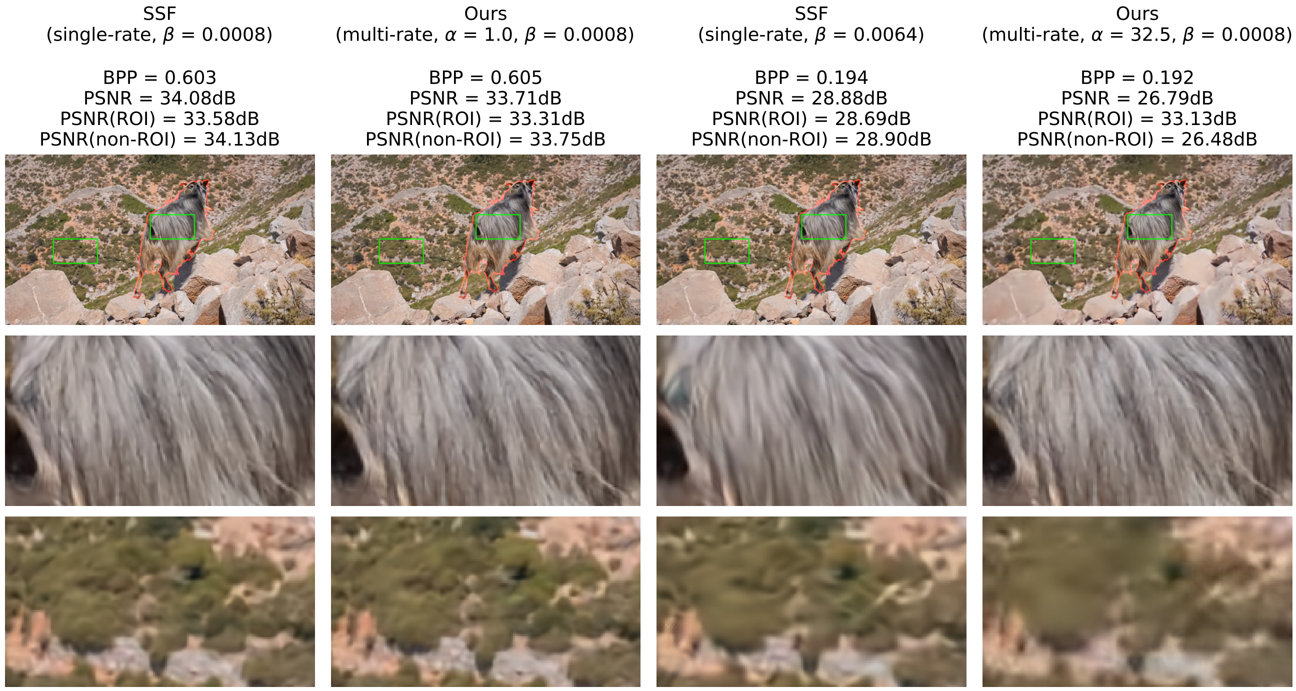
C.2. Effect of varying β and α

In Figures 9, 10 we show a frame from the "Soapbox" video (same frame as Fig. 5), reconstructed by our model using different values of β and α. The crop is chosen so that foreground and background are visible simultaneously. It can be seen that β (varies across columns) chooses a general tradeoff between rate and distortion, while α essentially trades off rate and background distortion, while keeping foreground distortion fixed. Fig. 9 shows a large range of α values, so that changes are salient and effects are easier to understand. Fig. 10 on the other hand demonstrates a range of α's that we deem realistic for real-world use, i.e. one where the drop in background quality is small for any given β. As a matter of fact, for our model operating at a higher rate (left column, $\beta = 0.0001$), we are unable to visually distinguish the background going from $40.78\,\text{dB}$ PSNR to $37.98\,\text{dB}$ in the chosen crop. At the same time this results in savings of $0.250\,\text{bpp}$ (29.7%), while foreground quality remains constant.

D. Additional analyses

D.1. Comparison to non-ROI codec literature

A desirable property of a ROI-based codec is that it matches performance of a non-ROI codec when the ROI spans the entire frame. To place our result in context, we compare our proposed codec on the UVG dataset in Fig. 8, by showing rate-distortion curves for $\alpha = 1$ where the ROI spans the entire frame, and showing the BD-rate gains in



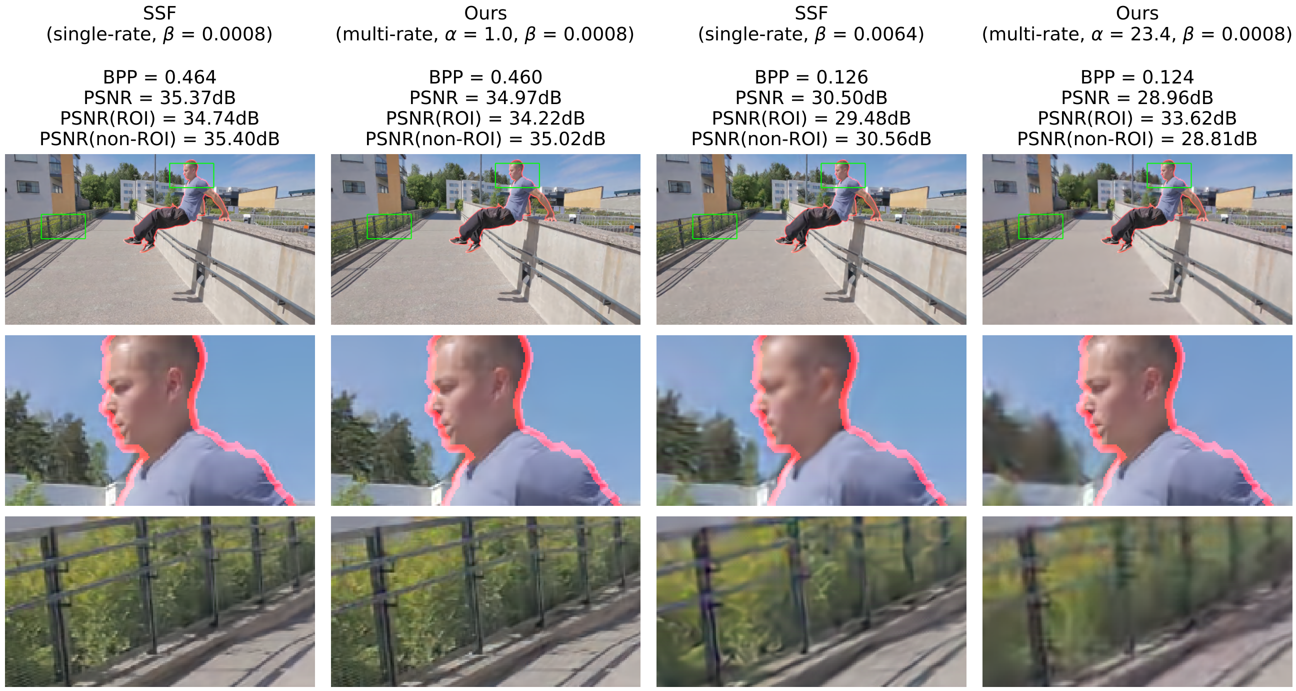
SSF (single-rate, β = 0.0008)
BPP = 0.603
PSNR = 34.08dB
PSNR(ROI) = 33.58dB
PSNR(non-ROI) = 34.13dB

Ours (multi-rate, α = 1.0, β = 0.0008)
BPP = 0.605
PSNR = 33.71dB
PSNR(ROI) = 33.31dB
PSNR(non-ROI) = 33.75dB

SSF (single-rate, β = 0.0064)
BPP = 0.194
PSNR = 28.88dB
PSNR(ROI) = 28.69dB
PSNR(non-ROI) = 28.90dB

Ours (multi-rate, α = 32.5, β = 0.0008)
BPP = 0.192
PSNR = 26.79dB
PSNR(ROI) = 33.13dB
PSNR(non-ROI) = 26.48dB

Figure 6: Example reconstructions from the ”goat” sequence from DAVIS val, compared to a single-rate SSF model.



SSF (single-rate, β = 0.0008)
BPP = 0.464
PSNR = 35.37dB
PSNR(ROI) = 34.74dB
PSNR(non-ROI) = 35.40dB

Ours (multi-rate, α = 1.0, β = 0.0008)
BPP = 0.460
PSNR = 34.97dB
PSNR(ROI) = 34.22dB
PSNR(non-ROI) = 35.02dB

SSF (single-rate, β = 0.0064)
BPP = 0.126
PSNR = 30.50dB
PSNR(ROI) = 29.48dB
PSNR(non-ROI) = 30.56dB

Ours (multi-rate, α = 23.4, β = 0.0008)
BPP = 0.124
PSNR = 28.96dB
PSNR(ROI) = 33.62dB
PSNR(non-ROI) = 28.81dB

Figure 7: Example reconstructions from the ”parkour” sequence from DAVIS val, compared to a single-rate SSF model.

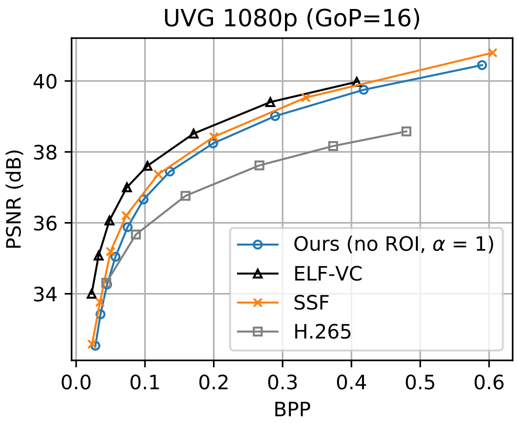


Figure 8: Performance on all UVG sequences at 1080p (GoP size 16)

Table. 1. We consider three baselines: 1) the SSF model, trained separately for every point, 2) H.265, using the ffmpeg implementation in *low-delay* setting, 3) ELF-VC [9], a recent variable-bitrate neural codec. Note that for consistency with ELF-VC, we use GoP size of 16 for all methods in this evaluation.

Performance with respect to SSF is only slightly reduced, illustrating that our model adds flexibility at little cost. ELF-VC demonstrates stronger performance, but our approach is largely model-agnostic and should translate well to convolution-based architectures. We emphasize that SSF no longer offers state-of-the-art performance, but it is a well tested baseline that allows us to demonstrate the benefits of our multi-distortion approach. Importantly, there is an established open-source implementation [1], while more recent architectures like ELF-VC do not offer one. We initially tried to reproduce ELF-VC to use it as a base model, but did not manage to train it in a stable manner. Nevertheless, we believe our findings should translate easily to other architectures, as the conditioning blocks (see Fig. 2 in the main body) can be added to any convolutional layer.

D.2. Constrained coding

The flexibility of our codec allows us to code under various rate and quality constraints, including maximum bitrate and minimum ROI quality. We show a proof of concept of these in Fig. 11, where β and α are selected using an oracle (i.e. we grid-search for optimal α and β values for each frame). The variable rate codec is able to reach the rate constraint well, but fails to achieve the 32dB PSNR threshold for the ROI. On the other hand, our model is able to meet the rate constraint by trading off non-ROI PSNR for ROI PSNR, all while meeting the rate constraint.

D.3. Performance on teleconferencing content

Teleconferencing is a commonly chosen example to motivate the use of semantics-aware coding: the foreground (i.e. person) is important, the background is of little interest. Fig. 12 shows example reconstructions for a teleconferencing sequence, taken from Pexels.com. As for Figures 6, 7, we take a single-rate SSF model trained with two different β's as baseline, and tune β and α for our model to match their rate, as described in Sec. C.1. Unlike before, even using the entire alpha range does not allow us to match the BPP for the lower-rate baseline, even though the two baselines have the smallest rate difference possible for the different models we trained. This indicates that a semantics-aware approach is a poor choice for this use case, and that even a model without foreground and background distinction can compress the contents remarkably well.

We observe that teleconferencing videos rarely contain extreme scene changes. However, background (any area apart from a human/object in the frame) occupies a significant part of the frame. Therefore, in case of the MR+MD SSF model, always transmitting a low-quality background hurts the performance. We propose a simple yet effective solution to alleviate this issue. We transmit a high-quality *I-frame* (at $\alpha = 1$) followed by *P-frame* at suitable α. This ensures that the background (which rarely changes) is transmitted in the best possible quality and transmitting *P-frames* at corresponding α also provides us with bitrate savings. We show in Sec. 5.6 the main text the quantitative effect of this simple change on HEVC E2.

In addition, we include a video example in the supplementary section titled *teleconferencing.mp4*. Here we show a crop of a video from the Pexels dataset in a 2×2 grid. *Top-left* is a variable bitrate model operating at a similar bitrate as our MR+MD model. *Top-right* is our model (LQ I-frame). *Bottom-left* is our model, however here we transmit a high-quality *I-frame* (HQ I-frame). *Bottom-left* is the corresponding semantic mask. We see that transmitting high-quality *I-frame* provides us with quality similar to the variable rate model but at a much lower rate ($0.050\,\mathrm{bpp}$ compared to $0.071\,\mathrm{bpp}$).

D.4. Effect of high-quality I-frames

We show the effect of error propagation for all datasets in Fig. 13, along with our suggested remedy of setting $\alpha = 1$ in I-frames. The row for HEVC E2 is a repeat of the results shown in the main body. While our approach is very helpful for HEVC E2 (static content similar to teleconferencing), saving approximately 80% BD-rate in the non-ROI while only incurring a penalty of ca. 5% BD-rate in the ROI, the situation is different for the other datasets, most strikingly for DAVIS val. Here we find that all savings in the non-ROI come at a similar cost in the ROI, contrary to the desired behaviour. We conclude that high-quality ($\alpha = 1$) I-frames

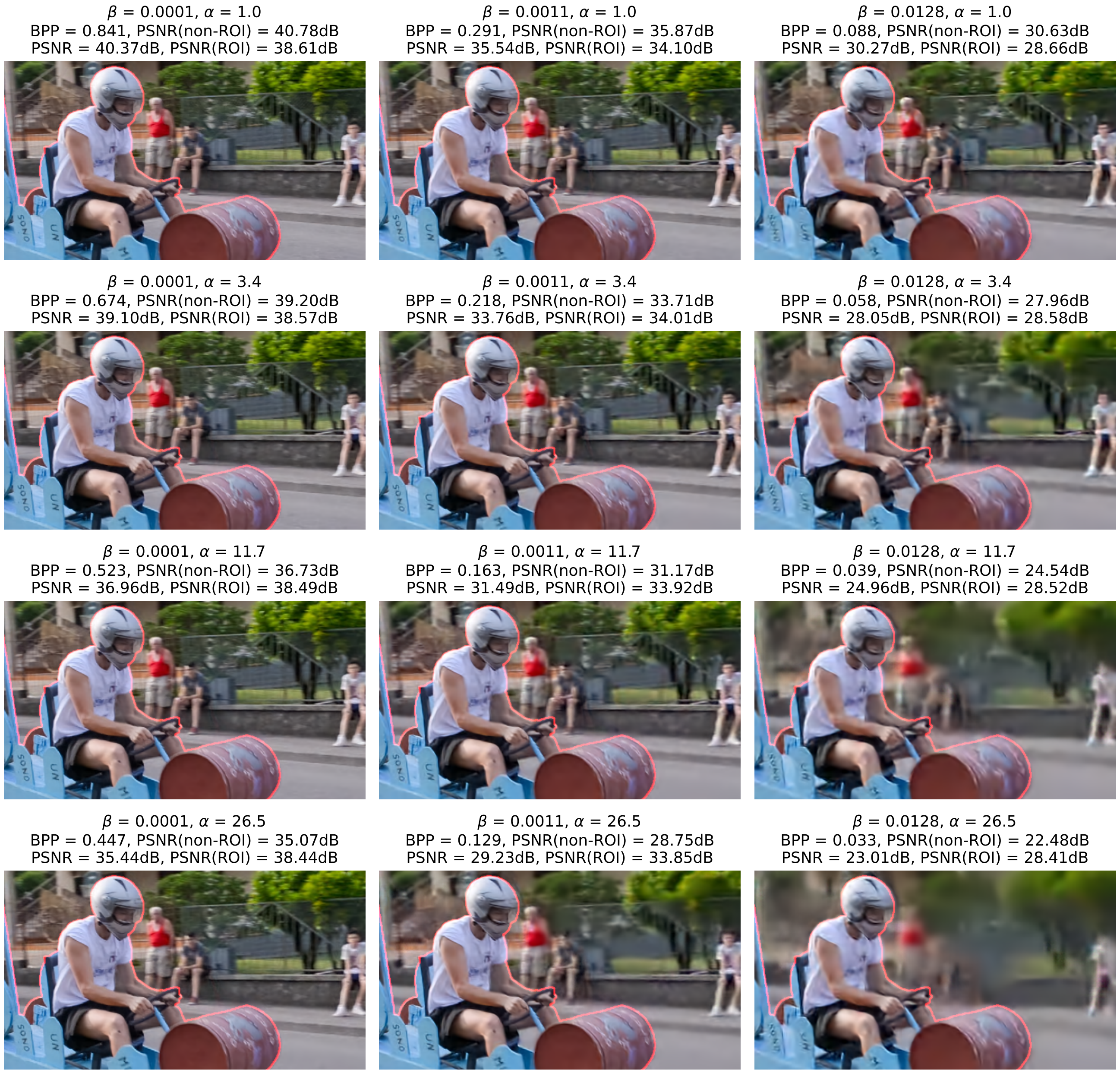
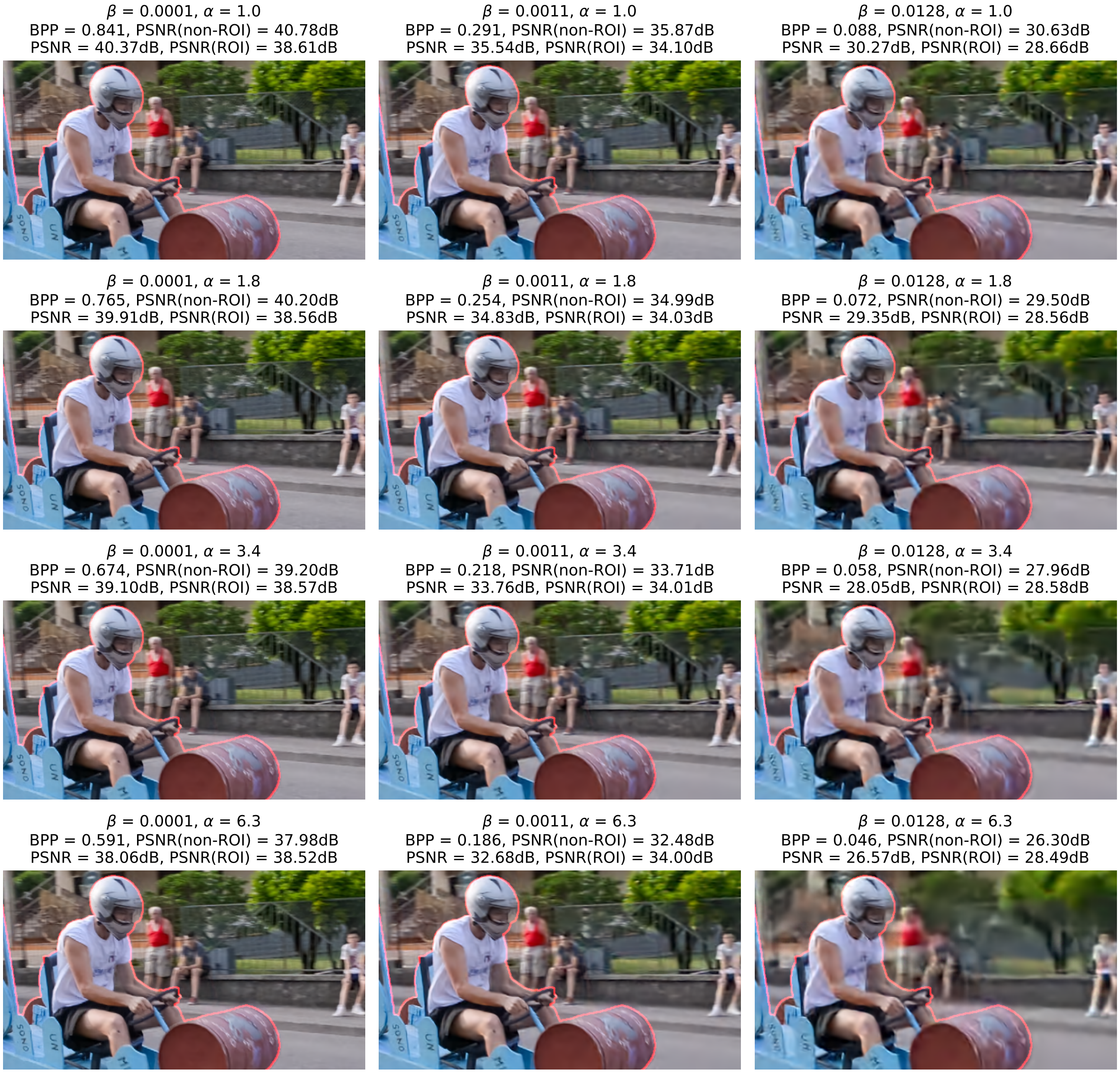

β = 0.0001, α = 1.0
BPP = 0.841, PSNR(non-ROI) = 40.78dB
PSNR = 40.37dB, PSNR(ROI) = 38.61dB
β = 0.0011, α = 1.0
BPP = 0.291, PSNR(non-ROI) = 35.87dB
PSNR = 35.54dB, PSNR(ROI) = 34.10dB
β = 0.0128, α = 1.0
BPP = 0.088, PSNR(non-ROI) = 30.63dB
PSNR = 30.27dB, PSNR(ROI) = 28.66dB
β = 0.0001, α = 3.4
BPP = 0.674, PSNR(non-ROI) = 39.20dB
PSNR = 39.10dB, PSNR(ROI) = 38.57dB
β = 0.0011, α = 3.4
BPP = 0.218, PSNR(non-ROI) = 33.71dB
PSNR = 33.76dB, PSNR(ROI) = 34.01dB
β = 0.0128, α = 3.4
BPP = 0.058, PSNR(non-ROI) = 27.96dB
PSNR = 28.05dB, PSNR(ROI) = 28.58dB
β = 0.0001, α = 11.7
BPP = 0.523, PSNR(non-ROI) = 36.73dB
PSNR = 36.96dB, PSNR(ROI) = 38.49dB
β = 0.0011, α = 11.7
BPP = 0.163, PSNR(non-ROI) = 31.17dB
PSNR = 31.49dB, PSNR(ROI) = 33.92dB
β = 0.0128, α = 11.7
BPP = 0.039, PSNR(non-ROI) = 24.54dB
PSNR = 24.96dB, PSNR(ROI) = 28.52dB
β = 0.0001, α = 26.5
BPP = 0.447, PSNR(non-ROI) = 35.07dB
PSNR = 35.44dB, PSNR(ROI) = 38.44dB
β = 0.0011, α = 26.5
BPP = 0.129, PSNR(non-ROI) = 28.75dB
PSNR = 29.23dB, PSNR(ROI) = 33.85dB
β = 0.0128, α = 26.5
BPP = 0.033, PSNR(non-ROI) = 22.48dB
PSNR = 23.01dB, PSNR(ROI) = 28.41dB

Figure 9: Effect of varying β and α in our model.

are only useful for content with mostly static background. We also include a column for $\alpha = 1$ P-frames as a sanity check. As expected, the curves coincide perfectly, as low-quality and high-quality I-frames are identical in this case.

D.5. RD performance for individual videos

In Figures 14, 15, 16 we show RD curves for all individual videos in the respective test sets. Each figure essentially corresponds to the middle column of Fig. 2 in the main body, using $\alpha = 26.5$. We find that there is large variation in performance between different videos, underlining the fact that the utility of semantics-aware coding depends largely on the use case.



β = 0.0001, α = 1.0
BPP = 0.841, PSNR(non-ROI) = 40.78dB
PSNR = 40.37dB, PSNR(ROI) = 38.61dB

β = 0.0011, α = 1.0
BPP = 0.291, PSNR(non-ROI) = 35.87dB
PSNR = 35.54dB, PSNR(ROI) = 34.10dB

β = 0.0128, α = 1.0
BPP = 0.088, PSNR(non-ROI) = 30.63dB
PSNR = 30.27dB, PSNR(ROI) = 28.66dB

β = 0.0001, α = 1.8
BPP = 0.765, PSNR(non-ROI) = 40.20dB
PSNR = 39.91dB, PSNR(ROI) = 38.56dB

β = 0.0011, α = 1.8
BPP = 0.254, PSNR(non-ROI) = 34.99dB
PSNR = 34.83dB, PSNR(ROI) = 34.03dB

β = 0.0128, α = 1.8
BPP = 0.072, PSNR(non-ROI) = 29.50dB
PSNR = 29.35dB, PSNR(ROI) = 28.56dB

β = 0.0001, α = 3.4
BPP = 0.674, PSNR(non-ROI) = 39.20dB
PSNR = 39.10dB, PSNR(ROI) = 38.57dB

β = 0.0011, α = 3.4
BPP = 0.218, PSNR(non-ROI) = 33.71dB
PSNR = 33.76dB, PSNR(ROI) = 34.01dB

β = 0.0128, α = 3.4
BPP = 0.058, PSNR(non-ROI) = 27.96dB
PSNR = 28.05dB, PSNR(ROI) = 28.58dB

β = 0.0001, α = 6.3
BPP = 0.591, PSNR(non-ROI) = 37.98dB
PSNR = 38.06dB, PSNR(ROI) = 38.52dB

β = 0.0011, α = 6.3
BPP = 0.186, PSNR(non-ROI) = 32.48dB
PSNR = 32.68dB, PSNR(ROI) = 34.00dB

β = 0.0128, α = 6.3
BPP = 0.046, PSNR(non-ROI) = 26.30dB
PSNR = 26.57dB, PSNR(ROI) = 28.49dB

Figure 10: Effect of varying β and α in our model, with smaller α values.

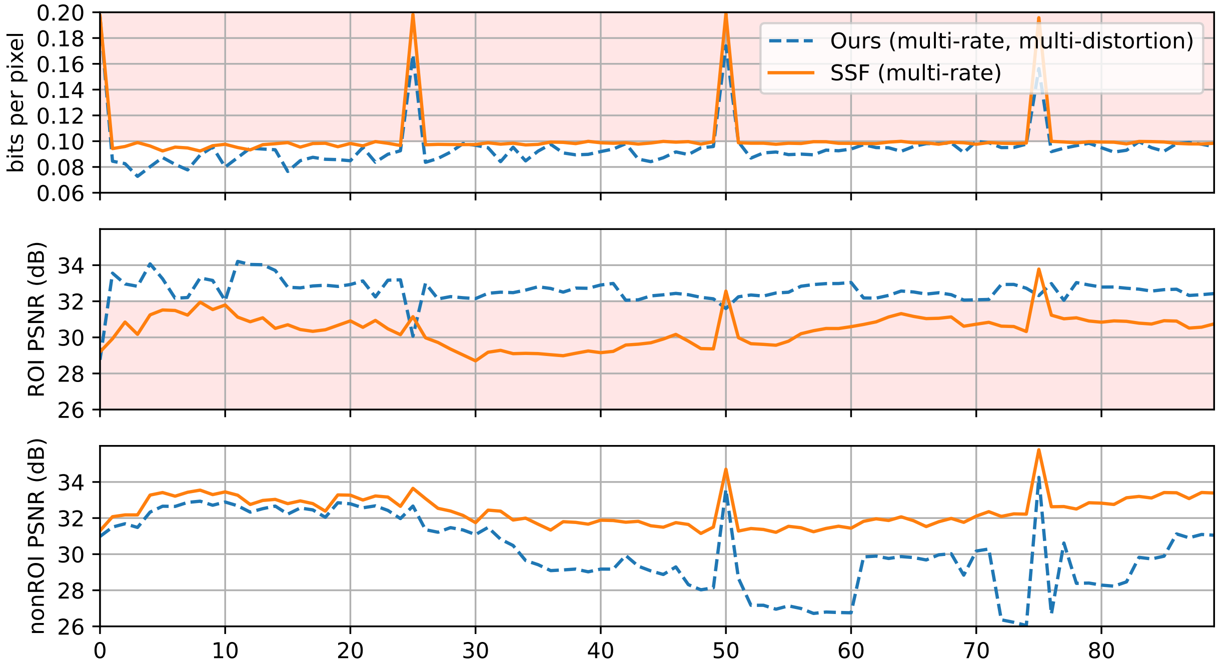


Figure 11: Fixed rate (≤ 0.1bpp for P-frames), fixed ROI quality (≥ 32dB) coding for video ‘Soapbox’, comparing our model to a multi-rate baseline SSF. The used group-of-pictures size is 25.

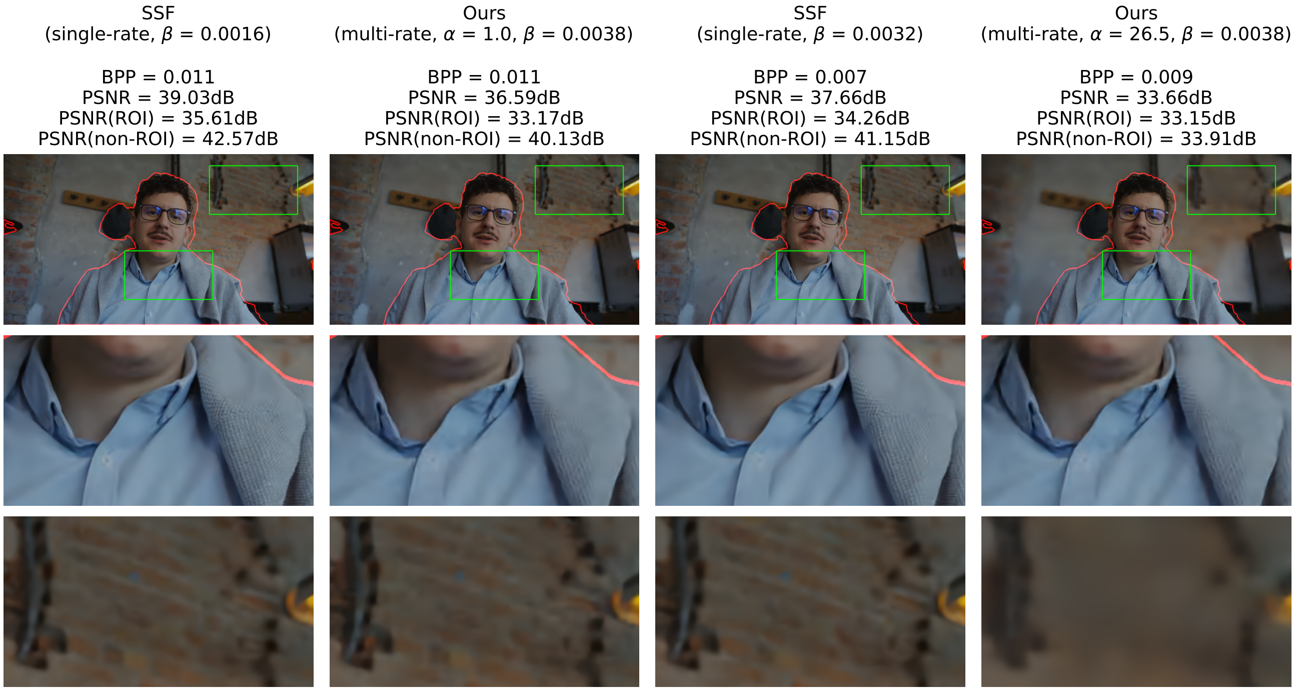


Figure 12: Example reconstructions for a teleconferencing sequence, taken from Pexels.com.

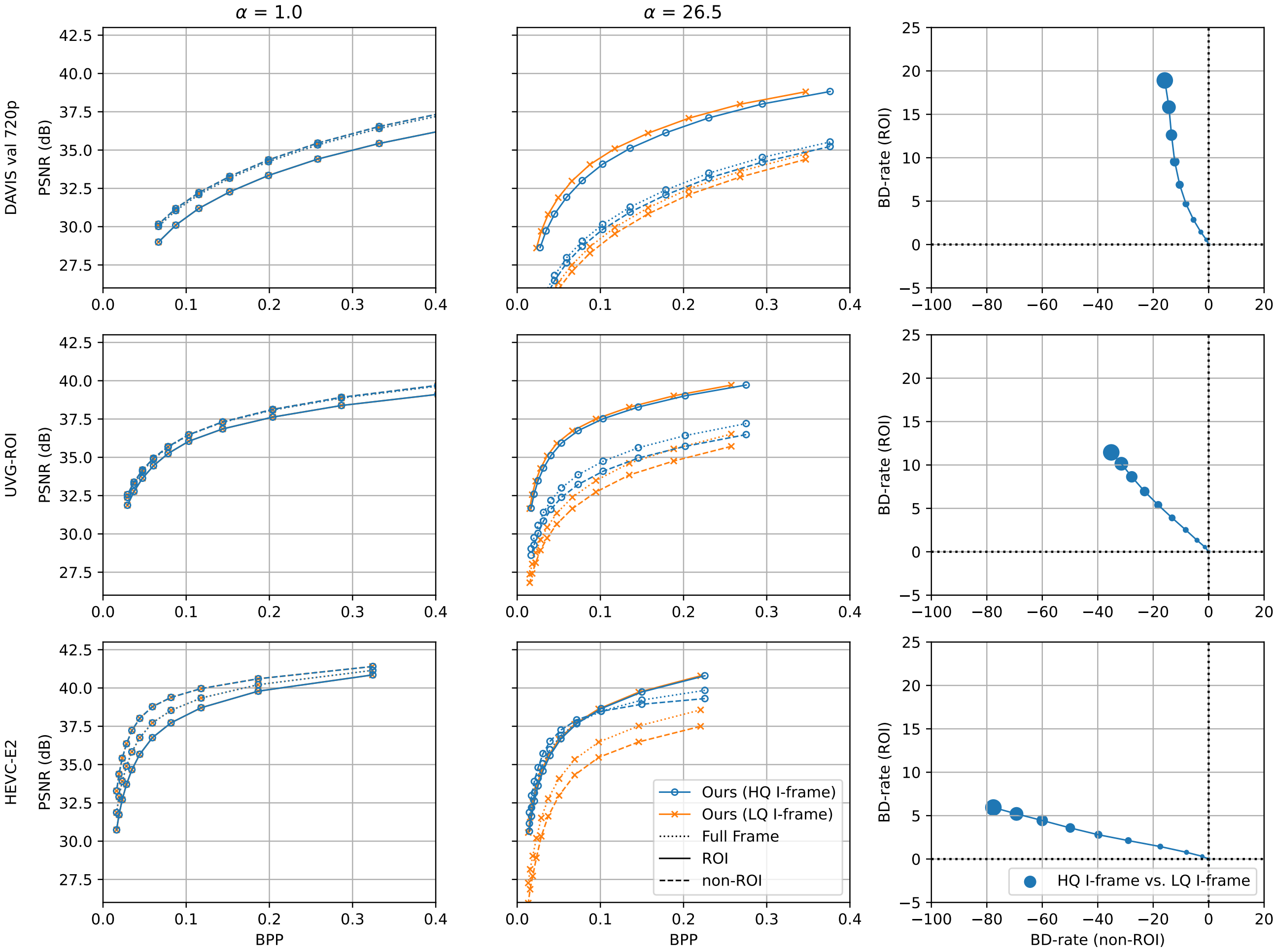


Figure 13: Effect of using $\alpha = 1$ I-frames (*HQ I-frame*) instead of the same α that's used in the subsequent P-frames (*LQ I-frame*)

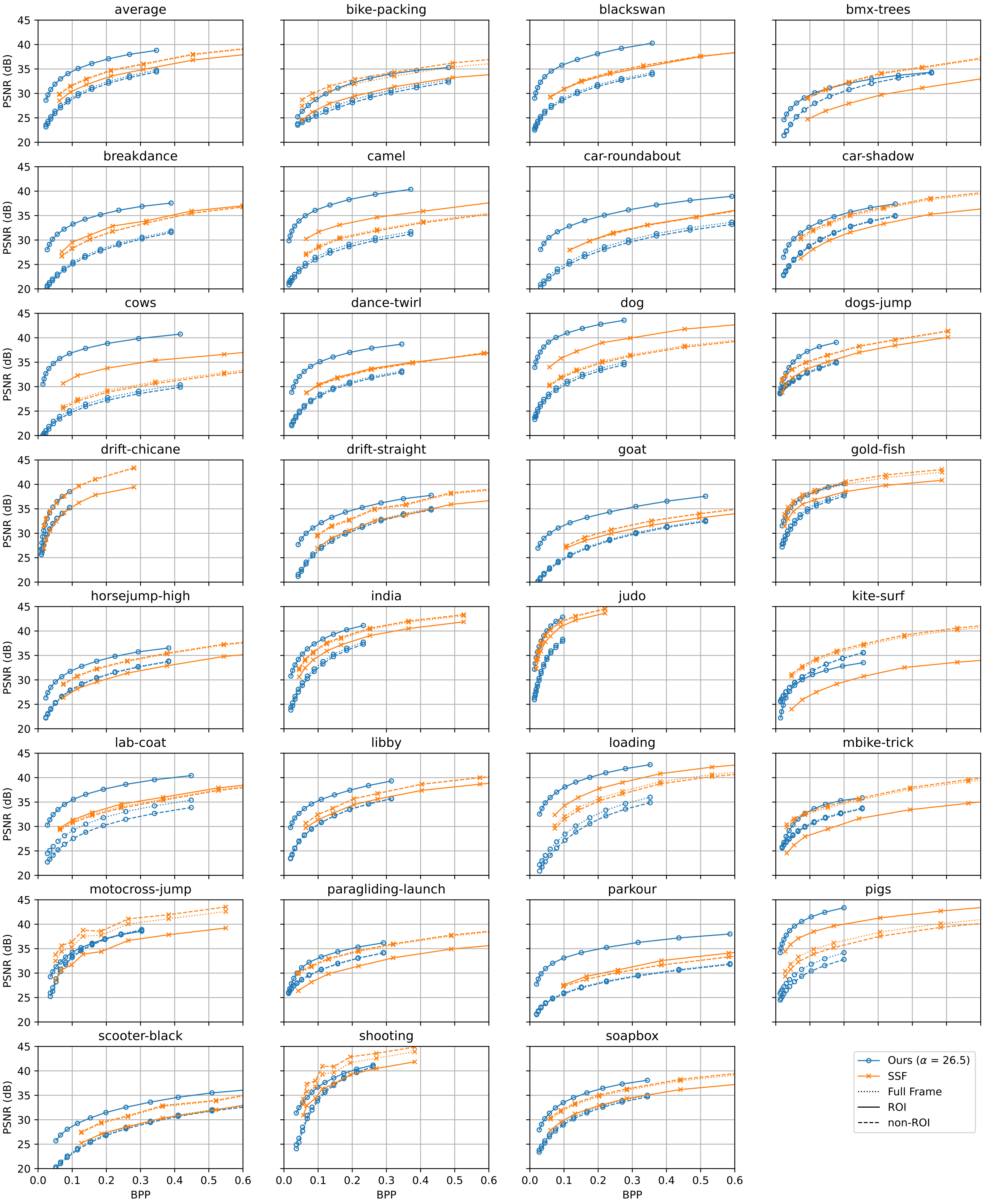


Figure 14: RD curves for individual videos in the DAVIS val set, $\alpha = 26.5$.

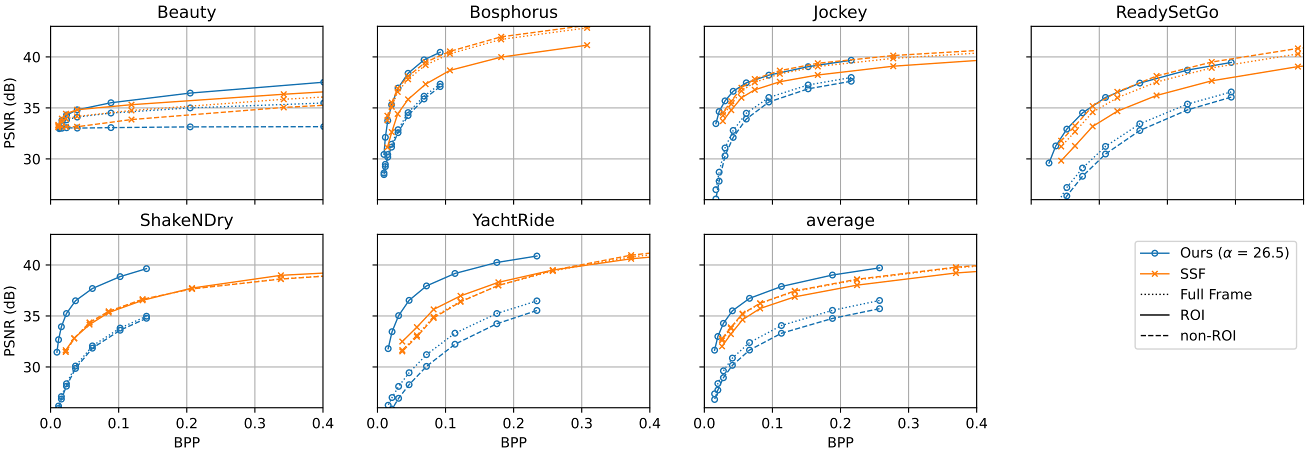


Figure 15: RD curves for individual videos in the UVG-ROI set, $\alpha = 26.5$.

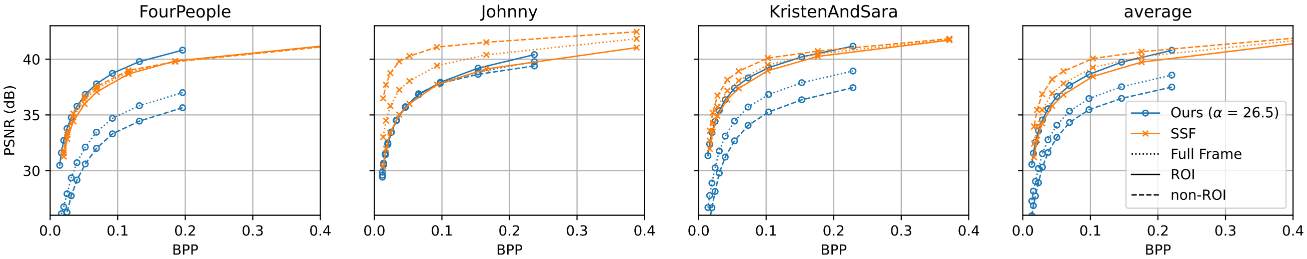


Figure 16: RD curves for individual videos in the HEVC-E2 set, $\alpha = 26.5$.

References

[1] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. CompressAI: a PyTorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.

[2] Frank Bossen et al. Common test conditions and software reference configurations. *JCTVC-L1100*, 12(7), 2013.

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018.

[4] Alex Clark. Pillow (pil fork) documentation, 2015.

[5] Alexandre Mercat, Marko Viitanen, and Jarno Vanne. UVG dataset: 50/120fps 4k sequences for video codec analysis and development. In *Proceedings of the 11th ACM Multimedia Systems Conference*, MMSys '20, page 297–302, New York, NY, USA, 2020. Association for Computing Machinery.

[6] Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985.

[7] Yura Perugachi-Diaz, Guillaume Sautière, Davide Abati, Yang Yang, Amirhossein Habibian, and Taco Cohen. Region-of-interest based neural video compression. *arXiv:2203.01978*, 2022.

[8] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[9] Oren Rippel, Alexander G Anderson, Kedar Tatwawadi, Sanjay Nair, Craig Lytle, and Lubomir Bourdev. Elf-vc: Efficient learned flexible-rate video coding. *Neural Information Processing Systems*, 2021.

[10] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.