Supplementary Material – Patch-VQ: 'Patching Up' the Video Quality Problem

A. Cropping Patches

Deciding number of scales for cropping v-patches: In a psychometric study, specifically based on evaluating video quality, a subject needs roughly 15-20 seconds to rate each content. This limited the number of v-patches we could collect ratings on, and thus we decided to only include **one scale** for each type of v-patches. Scale here defines the dimensions of the v-patches, or the proportion of the video data contained in the patches. For simplicity, we use the same scale (40% of original dimensions) for extracting the three types of v-patches. Additional examples of extracted v-patch triplets have been shown in Fig. 1.

Deciding size of v-patches: Empirically, sv-patches cropped at large scales are not local enough, and do not capture the local quality features satisfactorily. Alternately, smaller scales result in tv-patches too short in duration to collect reliable judgements. Similarly, the resulting stv-patches are too small and short to rate comprehensibly and reliably. We determined 40% to be the most suitable scale after examining v-patch samples.

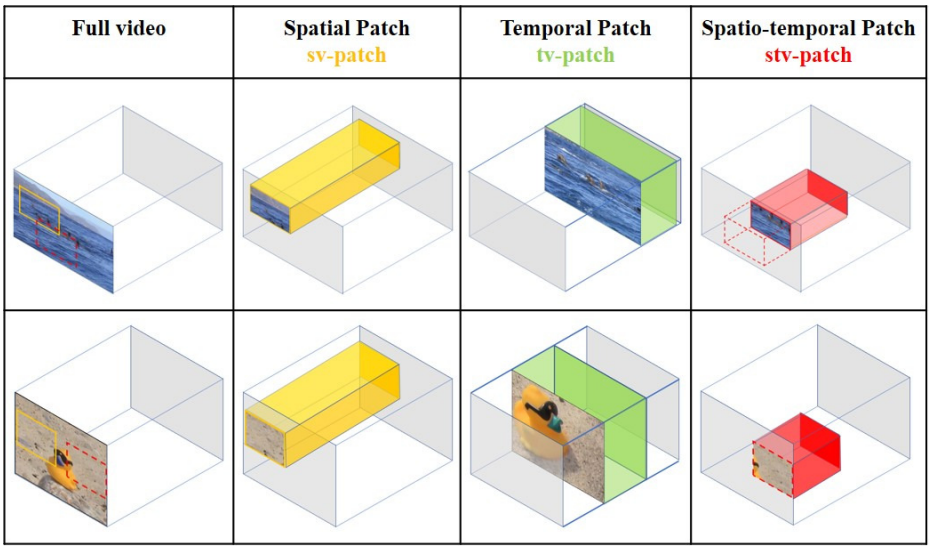


Fig. 1: **Examples of video patch** (v-patch) triplets cropped from random space-time volumes from two exemplar videos in the dataset. All v-patches are videos.

B. Dataset

B.1 Inter-subject consistency plots:

We have mentioned the average SRCC values, representative of the inter-subject consistencies, in Sec. 3.2.4. Along with that, we present the scatter plots of the two sets of subject MOS in Fig. 2. The narrow spread of the plots shows the high agreement, and hence higher consistency, among subject ratings. We also notice that the spread is highest (or, the correlation is lowest) in the case of stv-patches. This can be attributed to the fact that they account for only 6.4% of the video pixel volume, and sometimes, distortions prominent locally, might get masked or have little impact on perceived global video quality.

B.2 Consistency among subject demographics:

We utilized the subject data to study the effects of device parameters on MOS. The SRCC calculated between laptops and desktop computers (the most used devices in the study) was **0.7**, whereas that between videos viewed on phones and other devices was **0.5**. Although we collected relatively little data (3.7%) from phones, this reinforces the notion that perceptual video quality is impacted by viewing on a small device screen. We obtained the following correlations between the two major resolutions: 768×1366 and 640×360 (**0.76**); major viewing distances: less than 15 inches and 15-30 inches (**0.76**); major

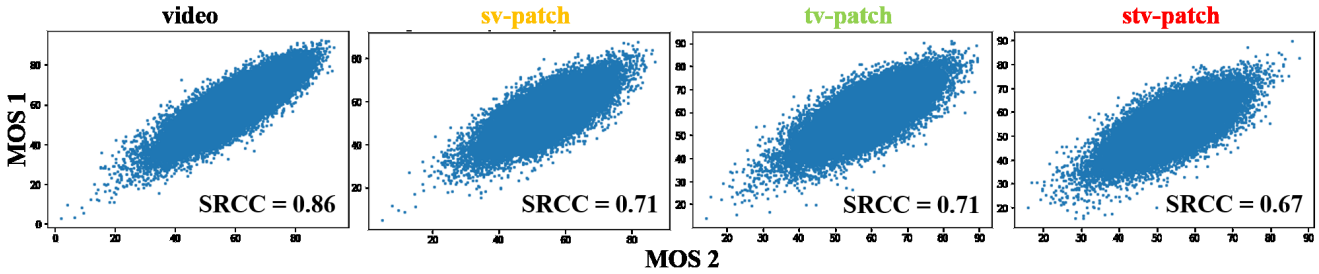


Fig. 2: **Inter-subject consistency:** Inter-subject scatter plot of MOS calculated between random 50% divisions of the human labels on all 39K videos (first from left) into disjoint subject sets. The same is plotted for the sv-patches (second), tv-patches (third) and stv-patches (fourth).

age groups: 20-30 and 30-40 (**0.79**); and genders (**0.8**), all of which are high, but low enough to be suggestive of further study. The consistency among the ratings from diverse subject demographics, when accumulated, result in the overall high consistency of the data (Sec. B.1), validating our data collection and cleaning methodologies.

B.3 Effect of playback delays on video quality:

Delays during playback could impact video quality [3]. We found that $> 96\%$ of the videos were viewed with delays < 1s., while 86% of the videos played without delays. By comparing the scores of the delayed videos against the "golden" scores, we found that device delays had negligible impact on the mean scores, and that eliminating scores associated with delays did not impact data consistency. Hence, we did not impose device delays as a rejection criteria.

B.4 Outlier rejection:

We removed the outliers in our data in two steps as described briefly in Sec. 3.2.3 - **outlier subject rejection** and **outlier score rejection**. The former rejection was video independent, whereas the latter was subject independent. Here, we elaborate the outlier score rejection, which was executed on all videos individually. We followed the standard outlier rejection techniques, but the technique applied was dependent on the score distribution. If, for a video, the scores were (approximately) Gaussian, the modified Z-scores method [1] was applied, which is based on calculating the standard deviation of the distribution. Calculating the kurtosis helped determine the normality of the score distribution. Alternately, if the scores were deemed to be not normal, then we applied the Tukey IQR [5] detection technique, which is based on calculating the interquartile range and is a more generalized method. Tuning the outlier rejection methods based on the nature of the score distribution yielded better consistency scores.

C. Modeling Details

For training PVQ (Sec. 4), we used the Adam optimizer with $\beta_1 = .9$ and $\beta_2 = .99$, a weight decay of .01. The initial learning rate was set to be 0.001 and we followed the 1cycle policy [4] to adjust the learning rate on the fly. We trained each model for 10 epochs and report the performance of the model on the two testing sets.

D. Amazon Mechanical Turk (AMT) Study

D.1 Study Requirements:

Each video batch (and thus each video) was published on AMT in four phases. The first two phases targeted "reliable" workers (with AMT ratings $> 95\%$, and $>$ 10,000 HITs), who helped eliminate inappropriate (violent or pornographic) content and static videos. In the latter two phases, we reduced the numbers to 75% and 1000, respectively.

As each subject was viewing the instructions, we monitored several parameters to ensure that they could effectively participate. The following eligibility criteria were imposed -

- **Browser Window Resolution:** At least 480p for mobile devices and 720p for others.
- **Browser Zoom:** Set to 100%.
- **Browsers:** Latest versions of Chrome, Firefox, Edge, Safari, and Chrome.
- **Loading Time:** Must be less than 20 secs for all the training videos.

In case they failed to meet any of the above criterion, subjects were prevented from progressing and informed accordingly. Apart from these, the subjects were also required to take a quiz reflecting their understanding of the instructions, and were allowed to proceed only if they answered at least five out of the six questions.

D.2 Interface:

The AMT interface comprised of a series of instruction pages, followed by the quiz, before they could start rating the videos. Workers were allowed to view the introductory page (Fig. 3) before accepting to participate in the study. If accepted, they had to go through the instruction pages (Fig. 4, 5, 6, 7), which were timed. During the instructions, we checked whether they satisfied the study criteria as described in Sec. D.1. Following the instruction pages, they had to pass the quiz (Fig. 8) in order to proceed to the training and testing phases. The task included rating the played video (Fig. 9) on a Likert scale [2] marked with *BAD*, *POOR*, *FAIR*, *GOOD*, and *EXCELLENT*, as demonstrated in Fig. 10. A similar interface was used for the v-patch sessions as well.

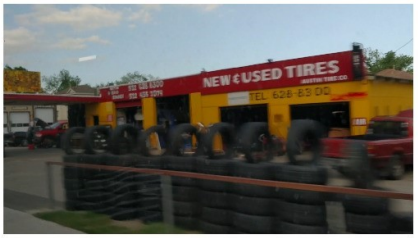
Subjective Quality Assessment of Videos

Please read the instructions carefully. You will be evaluated through a quiz after. We will be publishing this study continuously. **You can do as many HITs as you are qualified for**. So, you can skip the instructions and take the quiz here if you have done it before.

In this study, you will rate the quality of a set of videos.
Your quality ratings should reflect the quality of the videos, but not what the video is about. In other words, decide how badly the video is distorted, if at all.
For example, a well-composed but grainy, blurry or shaky video would likely be of low quality.
It is not important if the videographer did a poor job positioning people or objects in the video scene. In other words, the aesthetics are not important but the video quality is.
Here are a few example videos along with their quality opinions: **Bad**, **Poor**, **Fair**, **Good**, and **Excellent**.

BAD

POOR



FAIR

GOOD

EXCELLENT

You can proceed to the next page when the "Next" button appears.

Next

Fig. 3: **Introductory Page**

HOW TO RATE A VIDEO:

1. After each video has been played, a rating bar will appear, calibrated on a **continuous scale (0-100)** from BAD to EXCELLENT. Five pointers - "BAD," "POOR," "FAIR," "GOOD," and "EXCELLENT" are placed at equal intervals on top of the scale to guide you. The interface is as shown in the figure below.
2. Rate the video by using the mouse to move your rating to the score (position) you think best represents the quality of the video. NOTE THAT YOU MAY MOVE THE MARKER **ANYWHERE ON THE SLIDER**, **NOT ONLY** AT THE 5 POINTERS (BAD-EXCELLENT).
3. Drag the cursor along the scale and its final position will be considered as your response once you click **Submit**.
4. For every video we display, we have intentionally placed the marker at a random initial position.
5. You will not be able to submit your rating and proceed to the next video unless you have moved the cursor. Please do not give random ratings, because we will detect this and boot you from the study.
6. **Below the submit button**, you will have the option to **report** the video in case you feel the content has *nudity*, *violence* or any other *inappropriate content*. Please also report if you encounter a *static video* or a *still scene*, or if a video is *misoriented* (i.e. the video is captured vertically but oriented horizontally or vice versa). You can check the corresponding boxes to do so. This is not mandatory and you can proceed to the next video in case there is nothing to report.

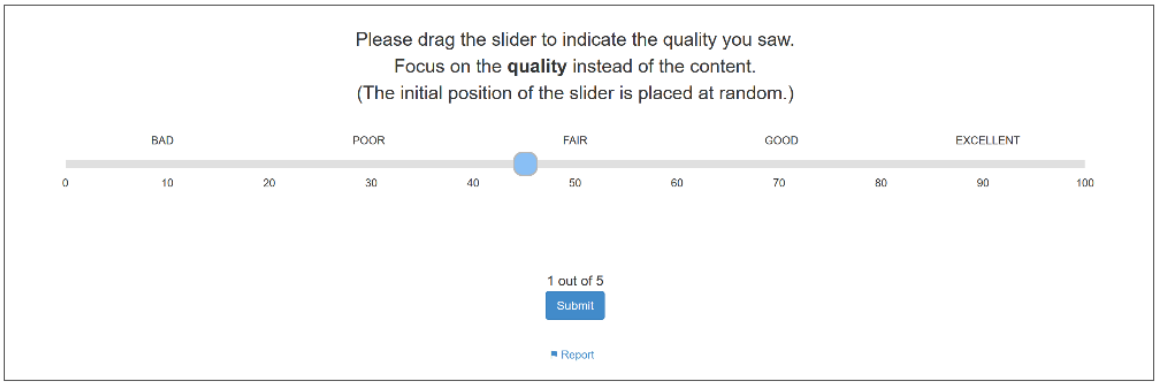


Fig. 4: **Instruction Page 1**

TRAINING AND TESTING PHASES:

The study has been divided into two phases - a **training phase** and a **testing phase**. The first few videos that you will see should help you acquaint yourself with the rating procedure and the range of qualities of videos. You'll be informed when this training phase is over and then you will move on to the testing phase.

Fig. 5: **Instruction Page 2**

ADDITIONAL INSTRUCTIONS:

- Please close any other tabs or windows that are open in your browser while participating in this study. Also, set your browser window to 100% zoom for the entire duration of the study.
- Please close all other applications that may be running on your device which may affect the browser performance.
- Please use the latest versions of any of the following browsers - Chrome, Firefox, Edge, Safari or Opera.
- Please move your chair to a comfortable viewing distance from where you can see the displayed videos.
- If you normally wear **corrective lenses** to view a monitor at this distance, **please use them during the study**, as abnormal vision will affect your perception of the video quality.
- Please switch off your mobile phone or other devices that might disturb or distract you during the experiment. Please ensure consistent network connectivity and an uninterrupted working environment. The session is continuous and cannot be paused.
- At the end of the study, you will be asked to fill in some pertinent survey questions which are integral to it.

Next

Fig. 6: **Instruction Page 3**

Ethics Policy

Thank you again for participating in our Amazon Turk study! One issue we would prefer not to bring up are Turk workers who do not take their task seriously, and instead *game* or *cheat* by trying to find ways of only appearing to do the task, to get paid without really doing the work. While most Amazon Turk workers are wonderful participants, the number of Turk workers that try to *cheat* has increased.

We therefore must tell you that we have sophisticated ways of finding whether a worker is working honestly or not. If a worker does not pass our tests, then their session will end, they will not be paid, and they will not be allowed to participate again, or in future studies!

There are other reasons why we might end your session early, e.g., if we find your set-up cannot download or play videos quickly. In those cases, we will not stop you from future studies, but we will ask you not to try the current study again.

IMPORTANT NOTE: If for some reason the video does not load, please return the HIT and contact us but DO NOT REFRESH the page

Next

Fig. 7: **Instruction Page 4**

QUIZ TIME!

The following quiz is to test your diligence and sincerity. Please choose the appropriate options:

Q1. Where can you find the rating slider?

○ **Below a video while it is playing**

○ **On the next page after the video has stopped playing**

○ **Top of the video while it is playing**

Q2. How do you rate a video using the slider?

○ **Click on the five reference positions shown above the scale**

○ **Drag the cursor along the rating scale to the appropriate position**

○ **Enter the rating value in the box below the scale**

Q3. You are evaluating each video based on its:

○ **Aesthetics (how good the video scene is framed)**

○ **Quality (how good the video looks)**

○ **Content (what is in the video)**

Q4. What kind of content, if present, do we want you to report? (Select all that apply)

□ **Low-light scenes**

□ **Still Scenes**

□ **Violence**

□ **Sports**

□ **Nudity**

Q5. How do you report a video?

○ **Return the HIT and email us immediately**

○ **Include it in the final comments at the end**

○ **Select the report option and choose accordingly**

Q6. What should you do if you normally wear corrective lens?

○ **Not wear it as that will be an interesting experiment**

○ **Wear it during the study, as not using it might affect your perception of quality**

○ **Not care as it does not matter for this study**

Submit

Fig. 8: **Quiz**

Fig. 9: **Video Playback**

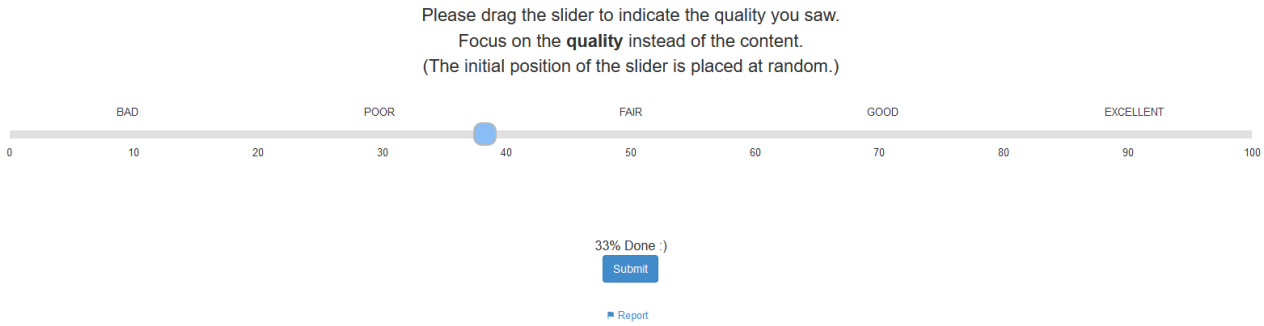


Fig. 10: **Rating Slider**

References

[1] B. Iglewicz and D. C. Hoaglin. Volume 16: How to Detect and Handle Outliers. *The ASQC Basic References in Quality Control: Statistical Techniques*, 1993.

[2] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, vol. 140, pp. 1-55, 1932.

[3] Z. Sinno and A.C. Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612-627, Feb. 2019. [Online] LIVE VQC Database: http://live.ece.utexas.edu/research/LIVEVQC/index.html.

[4] *fastai*. The 1cycle policy. [Online] Available: https://fastai1.fast.ai/callbacks.one_cycle.html#The-1cycle-policy.

[5] J. Tukey. *Exploratory data analysis*. Addison-Wesley Pub. Co, Reading, Mass, 1977.