6. Supplementary Material

The supplementary material consists of four sections. In Section 6.1, we provide qualitative results of the MDVC on another example video. The details on features extraction and implementation are described in Section 6.2 and 6.3. Finally, the comparison with other methods is shown in Section 6.4.

6.1. Qualitative Results (Another Example)

In Figure 6, we provide qualitative analysis of captioning on another video from ActivityNet Captions validation set to emphasize the importance of additional modalities for dense video captioning, namely, speech and audio. We compare the captioning proposed by MDVC (our model) conditioned on different sets of modalities: audio-only (A-only), visual-only (V-only), and including all modalities (S + A + V). Additionally, we provide the results of a captioning model proposed in Zhou *et al*. [59] (visual only) which showed the most promising results according to METEOR.

More precisely, the video (YouTube video id: `EGrXaq213Oc`) lasts two minutes and contains 12 human annotations. The video is an advertisement for snowboarding lessons for children. It shows examples of children successfully riding a snowboard on a hill and supportive adults that help them to learn. A lady narrates the video and appears in the shot a couple of times.

Generally, we may observe that MDVC with the audio modality alone (A-only) mostly describes that a woman is speaking which is correct according to the audio content yet the details about snowboarding and children are missing. This is expectedly challenging for the network as no related sound effects to snowboarding are present. In the meantime, the visual-only MDVC grasps the content well, however, misses important details like the gender of the speaker. While the multi-modal model MDVC borrows the advantages of both which results in more accurate captions. The benefits of several modalities stand out in captions for p_2 and p_{10} segments. Note that despite the appearance of the lady in the shot during p_{10}, the ground truth caption misses it yet our model manages to grasp it.

Yet, some limitations of the final model could be noticed as well. In particular, the content of some proposals is dissimilar to the generated captions, *e.g*. the color of the jacket (p_4, p_5), or when a lady is holding a snowboard with a child on it while the model predicts that she is holding a ski (p_7). Also, the impressive tricks on a snowboard were guessed simply as "ridding down a hill" which is not completely erroneous but still inaccurate (p_8). Overall, the model makes reasonable mistakes except for proposals p_3 and p_4. Finally, the generated captions provide more general description of a scene compared to the ground truth that is detailed and specific which could be a subject for future investigation.

6.2. Details on Feature Extraction

Before training, we pre-calculate the features for both audio and visual modalities. In particular, the audio features were extracted using VGGish [15] which was trained on AudioSet [12]. The input to the VGGish model is a 96×64 log mel-scaled spectrogram extracted for non-overlapping 0.96 seconds segments. The log mel-scaled spectrogram is obtained by applying *Short-Time Fourier Transform* on a 16 kHz mono audio track using a periodic *Hann* window with 25 ms length with 10 ms overlap. The output is a 128-d feature vector after an activation function and extracted before a classification layer. Therefore, the input to MDVC is a matrix with dimension $T_j^a \times 128$ where T_j^a is the number of features proposal p_j consists of.

The visual features were extracted using I3D [6] network which inputs a set of 24 RGB and optical flow frames extracted at 25 fps. The optical flow is extracted with PWC-Net [41]. First, each frame is resized such that the shortest side is 256 pixels. Then, the center region is cropped to obtain 224×224 frames. Both RGB and flow stacks are passed through the corresponding branch of I3D. The output of each branch are summed together producing 1024-d features for each stack of 24 frames. Hence, the resulting matrix has the shape: $T_j^v \times 1024$, where T_j^v is the number of features required for a proposal p_j.

We use 24 frames for I3D input to temporally match with the input of the audio modality as $\frac{24}{25} = 0.96$. Also note that I3D was pre-trained on the Kinetics dataset with inputs of 64 frames, while we use 24 frames. This is a valid approach since we employ the output of the second to the last layer after activation and average it on the temporal axis.

The input for speech modality is represented by temporally allocated text segments in the English language (one could think of them as subtitles). For a proposal p_j, we pick all segments that both: a) end after the proposal starting point, *and* b) start before the proposal ending point. This provides us with sufficient coverage of *what has been said* during the proposal segment. Similarly to captions, each word in a speech segment is represented as a number which corresponds to the word's order number in the vocabulary and then passed through the text embedding of size 512. We omit the subtitles that describe the sound like "[Applause]" and "[Music]" as we are only interested in the effect of the speech. Therefore, the speech transformer encoder inputs matrices of shape: $T_j^s \times 512$ where T_j^s is the number of words in corresponding speech for proposal p_j.

6.3. Implementation Details

Since no intermediate layers connecting the features and transformers are used, the dimension of the features transformers D_T corresponds to the size of the extracted features: 512, 128, and 1024 for speech, audio, and visual modalities, respectively. Each feature transformer has one

Method	METEOR
Seen full dataset	
Xiong [53] (RL)	7.08
Mun *et al*. [30] (RL)	8.82
Mun *et al*. [30] (without RL)	6.92
Seen part of the dataset	
MDVC, no missings	7.31

Table 4. The comparison with other dense video captioning methods on ActivityNet Captions validation set estimated with METEOR. The results are presented for the learned proposals.

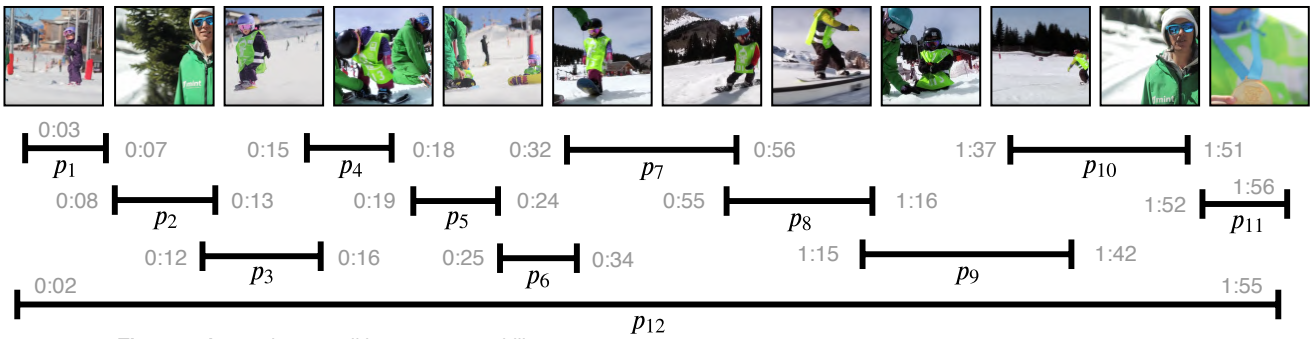
layer (L), while the internal layer in the position-wise fully-connected network has $D_P = 2048$ units for all modality transformers which was found to perform optimally. We use $H = 4$ heads in all multi-headed attention blocks. The captions and speech vocabulary sizes are 10,172 and 23,043, respectively.

In all experiments, except for the audio-only model, we use *Adam* optimizer [21], a batch containing features for 28 proposals, learning rate 10^{-5}, $\beta = (0.9, 0.99)$, smoothing parameter $\gamma = 0.7$. In the audio-only model, we apply two-layered transformer architecture with learning rate 10^{-4} and $\gamma = 0.2$. To regularize the weights of the model, in every experiment, *Dropout* [40] with $p = 0.1$ is applied to the outputs of positional encoding, in every sub-layer before adding a residual, and after the first internal layer of the multi-modal generator.

During the experimentation, models were trained for 200 epochs at most and stopped the training early if for 50 consecutive epochs the average METEOR score calculated on ground truth event proposals of both validation sets has not improved. At the end of the training, we employ the best model to estimate its performance on the learned temporal proposals. Usually the training for the best models culminated by 50^{th} epoch, *e.g*. the final model (MDVC (S + A + V)) was trained for 30 epochs which took, roughly, 15 hours on *one* consumer-type GPU (Nvidia GeForce RTX 2080 Ti). The code for training heavily relies on PyTorch framework and will be released upon publication.

6.4. Comparison with Other Methods

In Tab. 4, we present a comparison with another body of methods [53, 30] which were not included in the main comparison as they were using *Reinforcement Learning* (RL) approach to directly optimize the non-differentiable metric (METEOR). We believe that our method could also benefit from these as the ablation studies in [53, 30] show significant improvement. As it was anticipated, in general, methods which employ reinforcement learning perform better in terms of METEOR. Interestingly, our model still outperforms [53] which uses RL in the captioning module.



p_1
Zhou *et al.*: people are walking on a snowy hill
MDVC (A-only): A man is seen walking into a room and begins to walk around the area
MDVC (V-only): A man in a red shirt is standing on a snowy hill
MDVC (S + A + V): A man is seen standing on a field with a stick in his hands
Ground Truth: The small children are going on snowboarding lifts to begin their lessons

p_2
Zhou *et al.*: a man in a jacket is standing in a snow covered hill
MDVC (A-only): A woman is seen standing on a field and begins to talk to the camera
MDVC (V-only): A man in a red jacket is talking to the camera
MDVC (S + A + V): A woman is talking to the camera
Ground Truth: A woman dressed in a green winter coat and white woolen hat is talking about the snowboarding school

p_3
Zhou *et al.*: a man in a jacket is standing in a snowboard
MDVC (A-only): A woman is seen standing on a field and begins to talk to the camera
MDVC (V-only): A man in a red jacket is talking to the camera
MDVC (S + A + V): A woman is talking to the camera
Ground Truth: There's a small boy snowboarding down the slope

p_4
Zhou *et al.*: a man in a jacket is snowboarding down a hill
MDVC (A-only): A woman is seen standing in a room and begins to talk to the camera
MDVC (V-only): A man in a red jacket is standing in a snow covered with snow
MDVC (S + A + V): The man in the orange jacket is talking to the camera
Ground Truth: There are some adults helping small children get ready with their snowboarding equipment and gear

p_5
Zhou *et al.*: the man in a blue jacket is snowboarding down the hill
MDVC (A-only): A woman is seen speaking to the camera and leads into her holding a stick
MDVC (V-only): A man in a red jacket is snowboarding
MDVC (S + A + V): The man in the orange jacket is snowboarding down the hill
Ground Truth: The young snowboarder continues to go down the mountain slope steadily

p_6
Zhou *et al.*: a man in a jacket is standing on a snowboard
MDVC (A-only): A woman is seen speaking to the camera and leads into her holding a stick and speaking to the camera
MDVC (V-only): A man in a red jacket is standing on a snowy hill
MDVC (S + A + V): A woman is seen speaking to the camera while holding a camera and leads into her riding down
Ground Truth: An instructor from the snowboarding school is helping some young children get up after falling down

p_7
Zhou *et al.*: the man then rides down the hill on a snowboard followed by several people climbing up the
MDVC (A-only): The woman continues to talk and the camera pans around the yard
MDVC (V-only): The people are seen walking around a snowy hill and looking off into the distance
MDVC (S + A + V): A woman is seen speaking to the camera while holding up a ski and leads into them riding down a hill
Ground Truth: The instructors are helping and teaching the young students to begin snowboarding and teaching them various steps

p_8
Zhou *et al.*: the man then rides down a hill on a snowboard
MDVC (A-only): She then shows how to use the machine to cut the grass
MDVC (V-only): The man in the orange jacket is snowboarding
MDVC (S + A + V): The man is snowboarding down a hill
Ground Truth: One of the students does impressive stunts and moves while snowboarding

p_9
Zhou *et al.*: the man continues riding down the hill and ends with her skiing
MDVC (A-only): She then begins to talk to the camera and shows off the camera
MDVC (V-only): The people continue skiing down the slope and ends with a person walking down the slope
MDVC (S + A + V): A woman is seen speaking to the camera and leads into a person riding down a hill
Ground Truth: The instructors continue to teach their students

p_{10}
Zhou *et al.*: the man in a red jacket is talking to the camera
MDVC (A-only): A woman is seen speaking to the camera and leads into her holding a stick and speaking to the camera
MDVC (V-only): A man in a red jacket is talking to the camera
MDVC (S + A + V): A woman is seen speaking to the camera while holding up a camera and leads into a man riding down a hill
Ground Truth: The young snowboarders show off their talent as they go down the slopes

p_{11}
Zhou *et al.*: a woman in a jacket is sitting on a chair
MDVC (A-only): The man then puts his face on the ground
MDVC (V-only): A woman is seen speaking to the camera and leads into her holding up a bottle of water
MDVC (S + A + V): The camera pans to the camera and shows the camera
Ground Truth: One of the students shows the medal that he has won in a snowboarding competition.

p_{12}
Zhou *et al.*: people are snowboarding down a hill
MDVC (A-only): A woman is seen speaking to the camera while holding a stick and leads into her brushing her teeth
MDVC (V-only): A man is seen speaking to the camera and leads into several clips of people riding down a snowy hill
MDVC (S + A + V): A woman is snowboarding down a hill
Ground Truth: There are many young children in a snowboarding school

Figure 6. Another example of the qualitative results for a video in the validation set. In the video, a lady is shown speaking twice (in p_2 and p_{10}). Since MDVC is conditioned not only on visual (V) but also speech (S) and audio (A) modalities, it managed to hallucinate a caption containing a "woman" instead of a "man". We invite a reader to watch it on YouTube for a better impression (`EGrXaq213Oc`). Note: the frame size mimics the MDVC input; the scale of temporal segments is not precise. Best viewed in color.