Given a set of unlabelled videos, visual frames $F = \{f_1, \ldots, f_{N_f}\}$ are extracted.

Speech utterances $U = \{u_1, \ldots, u_{N_u}\}$ are aligned with F.

The immediate future utterance $W = \{w_1, \ldots, w_{N_w}\}$ where u_i and w_j are tokenized words in the transcribed utterances.

Utterances refer to a single sentence of transcribed speech.

Forward Generation: The model is trained to generate W given F and U using the following loss function.

$$\mathcal{L}_{FG} = -\sum_{i=1}^{N_w} \log P(w_i | w_1, \ldots, w_{i-1}, F, U)$$

This loss encourages the pretrained model to effectively encode temporally aligned multimodal inputs to predict the future utterance.

Backward Generation: The same loss is applied to generate U given F and W.

$$\mathcal{L}_{BG} = -\sum_{i=1}^{N_u} \log P(u_i | u_1, \ldots, u_{i-1}, F, W)$$

This loss encourages the network to generate a caption related to the visual contents.

The model is also trained with a masked language modeling loss (MLM) $\mathcal{L}_{MLM}(X)$, where X is the input utterance on which the masking is applied.

MLM loss is applied on both forward and backward input utterances $\mathcal{L}_{MLM}(U)$, and $\mathcal{L}_{MLM}(W)$, and are computed independently from bidirectional generation loss.

Given a multimodal video input consisting of F and text inputs $X = \{x_1, \ldots, x_{N_x}\}$. The model extracts features from the individual modalities independently.

When computing forward generation loss, X is set to temporally aligned U, and for computing backward generation loss, X is set to W.

Textual Encoder: The model extracts N_x contextualized textual embeddings $E = \{e_i\}$ from the input text using a BERT encoder.

Visual Encoder: Visual features are extracted directly from pixels using the transformer-based encoder ViViT, giving $T+1$ visual features $V = v_j$, where T is the number of tokens in the temporal dimension.

Given E, V, the model's multimodal encoder fuses the multimodal information using a co-attentional transformer, resulting in output multimodal features $\hat{E}, \hat{V}$.

Then given multimodal video features $C = \hat{E} \cup \hat{V}$ as context, the model autoregressively generates the output sentence Y using a transformer decoder.

Token y_i is then generated by encoding the previous $Y_i = \{y_0, \ldots, y_{i-1}\}$ tokens using a look-up table and positional embedding to produce $H_i = \{h_0, \ldots, h_{i-1}\}$.

Then using a single transformer, the context C and previous embedded tokens H_i are encoded, outputting $\tilde{C} \cup \tilde{H}_i$, where $\tilde{H}_i = \{\tilde{h}_0, \ldots, \tilde{h}_{i-1}\}$.

The next token y_i is predicted from $\tilde{h}_{i-1}$ using a linear projection with a softmax:

$$y_i = \text{argmax}(\text{softmax}(\Phi \tilde{h}_{i-1}))$$

where $\Phi \in \mathbb{R}^{vxd}$ is the linear projection matrix, and v is the vocabulary size.