Supplemental Materials for Show Me What and Tell Me How: Video Synthesis via Multimodal Conditioning

Ligong Han[1][2]* Jian Ren[1] Hsin-Ying Lee[1] Francesco Barbieri[1]
Kyle Olszewski[1] Shervin Minaee[1] Dimitris Metaxas[2] Sergey Tulyakov[1]
[1]Snap Inc. [2]Rutgers University

Content of Supplementary Materials

S1. More Details and Ablation for Methods

In this section, we introduce additional details of our methods. Specifically, we describe the settings for the masking strategies for Masked Sequence Modeling (MSM) in Sec. S1.1, different training methods for Relevance Estimation (REL) task in Sec. S1.2, augmentation performed on the task of Video consistency modeling (VID) in Sec. S1.3, additional discussion on improved mask-predict for video prediction in Sec. S1.4, and an ablation analysis on text augmentation in Sec. S1.5.

S1.1. Settings for Masking Strategies in MSM Task

In the main paper (Sec. 3.1), we introduce five masking strategies, *i.e.*, (I) i.i.d. masking; (II) masking all tokens; (III) block masking; (IV) the negation of block masking; and (V) randomly keeping some frames, to train the task of mask sequence modeling. In all of our experiments, if not specified, we apply strategies I - IV with probabilities as $[0.7, 0.1, 0.1, 0.1]$. For strategy V, we adopt it by randomly keeping k frames on top of the mask produced from strategies I - IV. We set the probability of strategy V as 0.2 and $k = T/2$, where T is the total number of frames.

S1.2. Training Methods for REL Task

We compare two training methods for the relevance estimation task. The first one is swapping the conditional inputs to get the negative sample, which we denote as `REL`=swap. The method is introduce in Sec. 3.1 of the main paper. The second method, `REL`=negative, is to sample a negative training data such that it has a different annotation as the positive one. This ensures that the negative sequence for REL is indeed negative, which is not guaranteed in the case of conditional swapping. As shown in Tab. 5, we empirically find that negative sampling achieves better performance than conditional swapping in the early stage but its FVD and F_8 score becomes inferior when the model converges. Thus, `REL`=swap is used in all experiments if not specified.

[*] Work done during an internship at Snap Inc.

Table 5. Analysis of the use of different training methods for the video relevance task on the Multimodal VoxCeleb dataset. Results from two iterations (50K and 100K) are reported for each method.

Resolution	Method	Iteration	FVD ↓	F_8 ↑	$F_{1/8}$ ↑
128×128	`REL`=swap	50K	123.147	0.921	0.888
		100K	**103.622**	**0.936**	0.895
	`REL`=negative	50K	109.471	0.931	0.903
		100K	117.128	0.922	**0.922**
256×256	`REL`=swap	50K	293.999	0.753	0.692
		100K	**191.910**	**0.781**	0.788
	`REL`=negative	50K	225.043	0.648	0.651
		100K	201.702	0.774	**0.864**

S1.3. Video Augmentation on VID Task

We propose a `VID` token to for modeling video consistency (Sec. 3.2 in the main paper). To learn the `VID` in a self-supervised way, we introduce four negative video augmentation methods. Here we illustrate more details for each augmentation strategy, shown in Fig. 7, including color jittering, affine transform, frame swapping, and frame shuffling. In all of our experiments, if not specified, we uniformly sample these strategies with probabilities as $[0.25, 0.25, 0.25, 0.25]$.

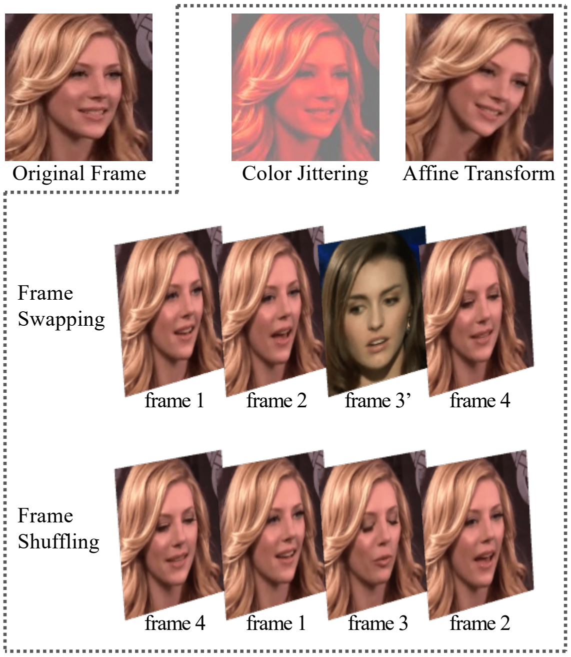


Figure 7. **Augmentation strategies for modeling video consistency.** *Top Row*: first column – original frame; second column – augmented with color jittering; third column – augmented with affine transform. *Second Row*: frame swapping such that the third frame is swapped by using a frame from another video. *Third Row*: frame shuffling such that the position of frames is randomly shuffled.

S1.4. More Details on Improved Mask-Predict

`SampleToken` and `SampleMask`. We introduce our algorithm for improved mask-predict in the main paper (Alg. 1). Here we provide more details of the two functions (`SampleToken` and `SampleMask`) used in the algorithm.

- `SampleToken` is given in Alg. 2, with PyTorch [8]-like functions. $\texttt{Gather}(\mathbf{p}, \mathbf{z})$ gathers values of $\mathbf{p}$, which is a matrix whose dimensions are the number of tokens by the number of words, along the token axis specified by indices $\mathbf{z}$.
- `SampleMask` is given in Alg. 3. The function `Find` collects the indices of the `True` elements; function $\texttt{Multinomial}(\mathbf{y}, n)$ samples n points without replacement from a multinomial specified by $\mathbf{y}$ and returns their indices; and function $\texttt{Scatter}(\mathbf{0}, \mathbf{j}, 1)$ sets values to 1 in a tensor initialized to $\mathbf{0}$ at locations specified by indices $\mathbf{j}$. Lines 1 - 5 in Alg. 3 sample n locations, according to $\mathbf{y}$, to be preserved, and the locations with $\mathbf{m}_{\text{PC}}$ equal to 1 are always selected.

Algorithm 2 `SampleToken`

```
Require: Logit $\tilde{\mathbf{p}}$ and noise level $\sigma$.
1: $\mathbf{g} \leftarrow \texttt{Gumbel}(0, 1)$ $i.i.d.$
2: $\mathbf{p} \leftarrow \texttt{Softmax}(\tilde{\mathbf{p}} + \sigma\mathbf{g})$
3: $\mathbf{z} \leftarrow \texttt{Multinomial}(\mathbf{p})$    ▷ sample from multinomial
4: $\mathbf{y} \leftarrow \texttt{Gather}(\mathbf{p}, \mathbf{z})$    ▷ collect probs for each token
5: return $\mathbf{z}$, $\mathbf{y}$
```

Algorithm 3 `SampleMask`

```
Require: Probabilities $\mathbf{y}$, preservation mask $\mathbf{m}_{\text{PC}}$, and the
    number of tokens to keep $n$.
1: $\mathbf{y}' \leftarrow \mathbf{y}\,[\mathbf{m} == 0]$    ▷ collect probs no need to preserve
2: $\mathbf{i}' \leftarrow \texttt{Find}(\mathbf{m} == 0)$    ▷ collect indices
3: $\mathbf{i} \leftarrow \texttt{Multinomial}(\text{Normalize}(\mathbf{y}'), n)$
4: $\mathbf{j} \leftarrow \mathbf{i}'\,[\mathbf{i}]$    ▷ slicing to get sampled indices
5: $\mathbf{m} \leftarrow \texttt{Scatter}(\mathbf{0}, \mathbf{j}, 1)$    ▷ populate indices
6: $\mathbf{m} \leftarrow \mathbf{m} \,|\, \mathbf{m}_{\text{PC}}$    ▷ elementwise OR
7: return $\mathbf{m}$
```

Mask Annealing. We define the piecewise linear mask annealing scheme $n^{(i)}$ (used in Alg. 1 in the main paper) as follows.

$$n^{(i)} = \begin{cases} N \cdot (\beta_1 + \frac{L_1 - i}{L_1 - 1} \cdot (\alpha_1 - \beta_1)) & \text{for} \quad 1 \leq i \leq L_1 \\ N \cdot \alpha_2 & \text{for} \quad L_1 < i \leq L_1 + L_2 \\ N \cdot \alpha_3 & \text{for} \quad i > L_1 + L_2 \end{cases} \tag{4}$$

where we set $L_1 = 10$, $L_2 = 10$, $\alpha_1 = 0.9$, $\beta_1 = 0.1$, $\alpha_2 = 0.125$, and $\alpha_3 = 0.0625$. We use the following values

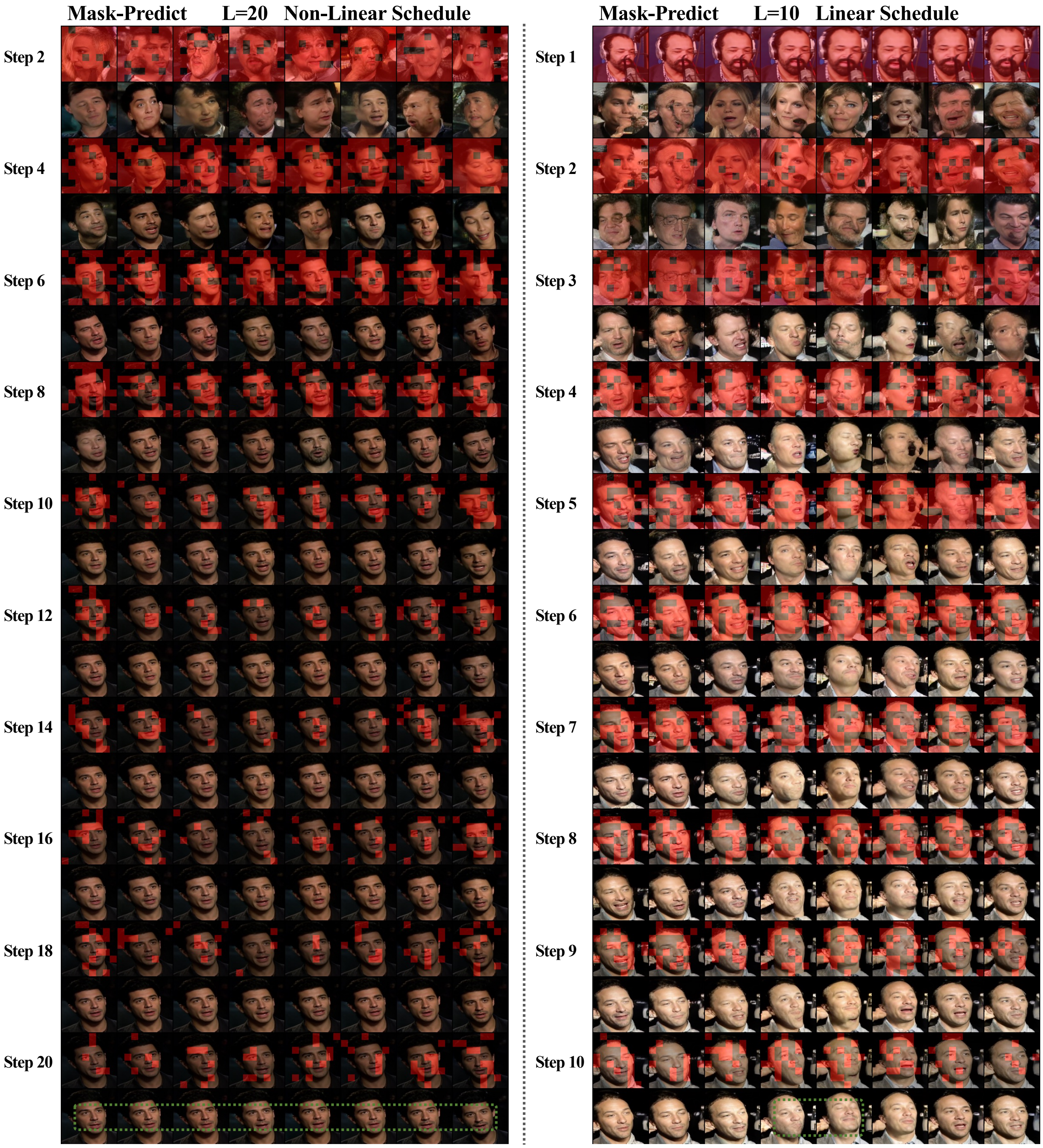


Figure 8. **Comparison between non-linear (ours) *vs.* linear schedule for mask annealing in mask-predict**. The mask-predict starts from a fully-masked sequence (Step 1, and the images displayed beneath red masks in Step 1 are real video frames). Patches with red color denote the corresponded tokens are masked. The images with red color are generated after the mask-predict at that step. *Left*: frames generated using our non-linear mask annealing scheme. The motion is vivid and frames have high quality (highlighted in dotted green box). *Right*: using a linear scheme ($L = L_1 = 10, \alpha_1 = 0.9, \beta_1 = 0.1$) to generate frames. Artifacts can be observed on the synthesized images (highlighted in dotted green box). Frames are synthesized from the model trained on the Multimodal VoxCeleb dataset.

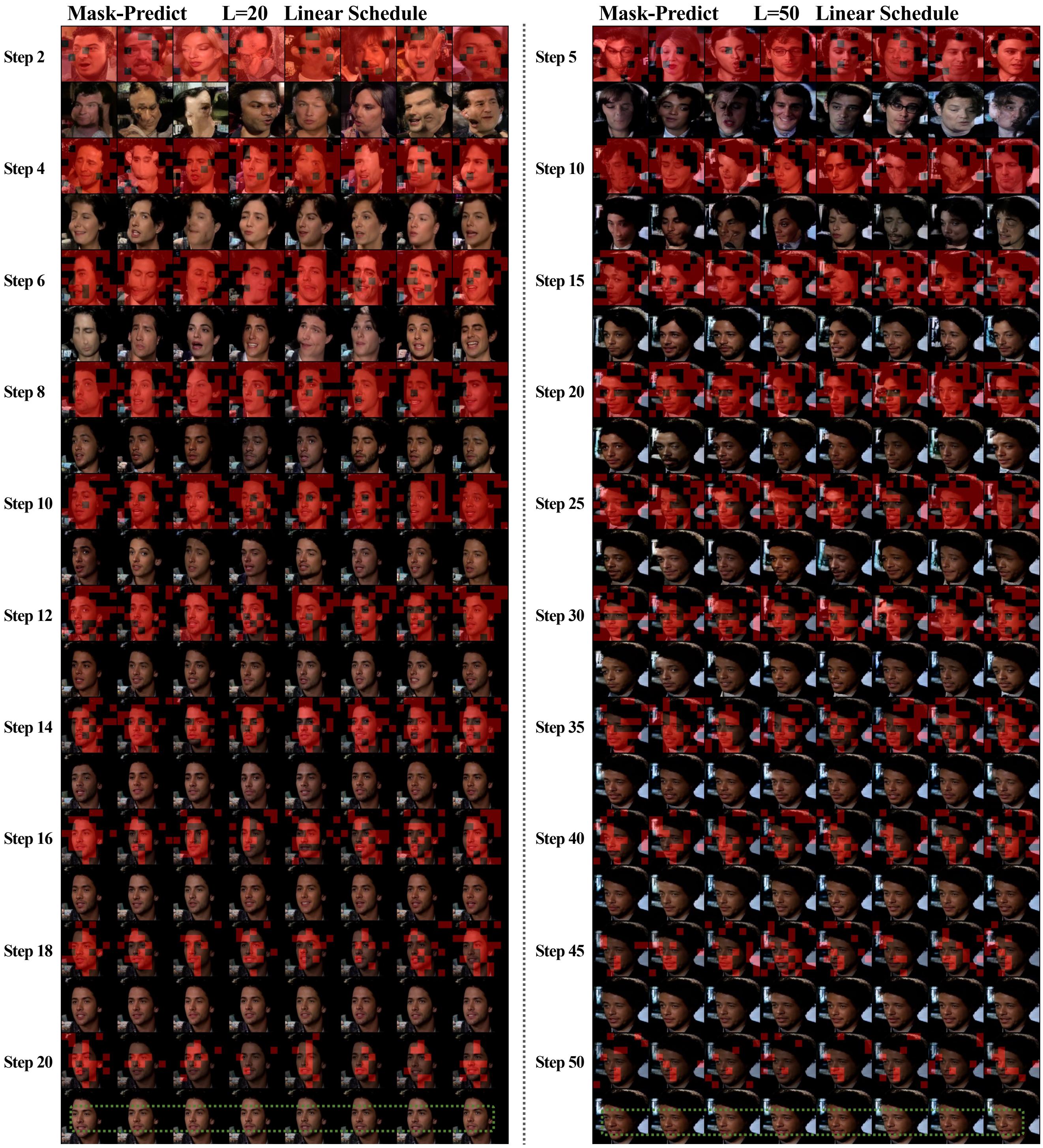
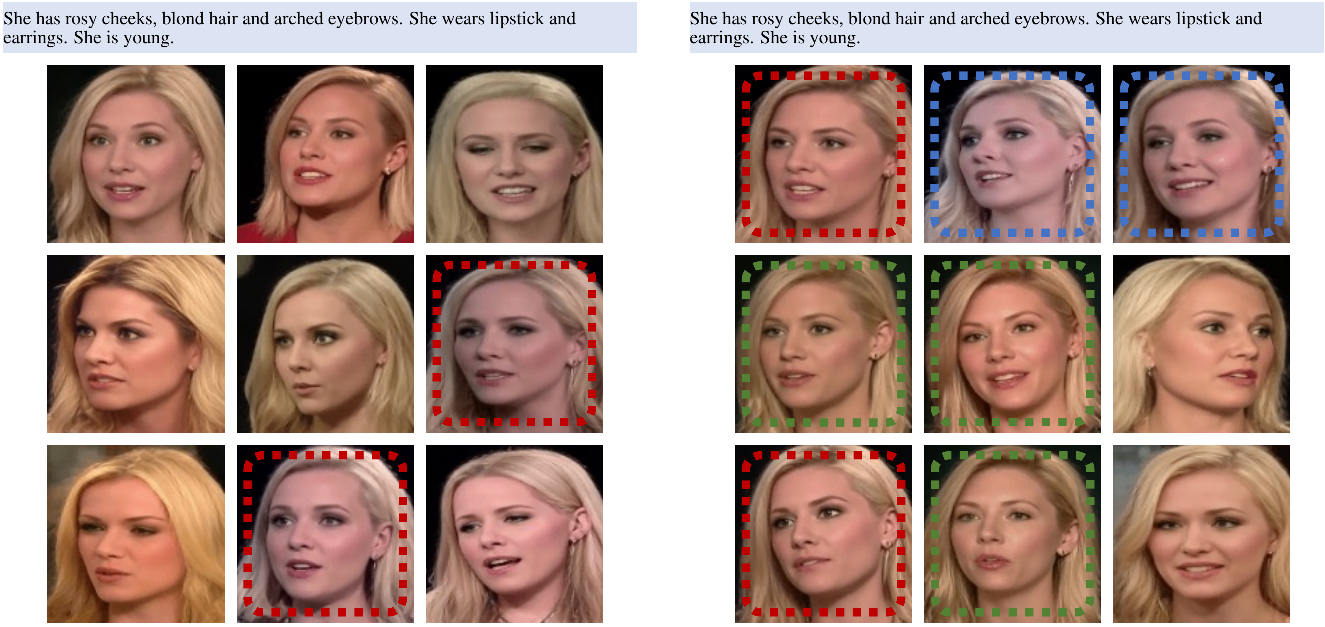


Figure 9. **Frames synthesized by using linear schedule for mask annealing in mask-predict**. The mask-predict starts from a fully-masked sequence (Step 1, and the images displayed beneath red masks in Step 1 are real video frames). Patches with red color denote the corresponded tokens are masked. The images with red color are generated after the mask-predict at that step. Two samples have the setting as $\alpha_1 = 0.9$ and $\beta_1 = 0.1$. Motion has been washed out, *i.e.*, frames in a sequence tends to be static and have similar appearance as illustrated in dotted green box, for the setting of $L = L_1 = 20$ (*Left*) and $L = L_1 = 50$ (*Right*). Frames are synthesized from the model trained on the Multimodal VoxCeleb dataset.


She has rosy cheeks, blond hair and arched eyebrows. She wears lipstick and earrings. She is young.
She has rosy cheeks, blond hair and arched eyebrows. She wears lipstick and earrings. She is young.

(a) *W/* noise annealing in mask-predict.

(b) *W/O* noise annealing in mask-predict.

Figure 10. **Comparison between *w/* (a) and *w/o* (b) noise annealing in mask-predict for text-to-video generation**. Each image in a subfigure is the first frame sampled from a synthesized video and each subfigure includes 9 videos. For the two subfigures, we use the same textual input to generate videos and apply dotted boxes with the same color to denote the synthesized videos with the same (or very similar) identity. Adding noise improves diversity as (a) only contains two images with the same (or very similar) identity. Frames are generated from a model trained on the Multimodal VoxCeleb dataset. We use linear mask annealing scheme with $L = 15$.

in experiments if not specified: $L_1 = 10, L_2 = 10, \alpha_1 = 0.9, \beta_1 = 0.1, \alpha_2 = 0.125, \alpha_3 = 0.0625$. The total step of mask-predict is L so that $L_1 + L_2 \leq L$.

Compared with linear annealing, our non-learning mask annealing can generate videos with vivid motion and less artifacts. Example samples for models trained on the Multimodal VoxCeleb dataset are illustrated in Fig. 8 and Fig. 9. Our method generates facial videos with high fidelity for L as 20 (Fig. 8 Left), while linear annealing generates low quality frames (Fig. 8, right) and static videos where motion can hardly be observed (Fig. 9).

Noise Annealing. We define the noise annealing schedule $\sigma^{(i)}$ as follows:

$$\sigma^{(i)} = \begin{cases} \beta_1 + \frac{L_1 - i}{L_1 - 1} \cdot (\alpha_1 - \beta_1) & \text{for} \quad 1 \leq i \leq L_1 \\ \alpha_2 & \text{for} \quad L_1 < i \leq L_1 + L_2 \\ \alpha_3 & \text{for} \quad i > L_1 + L_2 \end{cases} \tag{5}$$

where L_1, α_1, β_1 are reused from Eqn. 4 for simplicity of notation, but with different values. We set $L_1 = 10, L_2 = 5, \alpha_1 = 0.4, \beta_1 = 0.02, \alpha_2 = 0.01, \alpha_3 = 0$.

Adding noise improves diversity for generated videos, as shown in Fig. 10. However, there is a tradeoff between diversity and quality. Adding too much noise influences sample quality, which might be due to unconfident tokens that cannot be remasked.

Beam Search. We analysize different numbers of beams B employed in the beam search that is used in mask-predict. Results shown in Tab. 6 show that using $B = 15$ achieves the best results. Interestingly, we empirically find that increasing B from 15 to 20 causes performance drop. We hypothesize that this is due to the scores used for beam selection is not accurate. When B gets larger, the negative influence of inaccurate score estimation becomes more prominent. We use $B = 3$ in all experiments if not specified.

Early-Stop. Early-stop is proposed in previous text-to-image generation [11] to stop the mask-predict at the earlier iteration for faster inference. Here we analyze the use of early-stop in our work, and determine that it cannot improve the efficiency. We obtain the scores $S_{\texttt{REL}}$ and $S_{\texttt{VID}}$, calculated from two special tokens `RED` and `VID`, respectively. We denote S_{avg} as their averaged score and use S_{avg} to decide the iteration for stopping if the highest score does not change for 3 iterations. We first show the quantitative results in Tab. 6, where we find that early-stop does not improve the FVD at $B = 1,\ 3,\ 5$. We further provide visual images in Fig. 11. We can see that $S_{\texttt{REL}}$ is very high at the beginning, and peaks at step 3, $S_{\texttt{VID}}$ peaks at step 15, and the average score S_{avg} reaches the highest value at step 9. However, we can still observe artifacts at each step (3, 9, and 15). Therefore, using scores calculated from special tokens might not be a reliable signal for determining early-stop, and we thus decide not to use it in our implementation.

Table 6. **Analysis on Beam Searching and Early-Stop**. Metrics are evaluated on models trained on the Multimodal VoxCeleb dataset, with the different number of beams B and whether early-stop is enabled. The task is text-to-video generation.

B	Early-Stop	FVD ↓	F_8 ↑	$F_{1/8}$ ↑
1	✗	97.992	0.939	0.930
1	✓	97.957	0.917	0.928
3	✗	96.288	0.945	0.925
3	✓	99.899	0.930	0.929
5	✗	94.170	0.922	0.937
5	✓	97.908	0.923	0.925
10	✗	95.560	0.932	0.924
15	✗	92.828	0.933	0.933
20	✗	97.247	0.922	0.918

Table 7. Human preference evaluation for different methods on the Multimodal VoxCeleb dataset. The task is text-to-video generation.

Methods for Pairwise Comparison			Human Preference
MMVID	*vs.*	ART-V	**54.0%** : 46.0%
MMVID-TA	*vs.*	MMVID	**54.5%** : 45.5%
MMVID-TA	*vs.*	ART-V	**61.2%** : 38.8%

S1.5. Analysis on Text Augmentation

Sec. 3.4 of the main paper introduces text augmentation to improve the correlation between the generated videos and input textual controls. We also notice that text augmentation can help improve the diversity of the synthesized videos. We perform human evaluation using Amazon Mechanical Turk (AMT) to verify the quality and diversity of videos synthesized from various methods. We consider three comparisons, including *MMVID*, which is our baseline model, *MMVID-TA*, which uses text augmentation, and *ART-V*, which uses the autoregressive transformer. 600 synthesized videos on the Multimodal VoxCeleb dataset for the text-to-video generation task are presented to AMT, and the results are shown in Tab. 7. We can see using text augmentation can help improve the quality and diversity of the generated videos, as 61.2% users prefer the MMVID-TA over ART-V.

S2. More Experimental Details

In this section, we introduce more implementation details in experiments and additional experimental results.

S2.1. More Implementation Details

Training of Autoencoder. For each dataset at each resolution, we finetune an autoencoder from VQGAN model [4] pretrained on ImageNet [9], with $f = 16$ (which is the equivalent patch-size a single code corresponds to) and $|\mathcal{Z}| = 1024$ (which is the vocabulary size of the codebook).
Evaluation Metrics. To evaluate the model performance on the Shapes dataset, we train a classifier following the instructions of TFGAN [2] as the original model is not released. To have a fair comparison, we also retrain a TFGAN model for text-to-video generation.

S2.2. Dataset Statistics and Textual Controls

Shapes. For text-to-video and independent text-visual control experiments, we use the text descriptions provided in the original Moving Shapes dataset [3]. The texts are generated using a template such as *"A ⟨object⟩ is moving in ⟨motion⟩ path towards ⟨direction⟩"* or *"A ⟨object⟩ is moving in ⟨motion⟩ path in the ⟨direction⟩ direction"*. More details can be found in TFGAN [3].
MUG. The original MUG Facial Expression dataset [1] does not provide text descriptions for videos. To have a fair comparison, we follow the examples in TiVGAN [5] and manually label genders for all subjects, and generate corresponding text for each video given annotations. For example, given a video with annotations as "female" and "happiness", we generate the description as *"A women/young women/girl is making a happiness face"* or *"A women/young women/girl is performing a happiness expression"*. We randomly choose a word to describe gender from *"women", "young women"* and *"girl"*.
iPER. The iPER [6] dataset contains 30 subjects wearing 103 different clothes in total, resulting in 206 videos (every cloth is unique in appearance and has both an A-pose and a random pose recording). To test the generalization capability of the generation models to unseen motions, we split a held-out set of 10 videos which contains 10 unique appearances performing an A-pose. We further cut all videos into 100-frame clips and perform training and evaluation on these clips. The held-out 10 videos contain 93 clips. Quantitative metrics are evaluated on these 93 clips plus the same set of appearance performing a random pose (186 clips in total). Similar to MUG dataset, the texts are generated using a template such as *"Person ⟨person_ID⟩ dressed in ⟨cloth_ID⟩ is performing ⟨pose⟩ pose"*.
Multimodal VoxCeleb. We generate textual descriptions from annotated attributes for the Multimodal VoxCeleb dataset following previous work [10], especially this webpage[1]. The attribute combinations labeled from videos of Multimodal VoxCeleb shows a long-tail distribution

[1] https://github.com/IIGROUP/Multi-Modal-CelebA-HQ-Dataset/issues/3

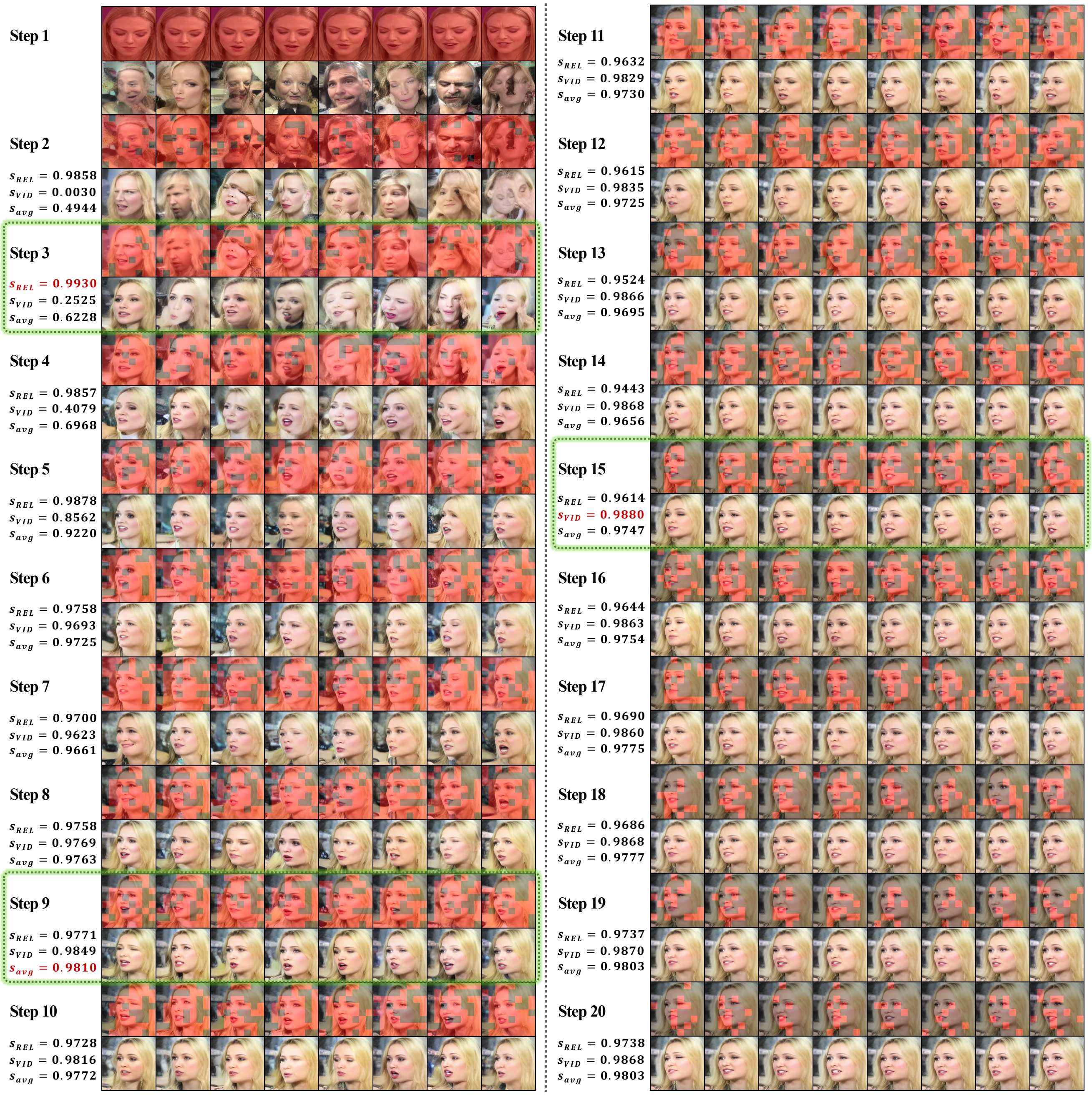


Figure 11. **Visual samples for analyzing early-stop.** REL score S_{REL} is very high at the beginning and peaks at step 3. VID score S_{VID} peaks at step 15. The average score S_{avg} reaches the highest value at step 9. Step 1 shows mask-predict starts from a fully-masked sequence and the images displayed beneath red masks in Step 1 are real video frames. $B = 1$ is used.

(Fig. 12). There are $13,706$ unique attribute combinations out of $19,522$ samples, and $11,259$ combinations have only one data point. This motivates us to use text dropout during training as we encourage the model not to memorize certain attribute combinations with one single data point.

S3. More Generated Videos

In this section, we provide more generated videos by our approach and other works. The thumbnail from each video is shown in the figures. Videos are also attached in the submitted file. *We provide an HTML page to visualize synthesized videos.*

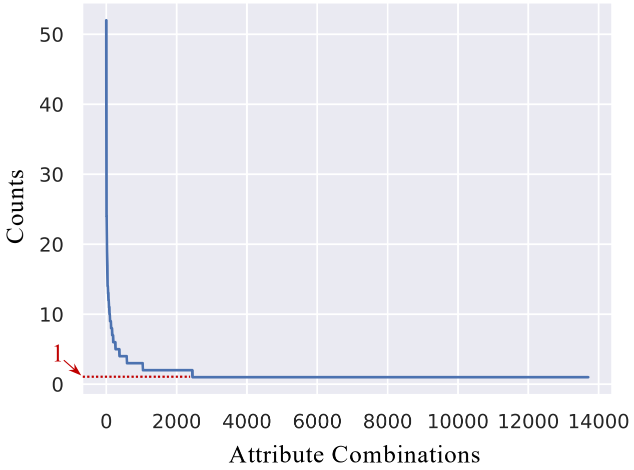


Figure 12. **Statistics of annotations for the Multimodal VoxCeleb dataset**. The attribute combinations show a long-tail distribution. There are $13,706$ unique attribute combinations out of $19,522$ samples, and $11,259$ combinations have only one data point.

Results on the Shapes dataset. We provide more results on the Shapes dataset.

- Fig. 13 shows the videos generated by our approach for the task of text-to-video generation.
- Fig. 14 shows the videos generated by our approach for the task of independent multimodal generation. The input control signals are text and a partially observed image (with the center masked out).
- Fig. 15 shows the videos generated by our approach for the task of dependent multimodal generation. The input control signals are text and image.

Results on the MUG dataset. Fig. 16 shows the video generated by our approach for the task of text-to-video generation.

Results on the iPER dataset. Fig. 17 shows the video generated by our approach for the task of text-to-video generation. We demonstrate long sequence generation in Fig. 17 by performing extrapolation. The process is repeated for each sequence 100 times, resulting in a 107-frame video. The textual input also controls the speed, where “slow” indicates videos with slow speed such that the motion is slow, while “fast” indicates the performed motion is fast.

Results on the Multimodal VoxCeleb dataset. We provide more results for models trained on the Multimodal VoxCeleb dataset.

- Fig. 18 shows the videos generated by our approach for the task of text-to-video generation.
- Fig. 19 shows the videos generated by our approach for the task of independent multimodal generation. The input control signals are text and a segmentation mask.
- Fig. 20 shows the videos generated by our approach for the task of independent multimodal generation. The input control signals are text and an artistic drawing.
- Fig. 21 shows the videos generated by our approach for the task of independent multimodal generation. The input control signals are text and a partially observed image.
- Fig. 22 shows the videos generated by our approach for the task of dependent multimodal generation. The input control signals are text, an image, and a segmentation mask.
- Fig. 23 shows the videos generated by our approach for the task of dependent multimodal generation. The input control signals are text, an artistic drawing, and a segmentation mask.
- Fig. 24 shows the videos generated by our approach for the task of dependent multimodal generation. The input control signals are text, an image (used for appearance), and a video (used for motion guidance, which can be better observed in our supplementary video).
- Fig. 25 shows the videos generated by methods w/ (*w/* RoBERTa) and w/o (*w/o* RoBERTa) using language embedding from RoBERTa [7] as text augmentation. Models are trained on the Multimodal VoxCeleb dataset for text-to-video generation.
- Fig. 26, Fig. 27, and Fig. 28 shows the videos synthesized by TFGAN for text-to-video generation, ART-V for text-to-video generation, and ART-V for independent multimodal generation, respectively. Artifacts can be observed from the generated videos.
- Fig. 31 shows more videos generated from partially occluded faces. Note that the occlusion pattern at test time is different from that at training time.
- Fig. 32 shows examples of nearest neighbor analysis on the Multimodal VoxCeleb dataset. We show generated samples by using: 1) contradicting conditions with text suggests a female and the image shows a beard (upper, the result is an unseen combination); and 2) various conditioning (lower, conditioning omitted) with their nearest neighbors in VoxCeleb found using face similarity scores[2].

S4. Limitation and Future Work

Higher Resolution Generation. We conduct experiments to generate higher resolution videos by performing experiments on the Multimodal VoxCeleb dataset to synthesize video with the resolution of 256×256. Synthesized videos for the task of text-to-video generation are shown in Fig. 29. We notice that artifacts can be found from some videos: *e.g.* we find synthesized samples are more likely to show

[2] Face Recognition Code: https://github.com/timesler/facenet-pytorch

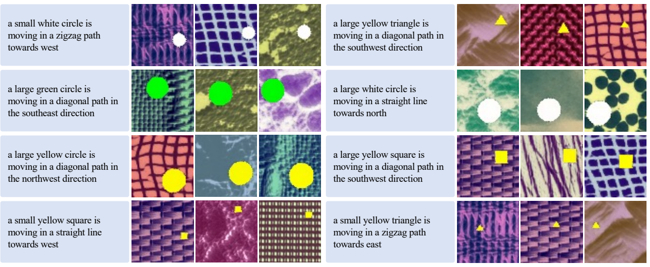

a small white circle is moving in a zigzag path towards west
a large yellow triangle is moving in a diagonal path in the southwest direction
a large green circle is moving in a diagonal path in the southeast direction
a large white circle is moving in a straight line towards north
a large yellow circle is moving in a diagonal path in the northwest direction
a large yellow square is moving in a diagonal path in the southwest direction
a small yellow square is moving in a straight line towards west
a small yellow triangle is moving in a zigzag path towards east

Figure 13. Example videos generated by our approach on the Shapes dataset for text-to-video generation. We show three synthesized videos for each input text condition.

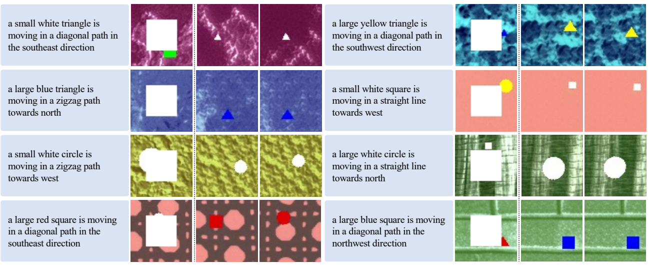

a small white triangle is moving in a diagonal path in the southeast direction
a large yellow triangle is moving in a diagonal path in the southwest direction
a large blue triangle is moving in a zigzag path towards north
a small white square is moving in a straight line towards west
a small white circle is moving in a zigzag path towards west
a large white circle is moving in a straight line towards north
a large red square is moving in a diagonal path in the southeast direction
a large blue square is moving in a diagonal path in the northwest direction

Figure 14. Example videos generated by our approach on the Shapes dataset for independent multimodal generation. The input control signals are text and a partially observed image (with the center masked out, shown in white color). We show two synthesized videos for each input multimodal condition.

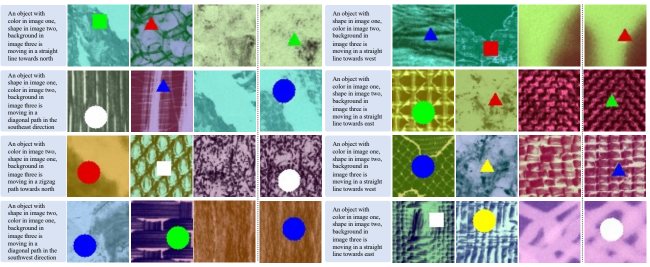

An object with color in image one, shape in image two, background in image three is moving in a straight line towards north
An object with shape in image one, color in image two, background in image three is moving in a straight line towards west
An object with shape in image one, color in image two, background in image three is moving in a diagonal path in the southeast direction
An object with color in image one, shape in image two, background in image three is moving in a straight line towards east
An object with color in image two, shape in image one, background in image three is moving in a zigzag path towards north
An object with color in image one, shape in image two, background in image three is moving in a straight line towards west
An object with shape in image two, color in image one, background in image three is moving in a diagonal path in the southwest direction
An object with color in image one, shape in image two, background in image three is moving in a straight line towards east

Figure 15. Example videos generated by our approach on the Shapes dataset for dependent multimodal generation. The input control signals are text and images. We show one synthesized video for each input multimodal condition.

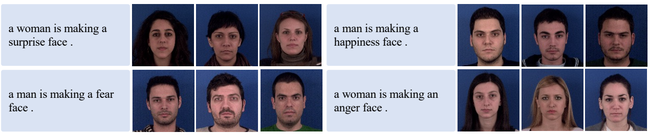

a woman is making a surprise face .
a man is making a happiness face .
a man is making a fear face .
a woman is making an anger face .

Figure 16. Example videos generated by our approach on the MUG dataset for text-to-video generation. We show three synthesized videos for each input text condition.

weird colors (Fig. 29, the third row and the second sample) or appear to be blurry (Fig. 29 the last row and the second and the fourth sample). We also find that the temporal consistency is worse than videos generated at the resolution of 128 × 128. One possible reason might be that each frame at a higher resolution requires a longer token sequence. Therefore, an image-level temporal regularization constraint might be necessary to improve the video consistency, which we leave for future work.

Longer Sequence Generation. For the task of long sequence generation, we notice that extrapolation might not always give reasonable motion patterns. For example, as shown in Fig. 17, the textual inputs from the first row struggle to generate diverse motions when the speed is given as "slow". This might be due to the temporal step size used to sample frames during training being short when the speed is "slow" and the frames cannot always cover diverse motion patterns. A future direction could be balancing the training set to cover enough motion patterns for the sampled frames.

Diversity of Non-Autoregressive Transformer. Compared with the autoregressive transformer, the non-autoregressive transformer can generate videos with better temporal consistency. However, we notice that the autoregressive transformer might generate more diverse videos, though many have low video quality. We apply text augmentation and improved mask-predict to improve the diversity of the non-autoregressive transformer. An interesting research direction is how to unify the training methods from the autoregressive and non-autoregressive transformer to enhance the non-autoregressive transformer itself for generating more diverse videos. We leave the direction as to future work.

S5. Ethical Implications

Our method can synthesize high-quality videos with multimodal inputs. However, a common concern among the works for high fidelity and realistic image and video generation is the purposely abusing the technology for nefarious objectives. The techniques developed for synthetic image and video detection can help alleviate such a problem by automatically finding artifacts that humans might not easily notice.

References

[1] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4. IEEE, 2010. 6

[2] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discrimina-

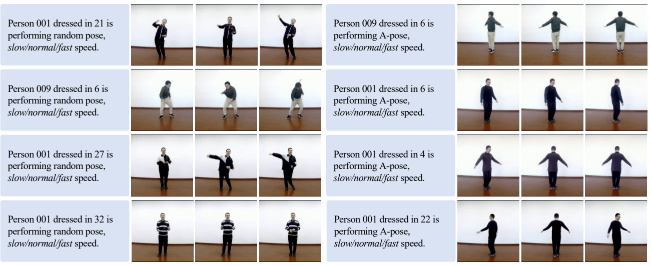

Person 001 dressed in 21 is performing random pose, *slow/normal/fast* speed.
Person 009 dressed in 6 is performing random pose, *slow/normal/fast* speed.
Person 001 dressed in 27 is performing random pose, *slow/normal/fast* speed.
Person 001 dressed in 32 is performing random pose, *slow/normal/fast* speed.
Person 009 dressed in 6 is performing A-pose, *slow/normal/fast* speed.
Person 001 dressed in 6 is performing A-pose, *slow/normal/fast* speed.
Person 001 dressed in 4 is performing A-pose, *slow/normal/fast* speed.
Person 001 dressed in 22 is performing A-pose, *slow/normal/fast* speed.

Figure 17. Example videos generated by our approach on the iPER dataset for long sequence generation. The extrapolation process is repeated for each sequence 100 times, resulting in a 107-frame video. The textual input also controls the speed, where "slow" indicates videos with slow speed such that the motion is slow, while "fast" indicates the performed motion is fast. We show one synthesized video for each input text condition. The first video following the text input corresponds to the "slow" condition, the second corresponds to the "normal", and the last corresponds to the "fast".

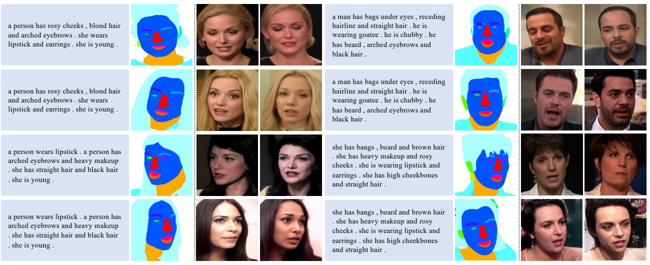

a person has rosy cheeks , blond hair and arched eyebrows . she wears lipstick and earrings . she is young .
a person has rosy cheeks , blond hair and arched eyebrows . she wears lipstick and earrings . she is young .
a person wears lipstick . a person has arched eyebrows and heavy makeup . she has straight hair and black hair . she is young .
a person wears lipstick . a person has arched eyebrows and heavy makeup . she has straight hair and black hair . she is young .
a man has bags under eyes , receding hairline and straight hair . he is wearing goatee . he is chubby . he has beard , arched eyebrows and black hair .
a man has bags under eyes , receding hairline and straight hair . he is wearing goatee . he is chubby . he has beard , arched eyebrows and black hair .
she has bangs , beard and brown hair . she has heavy makeup and rosy cheeks . she is wearing lipstick and earrings . she has high cheekbones and straight hair .
she has bangs , beard and brown hair . she has heavy makeup and rosy cheeks . she is wearing lipstick and earrings . she has high cheekbones and straight hair .

Figure 19. Example videos generated by our approach on the Multimodal VoxCeleb dataset for independent multimodal video generation. The input control signals are text and a segmentation mask. We show two synthesized videos for each input multimodal condition.

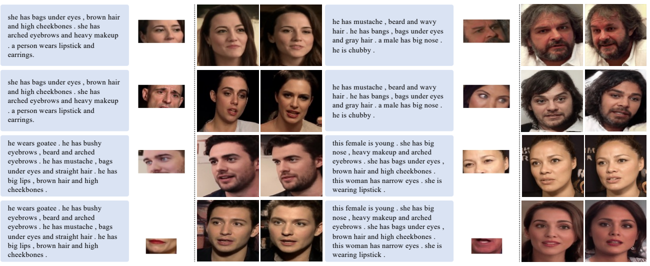

she has bags under eyes , brown hair and high cheekbones . she has arched eyebrows and heavy makeup . a person wears lipstick and earrings.
she has bags under eyes , brown hair and high cheekbones . she has arched eyebrows and heavy makeup . a person wears lipstick and earrings.
he wears goatee . he has bushy eyebrows , beard and arched eyebrows . he has mustache , bags under eyes and straight hair . he has big lips , brown hair and high cheekbones .
he wears goatee . he has bushy eyebrows , beard and arched eyebrows . he has mustache , bags under eyes and straight hair . he has big lips , brown hair and high cheekbones .
he has mustache , beard and wavy hair . he has bangs , bags under eyes and gray hair . a male has big nose . he is chubby .
he has mustache , beard and wavy hair . he has bangs , bags under eyes and gray hair . a male has big nose . he is chubby .
this female is young . she has big nose , heavy makeup and arched eyebrows . she has bags under eyes , brown hair and high cheekbones . this woman has narrow eyes . she is wearing lipstick .
this female is young . she has big nose , heavy makeup and arched eyebrows . she has bags under eyes , brown hair and high cheekbones . this woman has narrow eyes . she is wearing lipstick .

Figure 21. Example videos generated by our approach on the Multimodal VoxCeleb dataset for independent multimodal video generation. The input control signals are text and a partially observed image. We show two synthesized videos for each input condition.

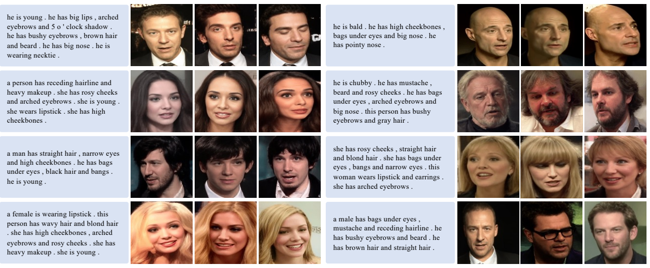

he is young . he has big lips , arched eyebrows and 5 o ' clock shadow . he has bushy eyebrows , brown hair and beard . he has big nose . he is wearing necktie .
a person has receding hairline and heavy makeup . she has rosy cheeks and arched eyebrows . she is young . she wears lipstick . she has high cheekbones .
a man has straight hair , narrow eyes and high cheekbones . he has bags under eyes , black hair and bangs . he is young .
a female is wearing lipstick . this person has wavy hair and blond hair . she has high cheekbones , arched eyebrows and rosy cheeks . she has heavy makeup . she is young .
he is bald . he has high cheekbones , bags under eyes and big nose . he has pointy nose .
he is chubby . he has mustache , beard and rosy cheeks . he has bags under eyes , arched eyebrows and big nose . this person has bushy eyebrows and gray hair .
she has rosy cheeks , straight hair and blond hair . she has bags under eyes , bangs and narrow eyes . this woman wears lipstick and earrings . she has arched eyebrows .
a male has bags under eyes , mustache and receding hairline . he has bushy eyebrows and beard . he has brown hair and straight hair .

Figure 18. Example videos generated by our approach on the Multimodal VoxCeleb dataset for text-to-video generation. We show three synthesized videos for each input text condition.

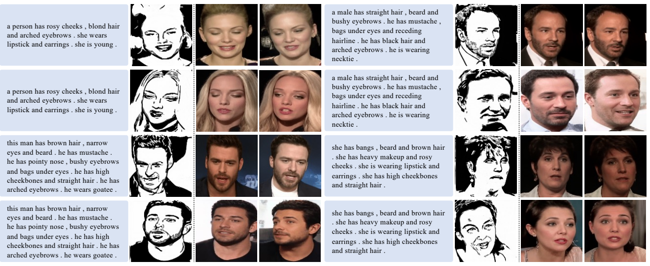

a person has rosy cheeks , blond hair and arched eyebrows . she wears lipstick and earrings . she is young .
a person has rosy cheeks , blond hair and arched eyebrows . she wears lipstick and earrings . she is young .
this man has brown hair , narrow eyes and beard . he has mustache . he has pointy nose , bushy eyebrows and bags under eyes . he has high cheekbones and straight hair . he has arched eyebrows . he wears goatee .
this man has brown hair , narrow eyes and beard . he has mustache . he has pointy nose , bushy eyebrows and bags under eyes . he has high cheekbones and straight hair . he has arched eyebrows . he wears goatee .
a male has straight hair , beard and bushy eyebrows . he has mustache , bags under eyes and receding hairline . he has black hair and arched eyebrows . he is wearing necktie .
a male has straight hair , beard and bushy eyebrows . he has mustache , bags under eyes and receding hairline . he has black hair and arched eyebrows . he is wearing necktie .
she has bangs , beard and brown hair . she has heavy makeup and rosy cheeks . she is wearing lipstick and earrings . she has high cheekbones and straight hair .
she has bangs , beard and brown hair . she has heavy makeup and rosy cheeks . she is wearing lipstick and earrings . she has high cheekbones and straight hair .

Figure 20. Example videos generated by our approach on the Multimodal VoxCeleb dataset for independent multimodal video generation. The input control signals are text and an artistic drawing. We show two synthesized videos for each input multimodal condition.

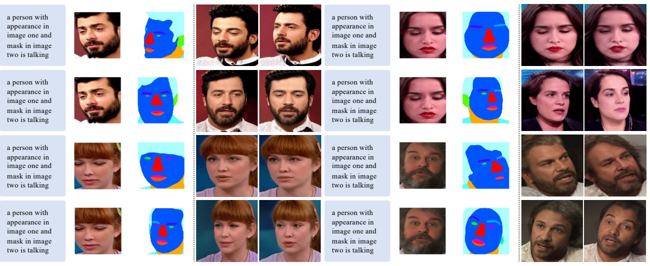

a person with appearance in image one and mask in image two is talking
a person with appearance in image one and mask in image two is talking
a person with appearance in image one and mask in image two is talking
a person with appearance in image one and mask in image two is talking
a person with appearance in image one and mask in image two is talking
a person with appearance in image one and mask in image two is talking
a person with appearance in image one and mask in image two is talking
a person with appearance in image one and mask in image two is talking

Figure 22. Example videos generated by our approach on the Multimodal VoxCeleb dataset for dependent multimodal video generation. The input control signals are text, an image, and a segmentation mask. We show two synthesized videos for each input condition.

tive filter generation for text-to-video synthesis. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1995–2001. International Joint Conferences on Artificial Intelligence Organization, 2019. 6

[3] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, page 2, 2019. 6

[4] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 6

[5] Doyeon Kim, Donggyu Joo, and Junmo Kim. Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access*, 8:153113–153122, 2020. 6

[6] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF Interna-*

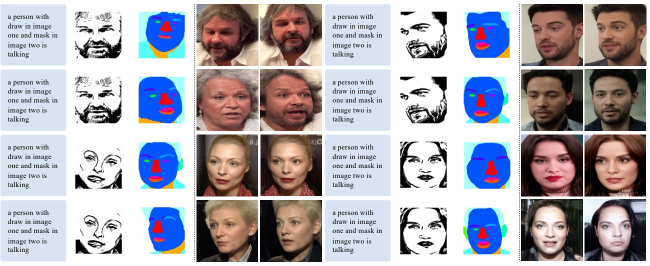

a person with draw in image one and mask in image two is talking
a person with draw in image one and mask in image two is talking
a person with draw in image one and mask in image two is talking
a person with draw in image one and mask in image two is talking
a person with draw in image one and mask in image two is talking
a person with draw in image one and mask in image two is talking
a person with draw in image one and mask in image two is talking
a person with draw in image one and mask in image two is talking

Figure 23. Example videos generated by our approach on the Multimodal VoxCeleb dataset for dependent multimodal video generation. The input control signals are text, an artistic drawing, and a segmentation mask. We show two synthesized videos for each input multimodal condition.

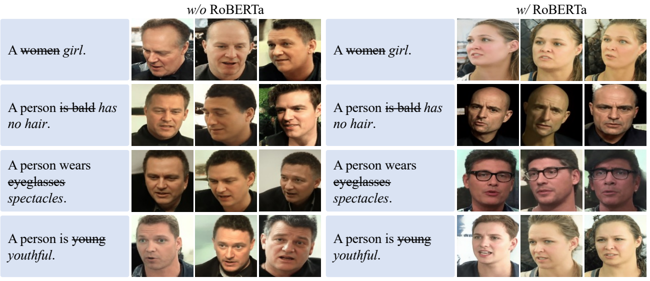

w/o RoBERTa
w/ RoBERTa
A ~~women~~ *girl*.
A person ~~is bald~~ *has no hair*.
A person wears ~~eyeglasses~~ *spectacles*.
A person is ~~young~~ *youthful*.
A ~~women~~ *girl*.
A person ~~is bald~~ *has no hair*.
A person wears ~~eyeglasses~~ *spectacles*.
A person is ~~young~~ *youthful*.

Figure 25. Example videos generated by methods w/ (*w/* RoBERTa) and w/o (*w/o* RoBERTa) using language embedding from RoBERTa as text augmentation. Models are trained on the Multimodal VoxCeleb dataset for text-to-video generation. We show three synthesized videos for each input text condition.

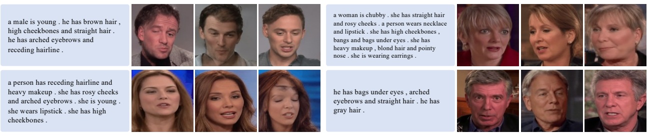

a male is young . he has brown hair , high cheekbones and straight hair . he has arched eyebrows and receding hairline .
a woman is chubby . she has straight hair and rosy cheeks . a person wears necklace and lipstick . she has high cheekbones , bangs and bags under eyes . she has heavy makeup , blond hair and pointy nose . she is wearing earrings .
a person has receding hairline and heavy makeup . she has rosy cheeks and arched eyebrows . she is young . she wears lipstick . she has high cheekbones .
he has bags under eyes , arched eyebrows and straight hair . he has gray hair .

Figure 27. Example videos generated by ART-V on the Multimodal VoxCeleb dataset for text-to-video generation. We show three synthesized videos for each input text condition.

tional Conference on Computer Vision, pages 5904–5913, 2019. 6

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 8

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 2

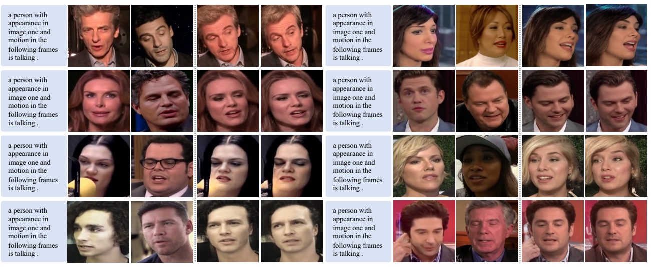

a person with appearance in image one and motion in the following frames is talking .
a person with appearance in image one and motion in the following frames is talking .
a person with appearance in image one and motion in the following frames is talking .
a person with appearance in image one and motion in the following frames is talking .
a person with appearance in image one and motion in the following frames is talking .
a person with appearance in image one and motion in the following frames is talking .
a person with appearance in image one and motion in the following frames is talking .
a person with appearance in image one and motion in the following frames is talking .

Figure 24. Example videos generated by our approach on the Multimodal VoxCeleb dataset for dependent multimodal video generation. The input control signals are text, an image (used for appearance), and a video (used for motion guidance, which can be better observed in our supplementary video). We show two synthesized videos for each input multimodal condition.

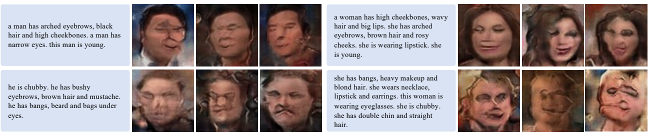

a man has arched eyebrows, black hair and high cheekbones. a man has narrow eyes. this man is young.
a woman has high cheekbones, wavy hair and big lips. she has arched eyebrows, brown hair and rosy cheeks. she is wearing lipstick. she is young.
he is chubby. he has bushy eyebrows, brown hair and mustache. he has bangs, beard and bags under eyes.
she has bangs, heavy makeup and blond hair. she wears necklace, lipstick and earrings. this woman is wearing eyeglasses. she is chubby. she has double chin and straight hair.

Figure 26. Example videos generated by TFGAN on the Multimodal VoxCeleb dataset for text-to-video generation. We show three synthesized videos for each input text condition.

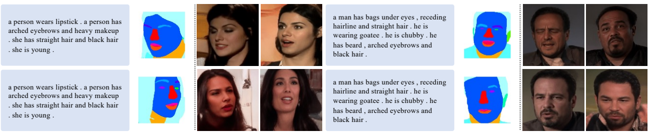
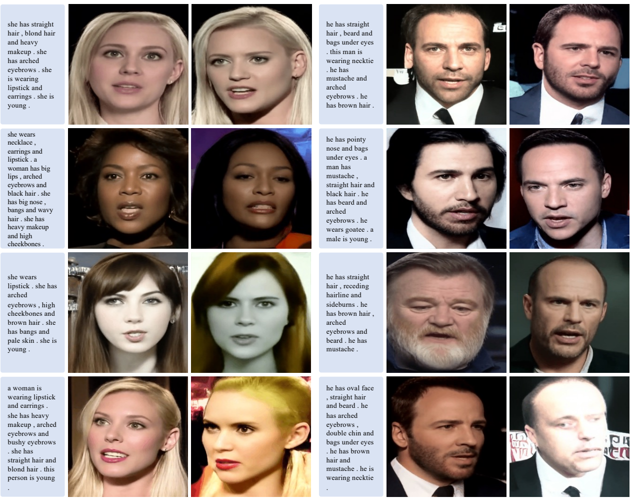

a person wears lipstick . a person has arched eyebrows and heavy makeup . she has straight hair and black hair . she is young .
a man has bags under eyes , receding hairline and straight hair . he is wearing goatee . he is chubby . he has beard , arched eyebrows and black hair .
a person wears lipstick . a person has arched eyebrows and heavy makeup . she has straight hair and black hair . she is young .
a man has bags under eyes , receding hairline and straight hair . he is wearing goatee . he is chubby . he has beard , arched eyebrows and black hair .

Figure 28. Example videos generated by ART-V on the Multimodal VoxCeleb dataset for independent multimodal video generation. The input control signals are text and a segmentation mask. We show two synthesized videos for each input multimodal condition.

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[10] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[11] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. Ufcbert: Unifying multi-modal controls for conditional image synthesis. *arXiv preprint arXiv:2105.14211*, 2021. 5



she has straight hair , blond hair and heavy makeup . she has arched eyebrows . she is wearing lipstick and earrings . she is young .

she wears necklace , earrings and lipstick . a woman has big lips , arched eyebrows and black hair . she has big nose , bangs and wavy hair . she has heavy makeup and high cheekbones .

she wears lipstick . she has arched eyebrows , high cheekbones and brown hair . she has bangs and pale skin . she is young .

a woman is wearing lipstick and earrings . she has heavy makeup , arched eyebrows and bushy eyebrows . she has straight hair and blond hair . this person is young .

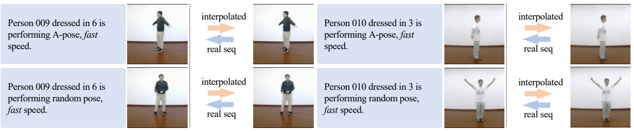
he has straight hair , beard and bags under eyes . this man is wearing necktie . he has mustache and arched eyebrows . he has brown hair .

he has pointy nose and bags under eyes . a man has mustache , straight hair and black hair . he has beard and arched eyebrows . he wears goatee . a male is young .

he has straight hair , receding hairline and sideburns . he has brown hair , arched eyebrows and beard . he has mustache .

he has oval face , straight hair and beard . he has arched eyebrows , double chin and bags under eyes . he has brown hair and mustache . he is wearing necktie .

Figure 29. Example videos generated by our approach on the Multimodal VoxCeleb dataset for text-to-video generation. Videos are synthesized at a resolution of 256×256. We show two synthesized videos for each input text condition. We use $L = 25, L_1 = 12$, $L_2 = 13$, $\alpha_1 = 0.9$, $\beta_1 = 0.1$, $\alpha_2 = 0.125$, and $\alpha_3 = 0.0625$ for mask-predict.



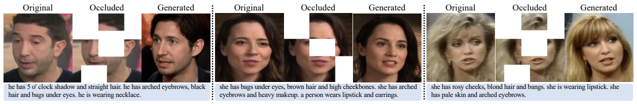
Person 009 dressed in 6 is performing A-pose, *fast* speed.

Person 010 dressed in 3 is performing A-pose, *fast* speed.

Person 009 dressed in 6 is performing random pose, *fast* speed.

Person 010 dressed in 3 is performing random pose, *fast* speed.

Figure 30. Example videos of our approach for video interpolation on iPER dataset.

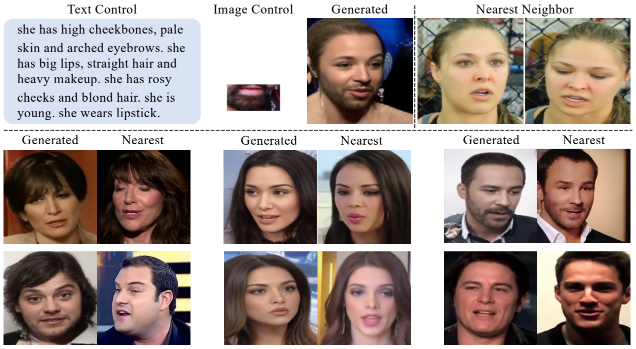


he has 5 o' clock shadow and straight hair. he has arched eyebrows, black hair and bags under eyes. he is wearing necklace.

she has bags under eyes, brown hair and high cheekbones. she has arched eyebrows and heavy makeup. a person wears lipstick and earrings.

she has rosy cheeks, blond hair and bangs. she is wearing lipstick. she has pale skin and arched eyebrows.

Figure 31. More example videos generated by our approach on the Multimodal VoxCeleb dataset for independent multimodal video generation. The input control signals are text and a partially observed image. Note that the occlusion pattern at test time is different from that at training time (with only either mouth or eyes and nose observable).



she has high cheekbones, pale skin and arched eyebrows. she has big lips, straight hair and heavy makeup. she has rosy cheeks and blond hair. she is young. she wears lipstick.

Figure 32. Examples of nearest neighbor analysis on the Multimodal VoxCeleb dataset. We show generated samples by using: 1) contradicting conditions with text suggests a female and the image shows a beard (upper, the result is an unseen combination); and 2) various conditioning (lower, conditioning omitted) with their nearest neighbors in VoxCeleb found using face similarity scores.