

Supervising 3D Talking Head Avatars with Analysis-by-Audio-Synthesis

Supplementary Material

7. Additional Experiments

7.1. Mesh-to-Speech

7.1.1 Per-Dataset Performance Analysis.

The performance of both mesh-to-speech and silent-video-to-speech models varies dramatically depending on the dataset. Datasets with limited number of utterances and vocabulary such as RAVDESS or GRID are less challenging, and the models can regress relatively accurately, sometimes to the point of perfect intelligibility. On the more challenging TCD-TIMIT dataset, which contains more complex vocabulary, words are often missed. However, despite the relatively high word error rate (WER) on TCD-TIMIT, the produced sounds remain plausible given the input animation. Please see Tab. 4 for per-dataset breakdown and refer to the supplementary video for an audible comparison.

DATASET	Model:	STOI \uparrow	ESTOI \uparrow	PESQ-WB \uparrow	PESQ-NB \uparrow	WER \downarrow
GRID	exp2speech \mathbf{x}	0.533	0.290	1.298	1.552	0.553
	face2speech \mathbf{V}	0.482	0.211	1.261	1.513	0.783
	mouth2speech $\mathbf{V}_{m\in\mathcal{M}}$	0.538	0.295	1.288	1.539	0.559
	Choi et al. (finetuned)	0.590	0.371	1.324	1.597	0.428
	Choi et al.	0.426	0.158	1.137	1.345	1.202
RAVDESS	exp2speech \mathbf{x}	0.460	0.284	1.159	1.262	0.145
	face2speech \mathbf{V}	0.474	0.293	1.164	1.266	0.173
	mouth2speech $\mathbf{V}_{m\in\mathcal{M}}$	0.510	0.328	1.161	1.262	0.131
	Choi et al. (finetuned)	0.548	0.375	1.174	1.277	0.060
	Choi et al.	0.452	0.245	1.074	1.149	0.774
TCD-TIMIT	exp2speech \mathbf{x}	0.409	0.206	1.146	1.249	0.929
	face2speech \mathbf{V}	0.368	0.139	1.117	1.228	1.431
	mouth2speech $\mathbf{V}_{m\in\mathcal{M}}$	0.389	0.174	1.131	1.243	1.070
	Choi et al. (finetuned)	0.436	0.257	1.168	1.293	0.579
	Choi et al.	0.172	0.041	1.103	1.254	0.641
COMBINED	exp2speech \mathbf{x}	0.502	0.273	1.257	1.468	0.648
	face2speech \mathbf{V}	0.458	0.203	1.225	1.437	0.953
	mouth2speech $\mathbf{V}_{m\in\mathcal{M}}$	0.506	0.272	1.246	1.457	0.678
	Choi et al. (finetuned)	0.555	0.348	1.281	1.511	0.437
	Choi et al. (orig)	0.376	0.141	1.126	1.313	1.011

Table 4. **Quantitative comparison of mesh-to-speech with silent-video-to-speech.** While Choi et al. outperforms THUNDER in most cases, the performance is comparable. The performance of both methods depends dramatically on the test dataset. Inputs from datasets with limited vocabulary (GRID, RAVDESS) tend to produce lower word error rates than datasets with richer vocabulary (TCD-TIMIT).

7.2. Speech-Driven Facial Animation

7.2.1 Experiments on TFHP

To verify whether the performance of THUNDER and mesh-to-speech translates to other types of data, we also conduct a comprehensive evaluation on TFHP (the DiffPoseTalk dataset [74]). TFHP was reconstructed from YouTube videos. The dataset contains unscripted audios with unconstrained longer sequences. The reconstructions were obtained using MICA [100] and SPECTRE [29] and look qualitatively very

different compared to our EMICA reconstructions of GRID, RAVDESS and TCD-TIMIT datasets. Can THUNDER work also in this setting?

We train both M2S and THUNDER on TFHP and compare it to the official DiffPoseTalk release. To achieve maximum fairness of comparison, we adopt DiffPoseTalk’s number of diffusion steps ($D = 500$) and their cosine diffusion schedule in THUNDER. We also use their dataset normalization statistics (i.e. normalize the FLAME coefficients by subtracting mean and dividing by standard deviation). Furthermore, in order to equalize the input conditions, we trained THUNDER with an *extra input condition* - the FLAME identity shape vector β , which is also what the original DiffPoseTalk is trained with. We train a number of models to cover all possible scenarios:

(1) Models trained without head pose. Consistently to the experiments on THUNDERSET, we find that THUNDER models benefit from the M2S loss, which can be observed in improvement on all lip-sync metrics. The improvement is more dramatic in THUNDER-F models, but manifests itself also in the models with trainable audio encoders (THUNDER-F). Similarly to our experiments on THUNDERSET, the application of M2S comes with a small reduction in upper-face diversity. Finally, THUNDER-T outperforms DiffPoseTalk on all lip-sync metrics. The metrics can be seen in Tab. 5, experiment 1.

(2) Models trained with head pose. Since TFHP comes with head pose annotations that correspond to natural head movement during spontaneous speech, we investigate if THUNDER models can also be trained to produce head pose. When it comes to lip-sync and face diversity, all findings are consistent with that of the experiment without head pose. Additionally, we find that THUNDER is capable of producing head pose with higher beat alignment (BA) and higher diversity (DIV) than DiffPoseTalk. This suggests that the M2S loss does not interfere with head pose generation. See Tab. 5, experiment 2 for more detail. Note that the beat alignment and diversity were computed with the global rotation representation converted into the 6D rotation representation.

(3) Models trained with head pose and speaking style conditioning. Finally, we train THUNDER models with the DiffPoseTalk contrastive style features as conditions. The results can be found in See Tab. 5, experiment 3. We find that conditioning with the style vector results in better lip-sync

Experiment	Name	Backbone	Input	Lip-Sync			Upper diversity [mm]		Lip diversity [mm]		Face Dynamic Dev.		Head Pose		Chae et al. [8]		
				LVE [mm] ↓	L-CCC ↑	L-PCC ↑	DTW [mm] ↓	S-DIV-U ↑	T-DIV-U ↑	S-DIV-L ↓	T-DIV-L ↓	T-FDD-U ↓	T-FDD-L ↓	BA × 10 ⁻³ ↑	DIV ↑	MTM [ms] ↓	PLRS ↑
(1) models trained without head pose	THUNDER-F w/o m2s	Wav2Vec2 frozen	shape, audio	1.21	0.316	0.42	2.37	0.299	0.386	2.7	3.71	0.0739	0.79	n/a	n/a	125.141	0.175
	THUNDER-F	Wav2Vec2 frozen	shape, audio	1.05	0.317	0.441	1.93	0.195	0.328	1.5	3.12	0.0627	0.759	n/a	n/a	103.185	0.269
	DiffPoseTalk	HuBERT trainable	shape, audio	1.02	0.41	0.526	1.84	0.272	0.383	1.7	2.98	0.0764	0.679	n/a	n/a	93.485	0.265
	THUNDER-T w/o m2s	Wav2Vec2 trainable	shape, audio	1.03	0.388	0.525	1.84	0.265	0.364	1.58	2.95	0.0679	0.609	n/a	n/a	94.132	0.247
(2) models trained with head pose	THUNDER-T	Wav2Vec2 trainable	shape, audio	0.963	0.415	0.55	1.77	0.228	0.341	1.09	2.79	0.0592	0.59	n/a	n/a	88.908	0.329
	THUNDER-F w/o m2s	Wav2Vec2 frozen	shape, audio	1.2	0.285	0.388	2.26	0.258	0.349	2.6	3.58	0.0649	0.76	6.02	3.07	134.403	0.153
	THUNDER-F	Wav2Vec2 frozen	shape, audio	1.07	0.298	0.435	1.95	0.203	0.321	1.5	2.98	0.0707	0.95	6.13	3.78	105.536	0.292
	DiffPoseTalk	HuBERT trainable	shape, audio	1.01	0.414	0.541	1.85	0.247	0.352	1.58	3.02	0.0641	0.618	5.54	2.16	90.005	0.251
(3) models trained with head pose and style condition	THUNDER-T w/o m2s	Wav2Vec2 trainable	shape, audio	1.12	0.358	0.488	1.86	0.241	0.351	1.87	3.21	0.0597	0.661	6.25	3.49	101.232	0.230
	THUNDER-T	Wav2Vec2 trainable	shape, audio	0.987	0.422	0.55	1.81	0.244	0.341	1.35	2.89	0.0578	0.55	5.77	4.83	89.550	0.320
	DiffPoseTalk	HuBERT trainable	audio, shape, style	0.897	0.439	0.555	1.77	0.191	0.319	1.27	2.86	0.0632	0.572	5.57	1.43	88.647	0.256
	THUNDER-F w/o m2s	Wav2Vec2 trainable	audio, shape, style	0.899	0.443	0.545	1.73	0.206	0.31	1.45	2.91	0.0549	0.573	5.72	3.3	94.046	0.261
(4) models trained with perc. loss Chae et al. [8]	THUNDER-T	Wav2Vec2 trainable	audio, shape, style	0.891	0.442	0.567	1.72	0.186	0.31	0.941	2.81	0.0598	0.625	5.98	1.96	89.108	0.327
	THUNDER-F w/o m2s	Wav2Vec2 frozen	audio, shape, style	1.02	0.309	0.386	2.04	0.227	0.323	2.39	3.42	0.0646	0.726	5.85	2.64	134.639	0.124
	THUNDER-F	Wav2Vec2 frozen	audio, shape, style	0.917	0.365	0.465	1.85	0.181	0.312	1.49	2.978	0.0679	0.755	5.92	2.4	103.993	0.280
	THUNDER-T w/o m2s, perc.	Wav2Vec2 trainable	audio, shape	1.117	0.358	0.488	1.864	0.241	0.351	1.87	3.21	0.060	0.661	6.25	3.49	101.232	0.230
	THUNDER-T w/o m2s, w/ perc.	Wav2Vec2 trainable	audio, shape	1.117	0.325	0.485	1.991	0.246	0.344	1.60	3.00	0.064	0.783	5.71	2.99	100.658	0.472
	THUNDER-T w/ m2s, w/o perc.	Wav2Vec2 trainable	audio, shape	0.987	0.422	0.550	1.811	0.244	0.341	1.35	2.89	0.058	0.550	5.77	4.83	89.550	0.320
	THUNDER-T w/ m2s, perc.	Wav2Vec2 trainable	audio, shape	1.079	0.365	0.537	1.869	0.224	0.329	1.11	2.74	0.064	0.777	5.92	2.88	88.862	0.469
	THUNDER-F w/o m2s, perc.	Wav2Vec2 frozen	audio, shape	1.2	0.285	0.388	2.26	0.258	0.349	2.6	3.58	0.0649	0.76	6.02	3.07	134.403	0.153
(5) models trained with perc. loss Chae et al. [8]	THUNDER-F w/o m2s, w/ perc.	Wav2Vec2 frozen	audio, shape	1.2	0.318	0.462	2.27	0.28	0.364	2.07	3.11	0.0652	0.704	5.98	3.66	107.130	0.455
	THUNDER-F w/ m2s, w/o perc.	Wav2Vec2 frozen	audio, shape	1.07	0.298	0.435	1.95	0.203	0.321	1.5	2.98	0.0707	0.95	6.13	3.78	105.536	0.292
	THUNDER-F w/ m2s, perc.	Wav2Vec2 frozen	audio, shape	1.16	0.348	0.498	2.07	0.33	0.375	2.13	3.23	0.0751	0.687	5.9	4.48	99.495	0.479

Table 5. **Experiment on TFHP.** We compare THUNDER-T, THUNDER w/o mesh-to-speech and DiffPoseTalk (all trained on TFHP). Similarly to our other experiments, THUNDER-T results in superior lip-sync metrics, while trading off upper-face diversity.

performance, lower face diversity, higher beat alignment and lower pose diversity compared to measurements in See Tab. 5, experiment 2. All of these are consistent with expectations, since passing in a style specification reduces the distribution of possible outcomes. Regardless of the above, the application of M2S has the expected effect - improved lip-sync metrics and lower diversity.

(4) Co-supervising with a perceptual loss. A recent paper by Chae et al. [8] proposes a new self-supervised speech-mesh representation, in form of a bi-modal masked autoencoder. The architecture consists of two encoders and two decoders (one for the mesh modality and one for the audio modality). The model is trained with two loss terms - the autoencoder reconstruction losses for both modalities and the InfoNCE loss which brings the two modalities into one coherent features space. This is similar to what CLIP [66] does with images and text. The authors show that their representation can be used in training other speech-driven animation methods. They use their frozen autoencoder in training, passing the input audio and output geometry in, extracting the features for both. Then, they apply the same InfoNCE loss between the two modalities and use it to co-supervise the talking head avatar system. In this experiment, we incorporate this perceptual loss into the THUNDER training. The training loss is then given as:

$$\mathcal{L}_{total} = w_{m2s}\mathcal{L}_{m2s} + w_{perc}\mathcal{L}_{perc} + \mathcal{L}_{rec} + \mathcal{L}_{vel}, \quad (9)$$

where $w_{perc}\mathcal{L}_{perc}$ is the perceptual term. We follow the authors and set the weight $w_{perc} = 0.1$, a factor of 7 orders of magnitude lower than the vertex loss, which puts the computed losses on a similar scale and hence results in optimal effect (i.e not too strong or too weak). For completeness, we also report the new metrics proposed by Chae et al. [8] - mean temporal misalignment (MTM) and Perceptual Lip Readability Score (PLRS) along with the other metrics used

in this paper. MTM measures the temporal misalignment using a Derivative Dynamic Time Warping (DDTW). SLRS is the same InfoNCE-based perceptual loss used for training, but computed by a different instance of the same architecture. The authors released two instances - one to be used as the perceptual loss and one to be used for evaluation with SLRS. We utilize both of the speech-mesh autoencoder models released by the authors accordingly. Tab. 5, experiment 4 reports the results of this experiment. In this experiment we investigate the following:

(a) *Does training THUNDER with mesh-to-speech result in improvement on metrics proposed by Chae et al. ?* Yes, all THUNDER in all settings (across experiments 1-4) improve THUNDER without mesh-to-speech and DiffPoseTalk on both MTM and PLRS. The improvement on MTM is expected, since MTM leverages DDTW, which is similar to the DTW metric [84] we use. Furthermore, improvement on PLRS indicates that the mesh-to-speech loss results in a perceptual improvement in quality.

(b) *Does training our diffusion-based speech-driven animation method with this perceptual loss improve performance?* Yes, THUNDER models trained without mesh-to-speech but with the perceptual loss improve considerably on PLRS. This is expected since PLRS is computed by a model with the same architecture as the network which provides the perceptual loss function. We also observe improvement on MTM. Furthermore, we observe improvement on our lip-sync related metrics in models with the frozen backbone (-F). Models with trainable backbones (-T) do not seem to improve on LVE, CCC, PCC or DTW when the perceptual loss is applied.

(c) *Can THUNDER be trained with both mesh-to-speech and the perceptual loss.* Yes. Both losses can be applied together. The mesh-to-speech loss has a strong impact on LVE, CCC, PCC, DTW and MTM, while the perceptual loss strongly affects PLRS, MTM and with -F models also improves CCC and PCC. THUNDER-F with both mesh-to-

Name	LVE [cm] ↓	LIP CCC ↑	LIP PCC ↑	DTW [cm] ↓	S-DIV-L [cm] ↓	MTM [ms] ↓	PLRS ↑
M2F-F* orig.	0.611	0.522	0.617	0.221	0.0956	41.0	0.27
M2F-F* w/ m2s orig.	0.551	0.558	0.679	0.212	0.027	57.5	0.283
M2F-F* angry	1.6	0.307	0.526	0.914	0.151	88.8	0.173
M2F-F* w/ m2s angry	1.46	0.348	0.59	0.741	0.08	61.7	0.176
M2F-F* calm	0.91	0.369	0.537	0.297	0.103	75.5	0.257
M2F-F* w/ m2s calm	0.86	0.459	0.638	0.302	0.0289	62.5	0.27
M2F-F* disgust	0.928	0.314	0.527	0.316	0.0675	77.9	0.249
M2F-F* w/ m2s disgust	0.874	0.426	0.627	0.305	0.019	66.6	0.296
M2F-F* fearful	1.01	0.38	0.618	0.315	0.101	60.8	0.263
M2F-F* w/ m2s fearful	1.03	0.353	0.627	0.303	0.0295	56.4	0.24
M2F-F* happy	1.31	0.201	0.394	0.484	0.152	82.0	0.209
M2F-F* w/ m2s happy	1.11	0.3	0.597	0.269	0.047	66.6	0.265
M2F-F* sad	0.96	0.308	0.531	0.363	0.108	77.3	0.257
M2F-F* w/ m2s sad	0.815	0.428	0.609	0.282	0.0371	56.9	0.287
M2F-F* surprised	1.04	0.391	0.576	0.494	0.0759	63.5	0.26
M2F-F* w/ m2s surprised	1.06	0.366	0.602	0.362	0.0278	53.7	0.254

Table 6. **Disentanglement effect.** We feed different input conditions to our versions of Media2Face* models and report the lip-sync metrics for both. *Orig.* denotes an image condition extracted from the video that corresponds to the audio. All other rows are results for denoising with other conditions that come from 7 different emotional images. Media2Face* with THUNDER exhibits superior lip-sync in most cases. Please note that since there is not Ground Truth available for any other setting except for passing the original image (ie. M2F orig.), the metrics are still computed against the original GT. That said, since PLRS is a perceptual score, it does not require GT. However, the consistent improvement in the lip metrics across all condition is still indicative of superior lip-sync.

speech and the perceptual loss appears to get the best of both worlds, while THUNDER-T with mesh-to-speech only is competitive with THUNDER-T with both losses. In any case, applying the mesh-to-speech is beneficial.

7.2.2 Disentanglement

Ideally, M2S should help preserve high-quality lip-sync in the presence of other editing conditions. In other words, the lip-sync should effectively be disentangled from other conditions. To validate this, we evaluate our re-implementation of Media2Face (THUNDERSET-trained) on the THUNDERSET test set, passing additional conditioning inputs extracted from emotional images from unrelated sequences. The conditioning images can be found in Fig. 6. For example results of Media2Face animations, please refer to the supplementary video. We find that training Media2Face with mesh-to-speech improves lip-sync. The metrics can be found in Tab. 6.

7.2.3 Extended Ablation and Sensitivity Analysis

Tab. 7 shows the complete ablation and sensitivity analysis of all THUNDER components on THUNDERSET. Here we give a comprehensive evaluation of our design decisions. (1) The input space for the mesh-to-speech loss, (2) the weight of the mesh-to-speech loss, and (3) the effect if finetuning of the input audio encoder.

(1) **On mesh-to-speech input space.** While all three modalities are comparable, *mouth2s* has scores best on the lip-sync metrics, likely thanks to the localized effect on the mouth.

Exp2s supervision performs slightly worse on lip-sync but results in higher sample diversity. Finally, applying supervision through the whole face to speech (*face2s*), while achieving comparable lip-sync scores, scores worse on diversity, specifically S-DIV-U. The metric can be found in Tab. 7, experiment 1.

(2) **On mesh-to-speech strength.** Applying mesh-to-speech loss, even weakly, results in improved lip-sync metrics (LVE, CCC, PCC and DTW). The stronger the loss, the more the lip-sync metrics improve. However, increasing weights of the mesh-to-speech loss come at the expense of generation diversity (metrics S-DIV, T-DIV). See Tab. 7, experiment 2.

(3) **On finetuning the audio encoder.** Nearly all recent methods finetune the input audio encoder or else they suffer from considerably impaired lip-sync. This results in a certain amount of overfitting to the training subjects’ voices, which manifests as lower diversity scores (S-DIV-U) compared to the models with frozen Wav2Vec (THUNDER-F). In essence, the model’s with trainable audio encoders become less stochastic and more deterministic. Remarkably, the application of the mesh-to-speech loss alleviates the necessity for finetuning Wav2Vec, producing significantly better lip-sync even without finetuning Wav2Vec (THUNDER-T), dramatically reducing the number of training parameters. Applying mesh-to-speech along with finetuning Wav2Vec results in further improvement of lip-sync metrics. Refer to Tab. 7, experiment 3.

Please refer to the supplementary video for visual examples of all the phenomena described above.

7.2.4 Perceptual Study

Here we provide the rest of our comprehensive perceptual evaluation of THUNDER. We run a perceptual study on Amazon Mechanical Turk. The participants are shown two videos side-by-side. One corresponding to THUNDER and one to a baseline. The left-right order is randomized. The

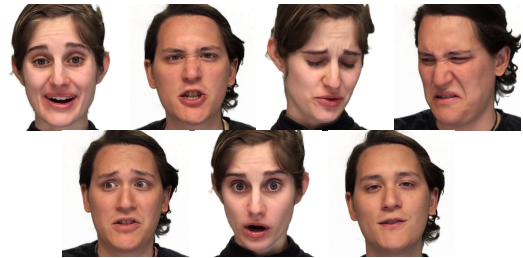


Figure 6. **Media2Face conditioning images.** These images were used as the conditions for the Media2Face* disentanglement experiment in Tab. 6. The images were selected out from the RAVDESS test set. Top row from left to right: happy, angry, sad, disgusted. Bottom row: fearful, surprised, calm.

participants must finish watching both videos before being allowed to rate the videos on a five-point Likert scale (strong preference for left, weak preference for left, indifference, weak preference for right and strong preference for right). We ask the participants to rate three aspects of the animation - lip-sync, dynamism and realism. The exact task description can be seen in the study template in Fig. 13. The participants are shown 20 comparisons, generated from test audios, which were randomly selected from the RAVDESS and TCD-TIMIT test sets. In addition to that, we repeat the first 4 comparisons at the end of the study and discard the first 4 responses. This gives the participants a few examples such that they can adjust to the task. Additionally, we include 4 catch trials, where one animation is selected from Ground Truth and the other animation is clearly wrong. We discard the participants that are wrong on more than one catch trial. The complete results of the perceptual study for THUNDER for lip-sync, dynamism and realism can be found in Figures 7, 8 and 9 respectively. The results for THUNDER-T are given in Figures 10, 11 and 12. The results validate our qualitative and quantitative findings - THUNDER models exhibit better lip-sync, realism and dynamism when compared to baselines.

7.2.5 Comparison to other methods.

Comparing different speech-driven animation methods is a difficult task, especially if trained on a different dataset and especially if the data was acquired in different ways (4D scans, pseudo-GT of different reconstruction methods, etc.). Having a fair apples-to-apples comparisons stemming

from measurements done against the (pseudo-) ground truth is therefore difficult. Regardless, for the sake of completeness, we compare THUNDER to the official releases of recent methods that were also trained on pseudo-GT, namely EMOTE [18] (a deterministic SOTA) and DiffPoseTalk [74] (stochastic SOTA). EMOTE is a FaceFormer-inspired architecture with additional input conditions for emotions and intensity, trained with a content-emotion disentanglement mechanism. It was trained on pseudo-GT reconstructions of MEAD. The pGT was acquired with the same INFERNO tracker, making the resulting metrics comparable. The macro-architecture of DiffPoseTalk’s diffusion is very similar to THUNDER and apart from the additional style conditioning input, it differs only in small details (such as conditioning with the FLAME shape vector, noise schedule, number of denoising steps, etc.). The most important difference is that it was trained on a different dataset, and the reconstructions were produced by a combination of SPECTRE [29] and MICA [100] and hence look qualitatively very different. As such, the DiffPoseTalk metrics likely dominated by the qualitative difference, and are only listed for completeness. Please note that for this experiment, DiffPoseTalk was run without the style condition (to match the setting of THUNDER). Finally, we include a FlameFormer trained on our data, and THUNDER trained without M2S. The results are reported in Tab. 8.

7.2.6 Qualitative Results

In this section we provide additional qualitative results and comparisons. Furthermore, we strongly recommend watch-

		Lip-Sync				Upper-face diversity		Lip diversity		Face Dynamic Dev.	
Experiment	Name	LVE ↓	L-CCC ↑	L-PCC ↑	DTW ↓	S-DIV-U ↑	T-DIV-U ↑	S-DIV-L ↓	T-DIV-L ↓	FDD-U ↓	FDD-L ↓
(1) THUNDER and mesh-to-speech input space	THUNDER w/o m2s	0.879	0.359	0.568	0.329	0.0419	0.044	0.21	0.254	<i>0.0118</i>	0.0932
	THUNDER w/ face2s	<i>0.804</i>	<i>0.411</i>	<i>0.633</i>	0.285	0.0297	<i>0.0409</i>	0.128	<i>0.237</i>	0.0117	<i>0.0827</i>
	THUNDER w/ exp2s	0.83	0.362	0.63	0.296	<i>0.0404</i>	0.0398	0.176	0.228	0.0125	0.0939
	THUNDER w/ mouth2s	0.802	0.426	0.639	0.29	0.0322	0.04	<i>0.134</i>	0.241	0.0122	0.0806
(2) THUNDER-F mesh-to-speech sensitivity analysis	THUNDER-F w/o m2s	0.879	0.359	0.568	0.329	<i>0.0419</i>	<i>0.044</i>	0.21	0.254	0.0118	<i>0.0932</i>
	THUNDER-F $w_{m2s} = 0.1$	0.887	<i>0.375</i>	<i>0.599</i>	0.309	0.0422	0.0445	0.211	0.258	0.0128	0.0936
	THUNDER-F $w_{m2s} = 0.5$	<i>0.805</i>	0.342	0.574	<i>0.294</i>	0.0387	0.0383	0.164	0.225	0.0131	0.103
	THUNDER-F $w_{m2s} = 1.0$	0.802	0.426	0.639	0.29	0.0322	0.04	0.134	0.241	<i>0.0122</i>	0.0806
	THUNDER-F $w_{m2s} = 5.0$	0.858	0.32	0.572	0.313	0.0257	0.0364	0.159	0.23	0.0143	0.105
	THUNDER-F $w_{m2s} = 10.0$	0.886	0.311	0.585	0.31	0.021	0.035	<i>0.146</i>	0.219	0.0155	0.105
(3) THUNDER-T mesh-to-speech sensitivity analysis	THUNDER-T w/o m2s	0.723	0.428	0.623	0.266	0.02	<i>0.0411</i>	<i>0.0656</i>	0.216	0.0118	0.0811
	THUNDER-T $w_{m2s} = 0.1$	0.79	0.417	0.63	0.252	0.0279	0.0431	0.0964	0.221	0.0117	0.0729
	THUNDER-T $w_{m2s} = 0.5$	0.693	<i>0.435</i>	<i>0.641</i>	0.263	<i>0.0221</i>	0.0391	0.0662	<i>0.201</i>	0.0118	<i>0.0782</i>
	THUNDER-T $w_{m2s} = 1.0$	<i>0.709</i>	0.445	0.66	<i>0.256</i>	0.021	0.039	0.0669	0.202	<i>0.0118</i>	0.0788
	THUNDER-T $w_{m2s} = 5.0$	0.844	0.343	0.614	0.262	0.0213	0.0351	0.0997	0.203	0.0139	0.0908
	THUNDER-T $w_{m2s} = 10.0$	0.737	0.408	0.621	0.301	0.0207	0.0344	0.0486	0.199	0.0147	0.089

Table 7. **Ablation study and sensitivity analysis.** Here we analyze the effect of the mesh-to-speech loss on the talking head avatar training. This experiment is conducted on THUNDERSET. The top section compares models supervised with mouth-mesh-to-speech (mouth2s), full-mesh-to-speech (face2s) and flame-to-speech (exp2s). While it is already present in the main paper text, we include it here for completeness. The mid and bottom sections analyze the effect of weight w_{m2s} , with either frozen (-F) or trainable (-T) Wav2Vec2. THUNDER-T models exhibit superior lip-sync performance but it comes at the expense of reduced diversity. Furthermore, for both THUNDER-F and THUNDER-T models, increasing w_{m2s} , while beneficial for lip animation quality, comes at an increasing expense of diversity. Please refer to the Sup. Video for qualitative comparison.

Name	Dataset	Input	Lip-Sync				Sample diversity		Temporal diversity		Face Dynamic Dev.	
			LVE ↓	L-CCC ↑	L-PCC ↑	DTW ↓	S-DIV-U ↑	T-DIV-U ↑	S-DIV-L ↓	T-DIV-L ↑	FDD-U ↓	FDD-L ↓
FlameFormer*	THUNDERSET	audio, one-hot speaker ID	0.809	0.368	0.57	0.291	0.0271	0.0372	0.132	0.239	0.0117	0.0909
EMOTE	MEAD	audio, one-hot speaker ID, emotion and intensity	1.06	0.221	0.442	0.466	0.0563	0.0401	0.248	0.24	0.0141	0.104
DiffPoseTalk	TFHP	audio, shape vector	1.15	0.255	0.38	0.316	0.0345	0.0479	0.226	0.308	0.0141	0.102
THUNDER-F w/o m2s	THUNDERSET	audio only	0.879	0.359	0.568	0.329	0.0419	0.044	0.21	0.254	0.0118	0.0932
THUNDER-F	THUNDERSET	audio only	0.802	0.426	0.639	0.29	0.0322	0.04	0.134	0.241	0.0122	0.0806

Table 8. **Quantitative comparison of THUNDER with other speech-driven avatar methods.** Asterisk* indicates we have re-implemented the baseline and trained it on our dataset. (1) *On lip-sync.* THUNDER outperforms all baselines on lip-sync metrics (LVE, CCC, PCC, DTW). (2) *On expression diversity.* DiffPoseTalk exhibits higher temporal face diversity T-DIV-U and L-DIV-U, likely due to the training data reconstructed from videos in-the-wild, which contains more dynamic motions than our in-the-lab reconstructions. EMOTE exhibits high sample upper face diversity S-DIV-U and expression diversity S-DIV-EXP which is due to the random sampling of speaking/emotion/intensity styles, which while high, may not be fitting to the particular input audio. THUNDER w/o m2s and THUNDER follow next in S-DIV-U. DiffPoseTalk (the official model) performs comparatively low on lip-sync and well on both temporal and sample diversity, but the result is likely skewed by the fact that it was trained on different dataset with different types of reconstruction.

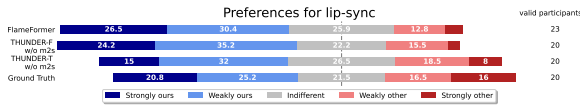


Figure 7. **THUNDER-F perceptual study - lip-sync.** The results of THUNDER-F against other methods. We compare THUNDER-F against methods with both trainable audio encoder (FlameFormer-T, THUNDER-T w/o mesh-to-speech) and frozen encoders (FlameFormer-F, THUNDER-F w/o mesh-to-speech). THUNDER-F outperforms other methods on lip-sync, including those with the trainable audio-encoder (denoted -T).

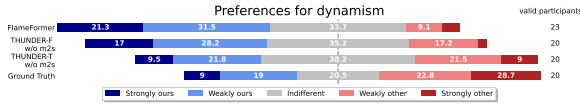


Figure 8. **THUNDER-F perceptual study - dynamism.** The participants found THUNDER-F considerably more dynamic than the deterministic FlameFormer and also more dynamic than THUNDER-F w/o mesh-to-speech. However, ground truth is still rated more dynamic than THUNDER-F.

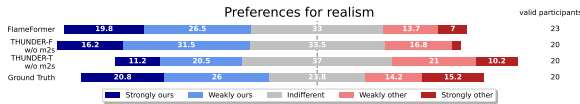


Figure 9. **THUNDER-F perceptual study - realism.** The participants found THUNDER-F considerably more dynamic than the deterministic FlameFormer and also more dynamic than THUNDER-F w/o mesh-to-speech. THUNDER-T w/o mesh-to-speech is rated about as realistic as THUNDER. Remarkably, the participants have preferred THUNDER’s generation over ground truth in terms of realism, which suggests the training dataset has been exhausted.

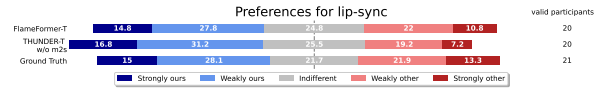


Figure 10. **THUNDER-T perceptual study - lip-sync.** The results of THUNDER-T against other methods. THUNDER-T outperforms its counterpart with mesh-to-speech and the deterministic FlameFormer on lip-sync quality. Additionally, the participants have a slight preference for THUNDER-T over ground truth. This suggests, that THUNDER-T saturates the lip-sync quality of our pseudo-GT.

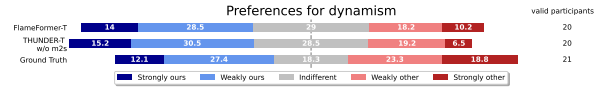


Figure 11. **THUNDER-T perceptual study - dynamism.** The participants rate THUNDER-T to be more dynamic than THUNDER-T without mesh-to-speech and also more dynamic than FlameFormer-T and THUNDER-T w/o mesh-to-speech. Pseudo-GT, however, is preferred over THUNDER-T, suggesting that THUNDER-T’s dynamism did not yet exhaust the dynamism of the training dataset.

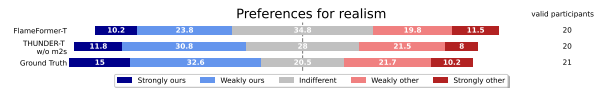


Figure 12. **THUNDER-T perceptual study - realism.** The participants prefer the degree of realism of THUNDER-T’s outputs over the deterministic FlameFormer-T and diffusion-based THUNDER without mesh-to-speech.

ing the supplemental video for an audible and in-motion

qualitative results and comparisons.

Qualitative Comparison on THUNDERSET. Fig. 14 is a

Decide which video has better facial animations

In this task you are presented with two videos of virtual characters animated by two different methods. Their facial motions and lip-sync are generated from the same audio.

Both videos have sound, please listen to them!


You will be asked to select your preference between the two videos according to two different criteria:

- 1. Which video has better synchronization of the lips with the speech?**
Please select the animations have more plausible lip-sync. Please pay attention to:
(1) Focus on the correct articulation of the individual syllables.
(2) Is the mouth opening and closing in sync with the speech?
(3) Is the mouth closed upon pronunciation of sounds like "m", "b" and "p"?
Please factor all of the above into your answer.
- 2. Which video has more lively facial expressions?**
Please select the video where the overall facial motion looks more lively and dynamic.
- 3. Which video has animations that are more realistic?**
Please select the video where the animation is more realistic. Are there any uncanny effects such as unrealistic or over-exaggerated expressions, lip penetration or face twists?

Possible answers:
Choose your preference on the scale from left to right, where the leftmost answer means strong preference for the video A, and the rightmost answer means strong preference for the video B, and the middle radio button means no preference.


Please press play in order to start the videos. **You need to watch and listen both videos at least once to be able to answer.**

video A



video A

video B



video B

1. Which video has better lip-synchronization with the spoken audio?

strong preference for video A
weak preference for video A
equally preferred
weak preference for video B
strong preference for video B

☐
☐
☐
☐
☐

2. Which video has more lively facial expressions?

strong preference for video A
weak preference for video A
equally preferred
weak preference for video B
strong preference for video B

☐
☐
☐
☐
☐

3. Which video is more realistic?

strong preference for video A
weak preference for video A
equally preferred
weak preference for video B
strong preference for video B

☐
☐
☐
☐
☐

To proceed, you must select an answer to all questions!

Submit

Figure 13. The perceptual study web template.

super-set of the figure listed in the main paper (Fig. 5).

Diversity. Input speech may originate from many different expression and even emotions. A well-trained stochastic model should be able to account for that and generate multiple plausible animations, possibly with different facial expressions. Fig. 15 demonstrates THUNDER’s ability to do

so.

Qualitative Comparison on TFHP. Here we provide qualitative results of models trained on TFHP (the same models from Sec. 7.2.1). We select a few utterances from various TFHP test-set sequences. We show results of models trained without head pose in Fig. 16, the *unposed* results of models

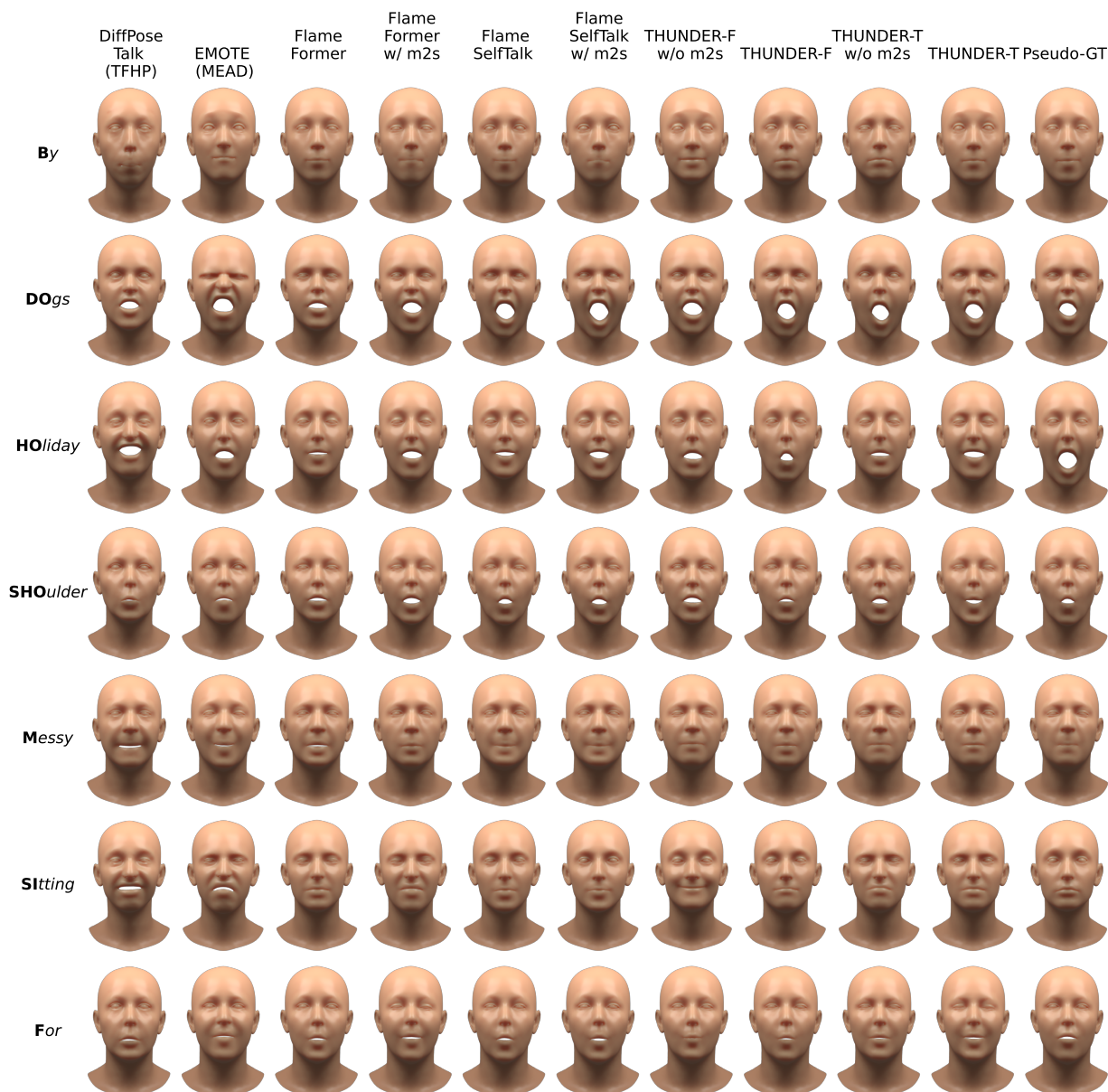


Figure 14. **Qualitative comparison on THUNDERSET.** This figure shows the comparison between baselines, our model and GT for selected utterances. Note that DiffPoseTalk was trained on TFHP and EMOTE on MEAD. Supplemental PDF and video contain more qualitative comparisons.

trained with pose in Fig. 17, and finally the same results with the generated pose in Fig. 18.

Supplementary Video. Our video contains exhaustive set of results which demonstrate the benefits of mesh-to-speech and analysis-by-audio-synthesis. The reader is encouraged to watch the video for our qualitative evaluation.

8. Architecture of Baselines

Fig. 19 shows the architecture of the baselines we reimplemented in this paper.

Media2Face. Our reimplement of Media2Face is built on our diffusion architecture. In training, we provide additional input of ImageBind [32] features extracted from images of the corresponding videos of the training set. Like the audio condition, this condition is dropped 20% of the time.

DiffPoseTalk. Our reimplement of DiffPoseTalk is also built on our diffusion architecture. It is trained in two stages. In the first stage we train the contrastive style encoder as proposed by the authors [74]. In training of DiffPoseTalk, we provide the additional style feature on the input. The

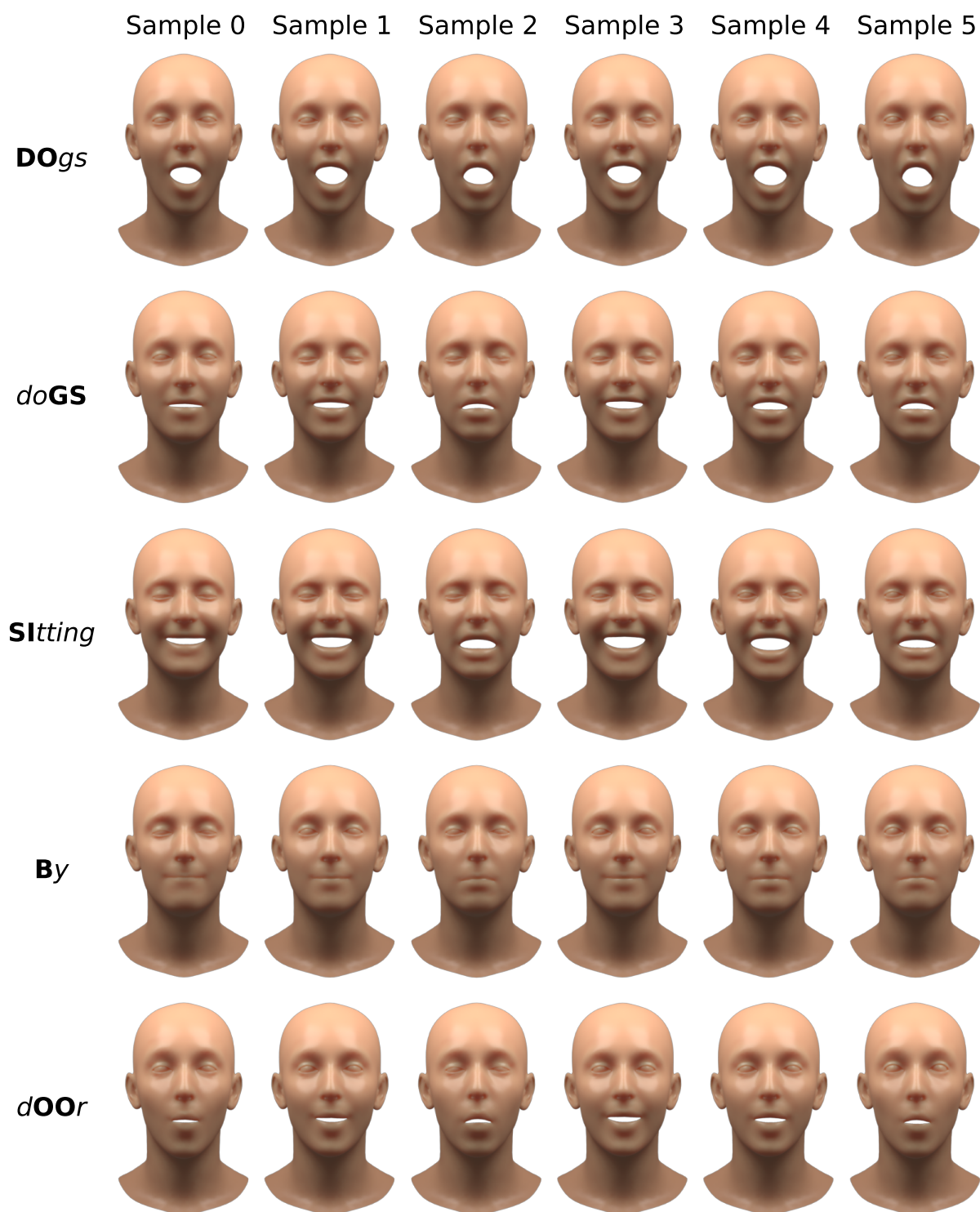


Figure 15. **Diversity of outputs.** THUNDER is capable of generating multiple plausible animations per audio. Each row of this figure contains 6 different generations of the same frame denoised from different initial noise samples (the audio condition remains the same for all).

style feature is extracted from the GT animation sequence.

Like the audio condition, the style condition is dropped 20%

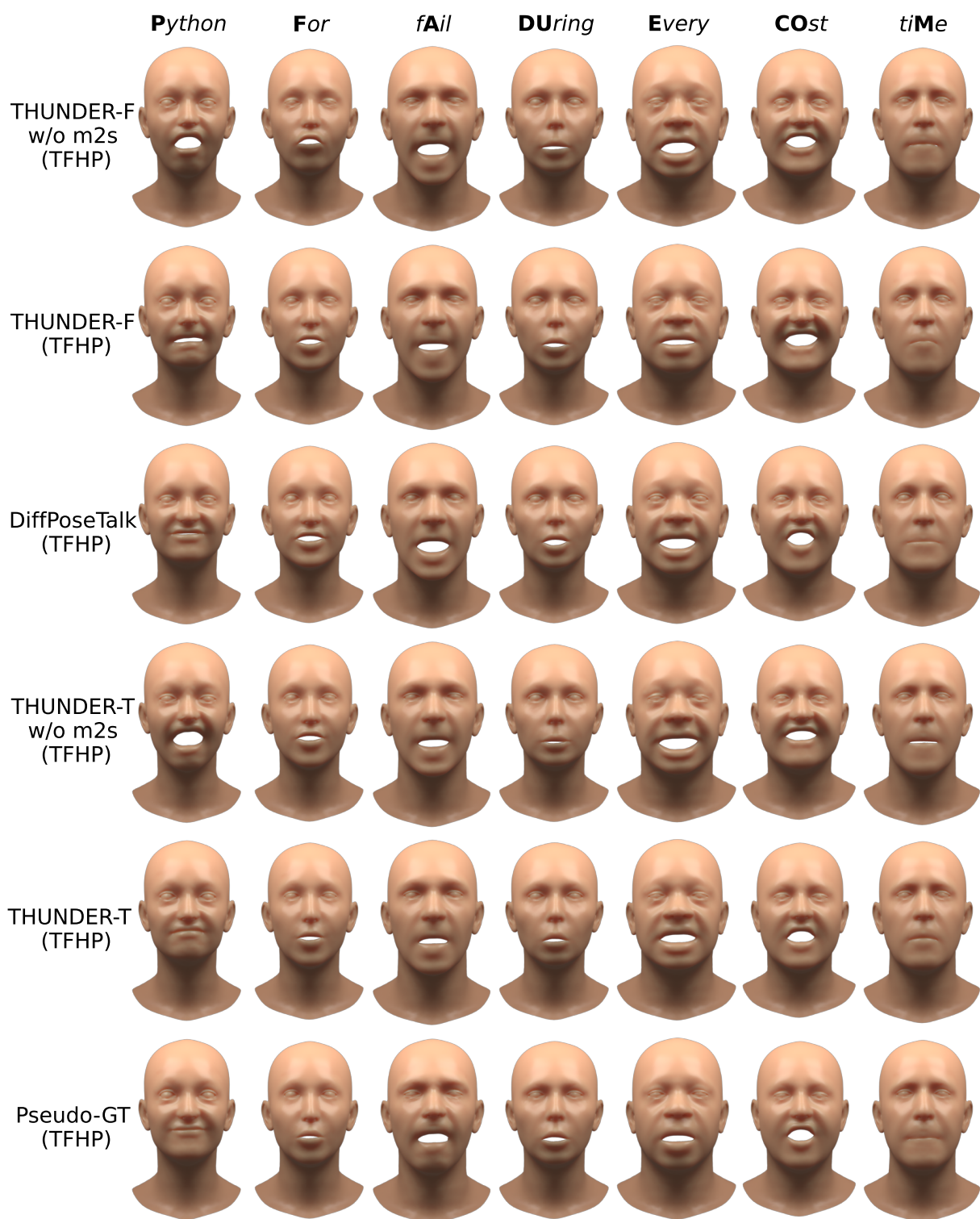


Figure 16. TFHP models trained without head pose.

of the time.

FlameFormer. Our FlameFormer architecture is based

on FaceFormer. Like FaceFormer, FlameFormer is a transformer-based deterministic speech-driven animation

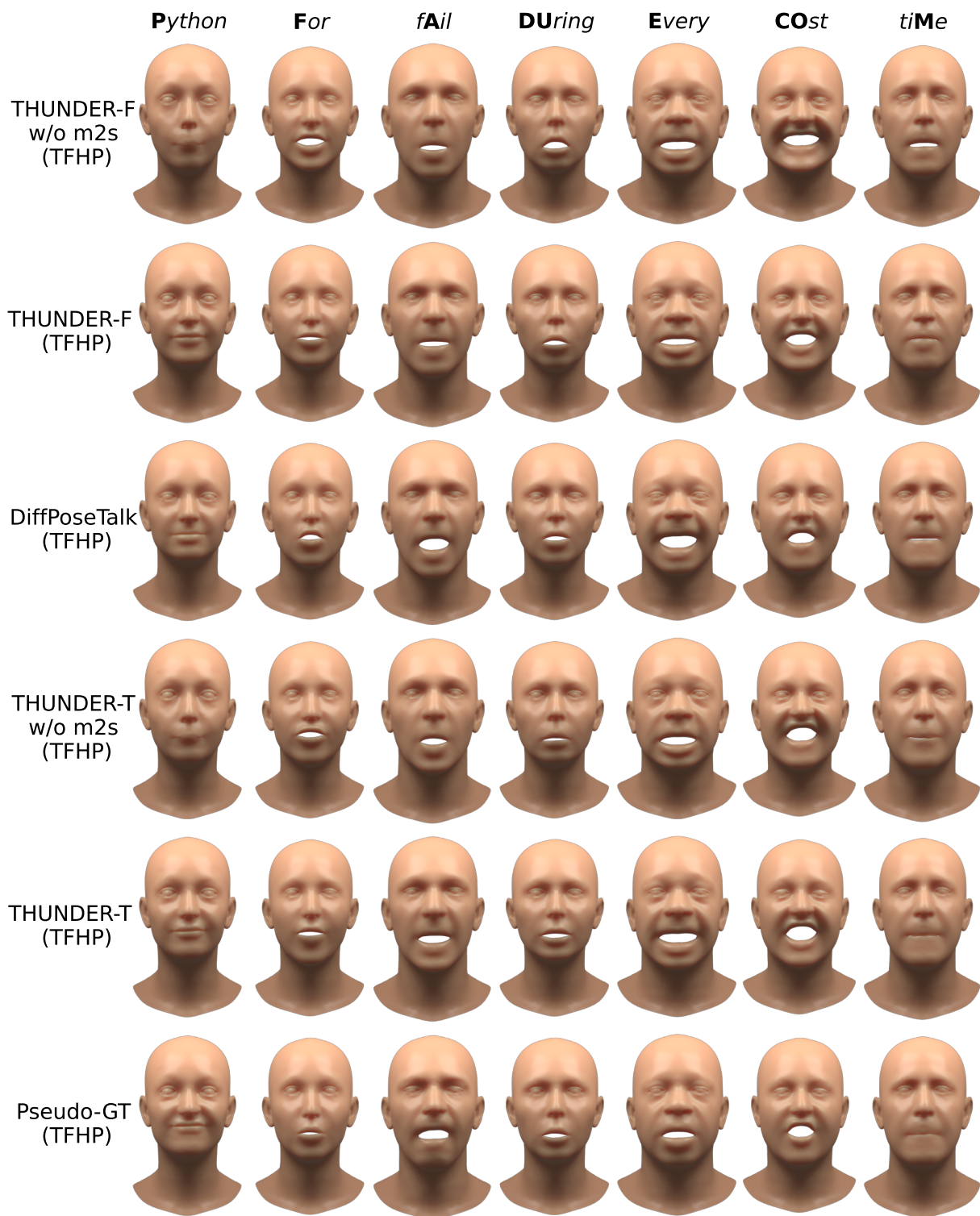


Figure 17. **Results of TFHP-trained models, trained with head pose.** The results are shown in canonical space, i. e. without the generated head pose.

network that takes additional categorical style condition on the input. Instead of predicting the full vertex space like the

original FaceFormer does, FlameFormer predicts FLAME 3DMM coefficients α . Additionally, instead of using the au-

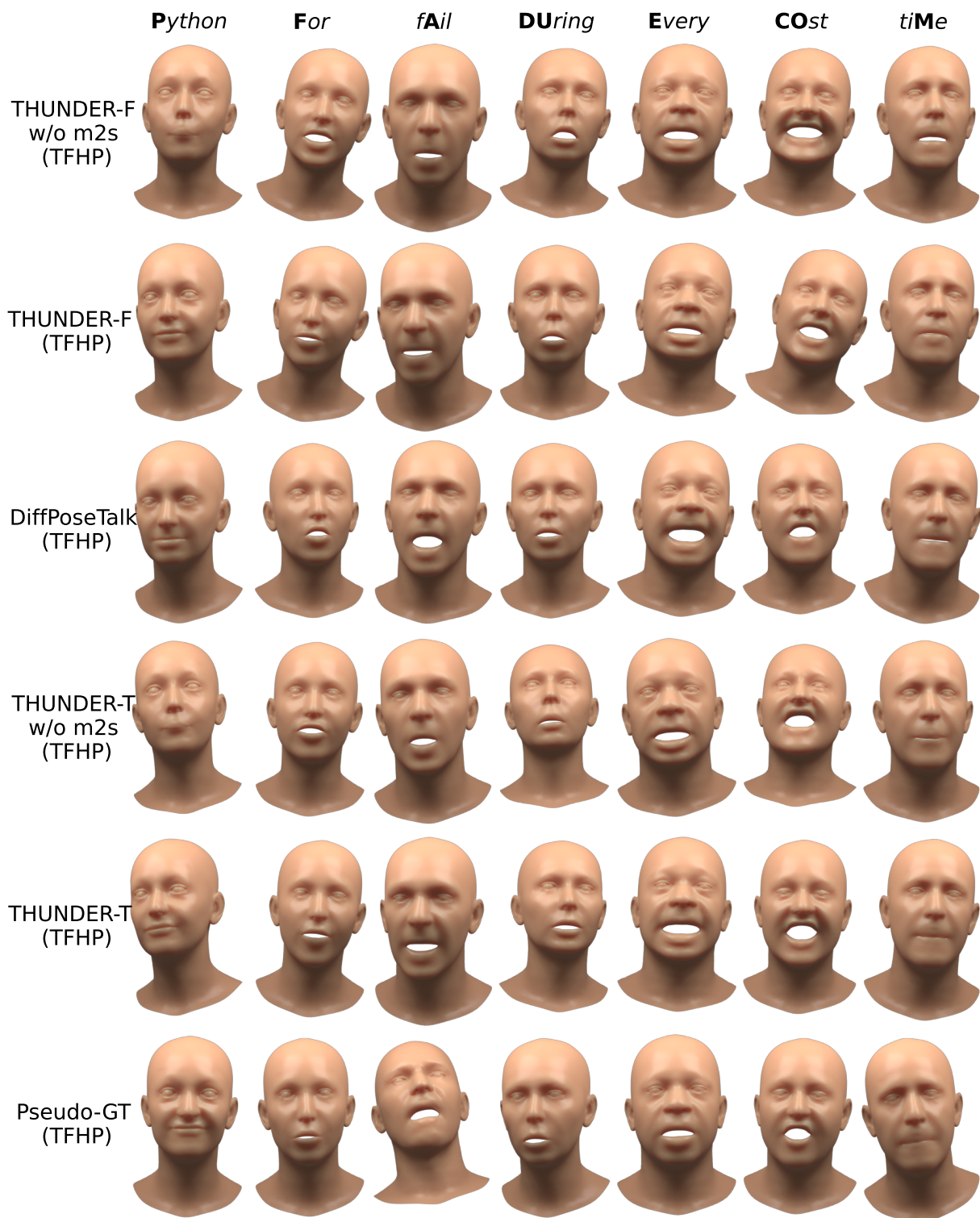


Figure 18. **TFHP models trained with head pose**. The results are shown posed with the predicted pose (or pseudo-GT pose in case of pseudo-GT).

toregressive loop with a transformer decoder, FlameFormer’s decoder is a transformer encoder, which eliminates the need

for the autoregressive formulation. This does not degrade performance and makes the model more efficient, as was

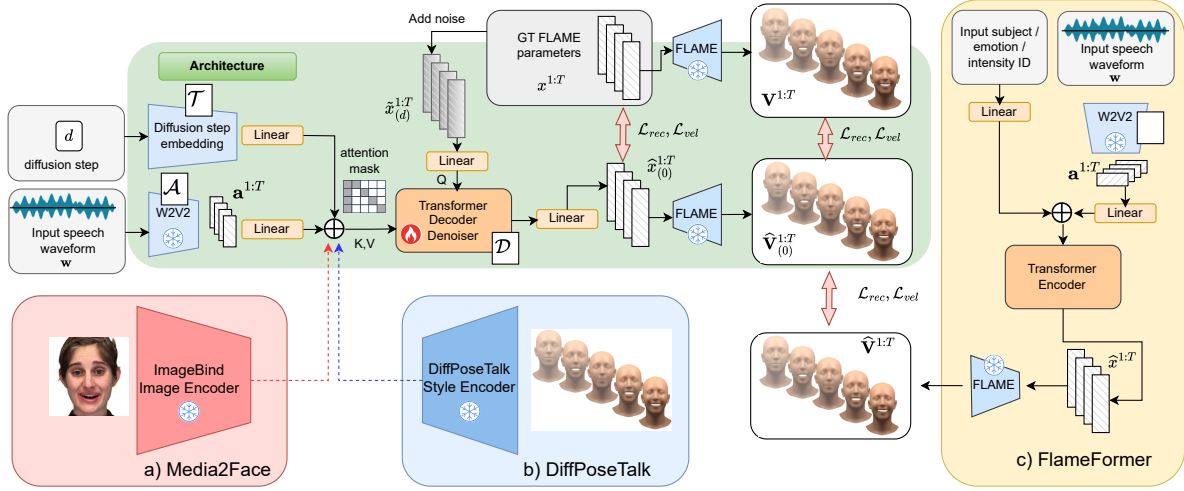


Figure 19. **Architecture of re-implemented baselines.** (a) Our version of Media2Face [98] utilized the same architecture as THUNDER but takes an extra ImageBind feature on the input. (c) Our re-implementation of DiffPoseTalk [74] is also based on THUNDER and takes an extra style feature from a pretrained style encoder. (c) FlameFormer is an adapted FaceFormer [26]. FlameFormer predicts FLAME expression instead of full vertex space. Furthermore, FlameFormer adapts a non-autoregressive BERT-like prediction mechanism (akin to EMOTE [18]).

already shown by Daněček et al. [18], which have proposed the FlameFormer baseline in their paper. Furthermore, we extend FlameFormer’s conditioning inputs - the subject identity (one-hot vector) is complemented with a one-hot vector for 8 basic emotions and three intensities.

9. Data acquisition

In this section we describe the rationale behind our choice of methodology to reconstruct the 3D faces from videos.

9.1. Choice of methodology

Acquiring high quality 4D scans of sufficient scale and richness is costly, time-consuming and requires specialized hardware. Hence, recent speech-driven animation works [18, 74, 95, 98] have turned to pseudo-GT recovered from videos. This choice comes with an important decision - which face reconstruction methodology to use. There are several options but two of the most viable options are:

1) *Optimization-based 3DDM fitting.* Approaches like these (for instance fitting with analysis-by-image-synthesis) have been proposed more than two decades ago [5, 6, 87]. While these options are viable and would produce good results, the optimization-based process is rather inefficient when deployed to many hours of video. This remains true even for the contemporary optimization-based trackers, such as the MICA Metrical Tracker [100] or Pixel3DMM [31].

2) *Off-the-shelf face 3DMM regressors.* Due to the computational intensity of optimization-based fitting and thanks to the recent advance in deep-learning based in-the-wild face reconstruction methods [17, 21, 28, 29, 68, 72, 80–

82, 97, 100]. While this is still an active area of research, and no method produces results that could be considered as accurate as 4D scans, recent years have brought enough advancement so that they can be employed for speech-driven animation dataset construction.

9.2. Discussion of existing FLAME regressors

Since THUNDER utilizes FLAME [51], our discussion will focus on FLAME regressors (trackers) only. There are several systems that have been used for talking head avatar research. No tracker is perfect. Each tracker produces different kinds of errors but despite that, the recent trackers are finally good enough to be used for talking head avatar research. In fact, a few previous methods have already made use of them [18, 55, 74]. Here we briefly discuss some of the most applicable candidates:

DECA [28] is the first FLAME-based regressor trained with the self-supervised “analysis-by-image-synthesis” loop. While this paper was SOTA at the time of its release, it still had considerable artifacts, especially when it comes to the richness of facial expression and emotions and quality of lip animations. Furthermore, the stability of the identity prediction was not stable when applied to videos.

EMOCA [17] is a follow-up of DECA, capable of reconstructing rich emotions thanks to its emotion-consistency loss. However, it can exaggerate expressions and the lip-sync is not good enough. EMOCaV2 later addressed the lip-sync issue by incorporating the perceptual lip reading loss from SPECTRE, obtaining the best of both worlds. However, similarly to DECA and SPECTRE, EMOCaV2 exhibits unstable identity prediction over the course of a video.

SPECTRE [29] applies a paradigm similar to EMOCA but leverages a lip-reading consistency loss instead. It produces good lip-readable animations, but unfortunately often exaggerates and over-articulates, is not capable of reconstructing rich emotional expressions and the disentanglement between expression and identity is rather poor, which often makes the predicted expression vector compensate the inaccuracies in identity predictions.

EMICA [19] is a combination of several methods' contributions. It combines the benefits of DECA, EMOCA and SPECTRE to achieve high quality emotions and lip-sync. Additionally, EMICA makes use of MICA [100] for a consistent identity prediction. The system is trained jointly and publicly available.

SMIRK [68] SMIRK is a recent reconstruction method which uses a neural renderer and in-the-loop neural data augmentation to produce high quality reconstructions.

9.3. Data acquisition in previous works

As a result of the tremendous progress of in-the-wild face reconstruction, some of the recent SOTA speech-driven animation methods have already turned to pseudo-GT as the primary source of data. For instance:

EMOTE [18] employed the EMICA model to reconstruct the MEAD [89] dataset.

Yang et al. [95] utilized both DECA and SPECTRE to reconstruct LRS3 [1].

DiffPoseTalk [74] employed MICA and SPECTRE separately and then combined the shape predictions of MICA and expression predictions of SPECTRE to reconstruct the TFHP dataset.

We opt for EMICA as it provides a good compromise - rich emotions, good lip animations, consistent shape identity and temporal consistency of the identity prediction. While not artifact-free, EMICA is a good option for building a large-scale dataset for 3D talking head avatar research.