How Video Meetings Change Your Expression —Supplementary Material—

Video visualizations of qualitative results can be found at the following link: facet.cs.columbia.edu The supplementary material is structured as follows: in Sec. A we provide dataset preparation details, and dataset statistics. In Sec. B we provide additional discussion on the metrics used in our paper for evaluating our models. In Sec. C we provide a note on the baselines. In Sec. D we provide further implementation details which we hope can be useful for reproducing the results, and also provide more information on how the fixed translator-set baseline was implemented. In Sec. E we provide some more ablation results. Finally, in Sec. F, we provide additional insights about the differences between F2F and VC through `FacET`, and in Sec. G, we provide latent distributions and dataset-level statistics for the presidents data.

A Dataset details

To study the differences between VC and F2F conversations, we collect the ZoomIn dataset, containing 240 hours of conversations in both VC and F2F settings. We obtain these videos from the YouTube channel "The Daily Talk Show - Australian Podcast", where the hosts (mostly two specific individuals) have conversations on a wide range of topics. Using background and eye gaze as cues, we are able to annotate the mode of conversation. We then run MediaPipe face landmark detector [1] on all the frames to obtain 3D facial keypoints. Some examples can be found in Fig. 1. Each sample clip in our dataset consists of detections on contiguous 176 frames from both the participants. In cases where the frame rate is higher, we apply linear interpolation on the keypoints to bring it down to 25 FPS. For the purpose of our experiments, we only consider those conversations where there are only two people.

Most of the videos in the ZoomIn dataset are of 1920 x 1080 resolution. We show some important statistics about our ZoomIn dataset in Tab. 2. We highlight two positive aspects about this data. First, the conversations that we call as video conferencing (VC) do not suffer from any quality issues of the second party. This

Table 2: Key statistics about the ZoomIn dataset.

Mode of conversation	Number of videos	Average length of videos (minutes)	Total duration of videos (hours)
Video conferencing	89	38.14	56.58
Face to face	279	39.44	184.70
Total	368	39.13	241.28

is because both the participants record their videos on their own systems, and then collate it together for best visual quality for their viewers. Second, in all the conversations, the cameras remains fixed. Hence both the participants have a camera in front of them at most times, and we can view the facial expressions of both the participants simultaneously during the course of the conversation.

Additionally, to study differences in communication styles across subjects, we collect a second dataset of presidential speaking styles. Specifically, we collect videos of statements and announcements by U.S. presidents Obama and Trump. We would like to highlight that the presidents dataset is much smaller (around $10\times$) as compared to the ZoomIn dataset, and suffers from issues like inconsistent and moving camera positions. Also, we perform undersampling to overcome the huge class imbalance in the presidential dataset.

To the best of our knowledge, this is a *first-of-its-kind* dataset that contains information about the variation of the mode of conversation between VC and F2F conversations. Please refer to facet.cs.columbia.edu for more on data.

B Metrics

Evaluating discovered differences is extremely challenging in settings without supervision as we cannot evaluate what will be discovered beforehand. Additionally, to the best of our knowledge, there are no existing prior works that tackle this problem of discovering differences from spatio-temporal keypoint data. Existing generative research (e.g., GANs) uses human evaluation or inception score. Firstly, recall that our task is to discover differences *without* access to labels for such differences. Since humans also do not have prior expert knowledge, we cannot run human studies. Secondly, our dataset (containing only 2 classes) has a few dominant patterns distinguishing them (e.g., 'Head Pitch'). As we show in Fig. 2, training a classifier directly results in good accuracy by only picking up on a few dominant features. Thus, getting a high or low inception score is also not a useful metric.

Therefore we evaluate how various design choices of our interpretable-by-design model result in finding good differences. The key metric we use to evaluate this is how well the translation function G_{XY} fools the discriminator while being interpretable. So we measure the the discriminator accuracy in distinguishing real samples in a domain versus the samples translated to that domain. A model that leads to *lower discriminator accuracy* can capture more modes of variations and is therefore a better translator. We hypothesize that in the long term for a discriminator to perform well, it needs to identify differences that our constrained generator cannot produce, but can be used by the discriminator for detecting fake samples. Those differences that the generator can learn to rectify, reach a state of equilibrium with the discriminator. Hence we find this metric to stabilize after a certain set of epochs, and therefore we provide the average over the last 100 epochs.

We find the accuracy of the discriminator trained at the *original* task of detecting video-calls (VC) vs face-to-face (F2F) conversations to be 99%, showing

Table 3: Statistical analysis for Tab. 1 (for $c = 2$ on ZoomIn dataset) with 95% confidence intervals from 8 runs.

Model Type		**Disc. Acc. (%) ↓**		
G_f	G_t (chunks)	Avg	F2F→VC	VC→F2F
Fixed	Fixed-size	94.24 ± 0.84	94.94 ± 0.74	93.57 ± 1.54
	Var.	92.51 ± 0.30	93.63 ± 0.63	91.50 ± 0.77
Pred.	Fixed-size	80.81 ± 1.33	80.38 ± 2.61	81.10 ± 2.64
	Var. (**FacET**)	**74.11 ± 0.89**	**74.70 ± 1.81**	**74.26 ± 1.01**

that the discriminator is strong and able to perform this task well. The discriminator accuracy of 73.16% (Tab. 1) shows that our method is able to fool the discriminator and that our method is able to detect and remove the differences between the domains. We also provide statistical analysis for Tab. 1 in Tab. 3.

C Baselines

Note that **Fixed translator-set** is a variation on G_f, whereas **No Partitions** and **Fixed translator-set** are variations on G_t. Therefore we can mix and match them for evaluation as well.

D Implementation Details

Our input data consists of by 478 3D facial keypoints. For the spatial β-VAE, we use a 5-layer MLP (with a LeakyReLU non-linearity) as the encoder as well as the decoder (with layer dimensions as 512, 512, 256, 256, 128; similarly in the reverse direction for the decoder). We begin training from solely using the training reconstruction loss (i.e. $\beta = 0$) and gradually increase it. We choose the highest β that does not compromise on reconstruction. We train the β-VAE with 16 latent codes but find that only 12 of them show any variation ($l = 12$), capturing various dimensions of expressions that can be seen in Fig. 1. We measured this by varying a latent between $(-3, +3)$, and measuring the mean L2 distance between reconstruction keypoints. For a β-VAE (with l=16), we observe that the 12 useful dimensions have a variation at least *138 times* the remaining 4. Lastly, we use Adam optimizer with a learning rate of 0.001.

For the translation function, we use 3-layer MLP (LeakyReLU non-linearity) for G_f and G_t. The discriminator is also a 3-layer MLP (dropout of 0.5). We use a factor of 8 to decrease the dimensionality in the subsequent layers as needed. We use Adam optimizer with a learning rate of 10^{-4}. We train for 1000 epochs on *ZoomIn* and for 2000 epochs on the presidents dataset. Our data is split in 90:10 train-test ratio. We train using an Nvidia RTX 2080Ti. For the translator function G_t, we explored different values of the temperature Q, with no impact

on performance. We set $Q = 0.12$ for our experiments. For training vid2vid (domain transfer), we use 8 Nvidia A6000 GPUs, and train for 50 epochs.

D.1 Fixed translator-set baseline implementation details

Instead of learning to predict a *translator* from the input, the fixed translator-set baseline instead chooses from a list of possible *translator*s. As stated in the main paper, this alternative model learns two things. First, is learns a classifier that chooses one out of p choices. Second, it also simultaneously learns a list of parameters to choose from $P \in \mathbb{R}^{p \times 2l}$.

Even in this case, the model is non-differentiable as the *choosing* operation using the output of the classifier is non-differentiable. However, we can approximate the arg max operation in the classifier with a softmax to make it differentiable. An additional issue with this approach is that, since there is no explicit supervision for the model to always predict a single choice out of p, the model cannot learn to predict a single high-confidence choice. Therefore, we additionally use an entropy loss term to enforce the model to produce *peaky* or high-confidence choices. If $s \in \mathbb{R}^p$ is the softmax probability score for each choice, the entropy loss can be written as:

$$\mathcal{L}_{\text{entropy}=-\sum_{i=1}^{p} s_i \log(s_i)}$$

The predicted parameter values can be written as the weighted sum of the p choices.

$$\omega^* = \sum_{i=1}^{p} s_i \omega_i \qquad\qquad \phi^* = \phi_{i=1}^{p} s_i \phi_i$$

Where, ω_i and ϕ_i are the i^{th} choice for multiplicative and additive factors. As the training progresses, the entropy loss forces the model to select only one out of p choices. The translation function G_{XY} can be be written as,

$$G_{XY}(z_x) = \omega^* \odot z_x + \phi^*$$

The entropy loss term is used along with the main adversarial loss term leading to a solution for the classifier and the learned list of parameters. In our experiments, we used $p = 32$.

E Ablations on the *translator*

To study the importance of each component of our *translator*, we perform an ablation by removing each component systematically. The results are shown in Tab. 4. In rows 1 & 2, the *translator* predicts the scaling factor ω only (no shift), denoted by Var. Chunks ($\times$). In rows 3 & 4, the *translator* predicts the shift ϕ only (no scaling), denoted by Var. Chunks ($+$).

For the ZoomIn dataset, comparing Var. Chunks ($\times$) with Var. Chunks ($+$), we observe that shift (ϕ) is consistently more important than scale (ω), as removing it results in worse performance than removing scale. However, this trend is

Table 4: Ablations over shifting and scaling in the *translator*.

	Model Type		**Discriminator Acc. (%) ↓**					
			ZoomIn			Presidents		
G_f	G_t	c	Avg	F2F→VC	VC→F2F	Avg	O→T	T→O
Predicted translator	Var. Chunks (×)	2	86.50	86.71	86.31	92.03	88.52	95.33
	Var. Chunks (×)	7	86.29	86.69	85.98	91.82	89.86	93.44
	Var. Chunks (+)	2	81.72	81.91	81.56	96.52	96.22	97.05
	Var. Chunks (+)	7	81.70	82.16	81.34	96.51	96.14	97.08
	Var. Chunks (**FacET**)	2	73.28	73.33	**73.50**	79.35	70.31	**88.70**
	Var. Chunks (**FacET**)	7	**73.16**	**72.65**	73.92	**78.14**	**67.50**	89.06

reversed for the Presidents dataset, where scale (ω) is more important than shift (ϕ). FacET consistently outperforms both these ablations across both datasets, highlighting the need for predicting both shift and scale to ensure good domain transfer.

F ZoomIn Observations

We present additional insights that FacET can discover about the differences in face-to-face and video conversations. Please refer to Fig. 5 and Fig. 3.

People move their head more during a F2F conversation. See Fig. 5 first two rows (or clusters). From the latent code #9 and #10 in both clusters (speaking with a laugh; listening with neutral expression) we notice that people move their heads a lot more in an F2F conversation. Note that this is different from saying "People can be facing at different points with respect to the camera in an F2F conversation". The latter observation can be made by the bimodal F2F distribution of latent #9 and #10. The F2F distribution is more spread out than the VC distribution. This suggests that not only are the views shifted w.r.t. the camera in F2F conversation, but the head also moves more.

People open their mouth more while listening in a F2F conversation. See Fig. 5 row 2 (cluster: "listening with a neutral expression"). This can be inferred by looking at latent #3 as well as the translated conversation where the original VC shows pursed lips while the translated F2F conversations shows slightly more opened lips. This is consistent with observations in neuroscience which show that people are more self-conscious due to the presence of a self-viewing screen in video chats, leading to increased self-regulating of facial expressions [2].

Eyebrow raises are bigger in VC. See Fig. 3. This again follows through the fact that "people emote bigger during in a VC" [3]. In the speaking with a raised eyebrow cluster, we observe that the VC distribution shifted towards the right (bigger eyebrow raise) for latent #11. This suggests that even though

people raise eyebrows in both modes of conversation, the magnitude of the raise is larger during a VC conversation.

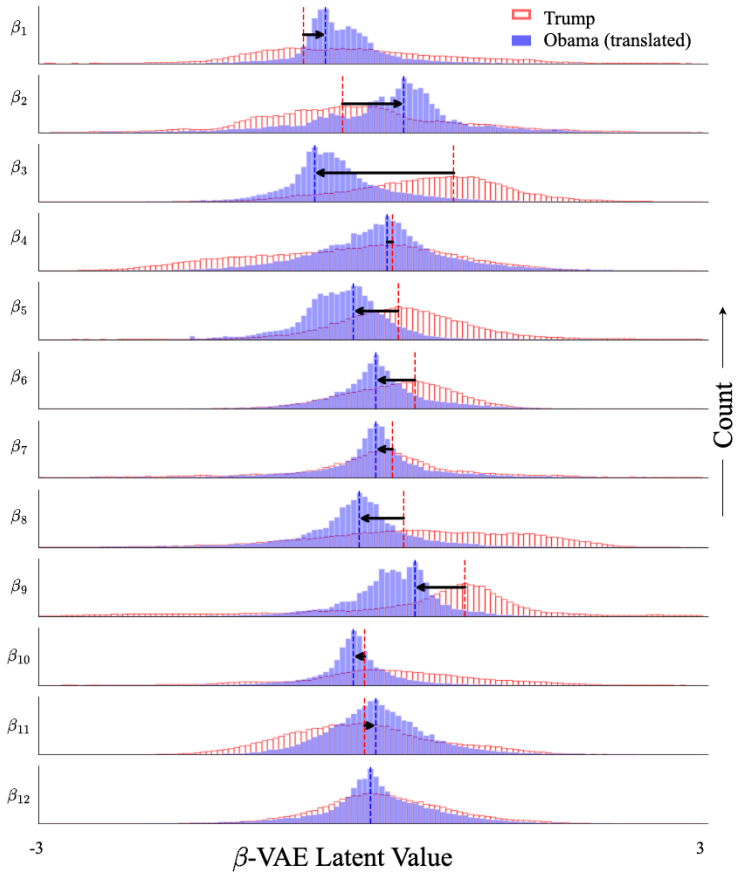

Fig. 9: Latents for presidents data. We vary each dimension of the 12-dimensional latent obtained through β-VAE encoding by keeping other dimensions fixed (rows). We show the dataset-level statistics for the desired latent, while the dotted line shows the modes. The direction and length of the arrows show the extent to which different latents change across domains on average. Please refer to facet.cs.columbia.edu to visualize the latents. This is analogous to Fig. 3 (which was for ZoomIn dataset).

Eyes are more closed in VC conversation. See Fig. 3. Although the latent codes #3 and #12 are not fully disentangled, we observe that going right on both corresponds to closing eyes. We observe that people close their eyes more during a VC as the F2F distribution shifts to the left. This can be partially attributed to zoom fatigue [4, 5], however, a deeper understanding is needed to better understand this. We cannot examine whether this is due to more blinking

or just more closed eyes. Our model lacks a disentangled latent for eyelid and eye movement, making it harder to delve deeper. We believe that better disentanglement (e.g., by using eye-specific loss-terms) and better keypoint tracking can lead to a more in-depth examination of this observation.

G Presidents Dataset Stats

Similar to Fig. 3 where we provided dataset-level statistics for the ZoomIn dataset, we also provide them for the Presidents data for completeness in Fig. 9.

References

1. C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, *et al.*, "Mediapipe: A framework for perceiving and processing reality," in *CVPR Workshop*, 2019. 1
2. S. Y. Shin, E. Ulusoy, K. Earle, G. Bente, and B. Van Der Heide, "The effects of self-viewing in video chat during interpersonal work conversations," *Journal of Computer-Mediated Communication*, vol. 28, p. zmac028, 11 2022. 5
3. F. Hill, "The gesture that encapsulates remote-work life," *The Atlantic*, July 20 2023. 5
4. H. Nesher Shoshan and W. Wehrt, "Understanding "zoom fatigue": A mixed-method approach," *Applied Psychology*, 2022. 6
5. G. Fauville, M. Luo, A. C. Queiroz, J. Bailenson, and J. Hancock, "Zoom exhaustion & fatigue scale," *SSRN Electronic Journal*, 2021. 6