How Video Meetings Change Your Expression

Sumit Sarin, Utkarsh Mall, Purva Tendulkar, and Carl Vondrick

Columbia University
{ss6712,um2171,pt2578,cv2428}@columbia.edu

Abstract. Do our facial expressions change when we speak over video calls? Given two unpaired sets of videos of people, we seek to automatically find spatio-temporal patterns that are distinctive of each set. Existing methods use discriminative approaches and perform post-hoc explainability analysis. Such methods are insufficient as they are unable to provide insights beyond obvious dataset biases, and the explanations are useful only if humans themselves are good at the task. Instead, we tackle the problem through the lens of generative domain translation: our method generates a detailed report of learned, input-dependent spatio-temporal features and the extent to which they vary between the domains. We demonstrate that our method can *discover* behavioral differences between conversing face-to-face (F2F) and on video-calls (VCs). We also show the applicability of our method on discovering differences in presidential communication styles. Additionally, we are able to predict temporal change-points in videos that decouple expressions in an *unsupervised* way, and increase the interpretability and usefulness of our model. Finally, our method, being generative, can be used to transform a video call to appear as if it were recorded in a F2F setting. Experiments and visualizations show our approach is able to discover a range of behaviors, taking a step towards deeper understanding of human behaviors. Video results, code and data can be found at facet.cs.columbia.edu.

Keywords: Facial Expressions · Interpretability · Video Conferencing

1 Introduction

Have you wondered how your facial expressions change when speaking to someone over a video call (VC) as compared to speaking face-to-face (F2F) (Fig. 1)? Studies have shown that there are significant differences between these two modes of communication – in terms of overall effectiveness, rapport building, cognitive load, energy consumption, etc. – which can manifest itself in changed expression patterns [1–5]. Since the recent intervention of COVID-19, our lives drastically changed as VC became the primary source of communication, resulting in unprecedented social phenomena such as *zoom fatigue* [6–10]. It is important to systematically study these changed patterns in order to build tools that can cope with such interventions – e.g., by evaluating and improving AR/VR technology so that our virtual experiences are as similar to reality as possible. Moreover, studying changes in human behavior, at an individual-level, as well as across

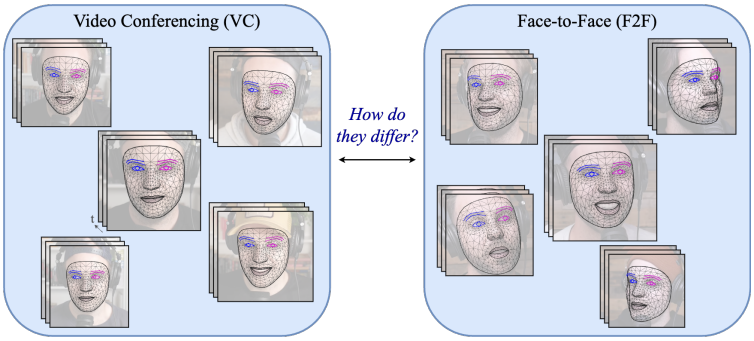

Fig. 1: What is the difference between the two domains? Given two unpaired sets of videos of persons speaking on VC (left) and F2F (right), our goal is to provide interpretable insights on how the motion sequence differs between the domains. See Figs. 3 & 5 for a detailed report generated by our approach.

a population, can be useful in a variety of disciplines including anthropology, sociology, as well as cognitive and social psychology [11–13].

Building models to understand facial expression changes in conversations is challenging for two key reasons. *First* is the dataset bias. Internet videos are an abundant source of natural and diverse data [14]. However, we find that internet videos of people on VC and F2F are heavily biased. VC often features direct camera gazes, while F2F shows side views. Such biases are not *inherently* harmful, but models that simply identify them are not sufficient to *discover* all the differences. Unsurprisingly, it is quite easy to build a classifier that can differentiate the two domains. Fig. 2 shows the weights of a linear classifier trained on disentangled facial features (see Sec. 3.1). With no temporal information, the classifier attains 88% accuracy on single frames alone! Moreover, these biases are implicit and apriori unbeknownst to us, making them near impossible to remove.

Second, our task is not simply that of classification, but rather *discovery of domain-specific differences.* We lack prior knowledge about the domains, and we wish to understand them better. Being able to predict the domain is perhaps less interesting than being able to understand the subtle spatio-temporal differences of facial expressions across the two domains. Much of the current literature focuses on post-hoc explainability of black-boxes [15–20]. However, such approaches are unsuitable for our goals because discriminative models (such as Fig. 2) trained on biased data only pick up these biases. Additionally, the explanations are useful only if humans themselves are good at the task.

We present `FacET`, a general-purpose framework to discover interpretable, spatio-temporal trends between two domains, that works even in the presence of dominant biases. In contrast to discriminative methods, `FacET` employs a generative approach to discover the differences. To study the aforementioned task of VC and F2F conversations, we first collect a large video dataset (240 hours) of such conversations, called the ZoomIn dataset. In addition, we also collect a second dataset of speaking styles of US presidents. We apply `FacET` on

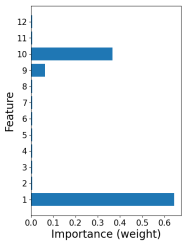

Fig. 2: Insufficiency of discriminative methods. We train a simple linear classifier on disentangled facial features to distinguish VC and F2F conversations (refer to Fig. 3 for the meanings of these features). We observe 88% classification accuracy even when using frames without the temporal information, with 'Head Pitch' (#1) and 'Head Tilt' (#10) being the most dominant features. A post-hoc explainability model cannot explain this discriminative model as it is trained on biased data.

these datasets, and reveal novel insights – e.g., we observe that speakers tend to laugh *bigger* in F2F as compared to VC. We also observe that President Trump raises his eyebrows more while listening compared to President Obama. Moreover, `FacET` has applications beyond revealing these insights – we show that `FacET` can be used to perform domain transfer – e.g., we can modify VC videos to look more like a F2F conversation (effectively "de-zoomifying" a video).

2 Related Work

We discuss the landscape of interpretability in computer vision by describing post-hoc explanations, interpretable architectures, as well as disentangled representations. We then discuss efforts in human facial expression understanding.

Interpretability through post-hoc explanations. Explainable AI (XAI) [21] methods seek to explain model decisions through visualizations [15–20], counterfactual explanations [22–24], feature importance [25] and sample importance [26–32] methods. These explanations are important to ensure algorithmic fairness [33], identify potential bias in the training data [34–36], and to ensure that the algorithms perform as expected [37–40]. However, post-hoc explanations are only meaningful if humans are good at solving these tasks and can check the model's explanation against their own understanding. Also they can only explain what the model has learned, which is problematic if the data is biased. They are more useful for explaining model decisions than data patterns, which is our focus.

Interpretability by design. A number of works have defined interpretability through a set of linear transforms between classes [41–46]. However these methods are limited to discriminative tasks where classes are distinguishable by humans. Generalized Additive Model [47] and its variants [48–52] are another class of methods which non-linearly transform input features separately. However, these methods suffer from complex training procedures and combinatorial explosion. Our method also falls in this category of interpretability-by-design, and leverages linear transformations in a generative translation setting.

Learning Disentangled Representations. Learning factorized representations to capture independent variations in data can correspond to human-defined

concepts and lead to interpretable models [53]. Learning disentangled representations using supervised data (apriori knowledge of the nature of generative factors) has been explored by many methods [54–57]. However, for practical applications where there is no supervision available for discovering the generative factors underlying the data, purely unsupervised approaches that use GANs [58–63] and VAEs [64–68] for learning disentangled representations are more useful. We use β-VAE [64] which is one such popular unsupervised method.

Facial Expressions Understanding. Existing literature on human facial expression understanding is limited to classifying emotions into known categories [69–75], or their combinations [76]. Several works have also studied person-specific speaking styles [69, 77] which link facial expression changes to personality traits. However, the continuous nature of facial expressions is not accurately modeled through such discrete representations. Recent works have also studied how a person's facial expressions change in response to the person they are interacting with [14, 78–80]. Yet, none of these works provide interpretable insights regarding inter or intra-subject patterns under varied conditions.

3 Method

A conventional approach to finding interpretable differences between two domains is to learn a discriminative model to distinguish and analyze them using post-hoc tools such as GradCAM [17]. However, discriminative models tend to focus on the first few significant modes of differences. Consequently, we cannot find all the important differences. This problem is further exacerbated if the dataset is biased, and significant differences appear only due to such biases. We posit that a generative model that can transform one domain to another should capture all possible modes of differences. Such a generative model can then be leveraged to analyse the different modes and their effect on the two domains.

We present `FacET` (**Fac**ial **E**xplanations through **T**ranslations). Given spatio-temporal facial keypoint data from two domains X and Y, `FacET` learns a translation function $G_{XY} : X \to Y$, that can convert samples from domain X to appear as if they were from domain Y. Our key insight is to *learn piecewise shift and scale transformations* [41–46] between X and Y which results in G_{XY} being interpretable, allowing us to explain the modes of differences between the two domains. To train `FacET`, we first learn a per-frame spatial representation to disentangle facial features (Sec. 3.1). Then, we train an interpretable G_{XY} on these features capturing spatio-temporal differences between the domains (Sec. 3.2). We use the trained G_{XY} to generate detailed reports (Sec. 3.3) highlighting key differences between the two domains. Finally, we showcase `FacET`'s ability to perform domain translation, e.g., by "de-zoomifying" a VC video (Sec. 3.4).

3.1 Disentangling Spatial Features

The first step towards an interpretable translation model is learning a disentangled spatial representation. We found that disentangled spatio-temporal features

are difficult to interpret for humans. Thus, we first learn to disentangle spatial features only. β-VAEs [64] have been found to generate disentangled representations that are useful in discovering patterns in a variety of domains [81–85]. We train a β-VAE on frame-level facial keypoints of our datasets. We obtain a spatial feature $z \in \mathbb{R}^l$ for a data point d, where d is a single frame in the clips of our dataset $X \cup Y$. Here l is the dimension of latent space. This can be achieved by learning an encoder (q) and decoder (p) by minimizing the β-VAE objective,

$$\mathcal{L}(X \cup Y) = -\mathbb{E}_{q(z|d)}[\log p(d|z)] + \beta D_{KL}(q(z|d)||p(z)) \tag{1}$$

The first term (marginal log-likelihood) optimizes for reconstruction quality, while the second term (KL-divergence) enforces disentanglement in the latent space z. Minimizing the KL divergence between the latent vectors and a unit Gaussian prior ($p(z) = \mathcal{N}(0, \mathbf{I})$), results in disentangled latent vectors. Higher β values result in better disentanglement at the cost of reconstruction.

3.2 Translation Function

We learn the translation function G_{XY} in the β-VAE's spatial latent space z. We aim to learn a G_{XY} that can potentially capture all differences between the two domains. To achieve this, we employ an adversarial discriminator objective. The goal of the discriminator is to keep finding differences between real and translated (fake) samples so that G_{XY} keeps improving. We denote the discriminator differentiating real and fake samples from domain Y as D_Y. For a clip $x \in X$, we denote its latent encoding as z^x (similarly, z^y for a sample $y \in Y$). By encoding individual frames of $x = \{x_1 \cdots x_t\}$ we can obtain the latent encodings $z^x = \{z_1^x, \cdots z_t^x\}$, where $z_i^x = q(x_i)$. The adversarial loss can be written as,

$$\mathcal{L}_{adv}(X, Y) = \mathbb{E}_{y \in Y}[\log D_Y(z^y)] + \mathbb{E}_{x \in X}[\log(1 - D_Y(G_{XY}(z^x)))] \tag{2}$$

Similar to other adversarial objectives [86, 87], we optimize this loss function by alternatively training the translation function G_{XY} and the discriminator D_Y. The translation function G_{XY} can be found using the following optimization,

$$G_{XY}^* = \arg\min_{G_{XY}} \max_{D} L_{adv}(X, Y) \tag{3}$$

Translation Function Architecture. If we do not constrain G_{XY} – for example, by using an MLP for G_{XY} – it would not be interpretable. Moreover, if the capacity of the network is high enough, the function could learn arbitrarily complex mapping between domains. For example, it could memorize a one-to-one mapping between examples from two domains, which is not a useful translation.

To tackle both these problems we need to constrain G_{XY}. Our key insight is that instead of directly learning to translate from domain X to Y, G_{XY} first predicts a *translator* function f, that when applied to a sample z^x results in the translated output $z^{y\prime} = f(z^x)$. The advantage of predicting such a *translator* is that its parameters are more interpretable than a black box model. For example,

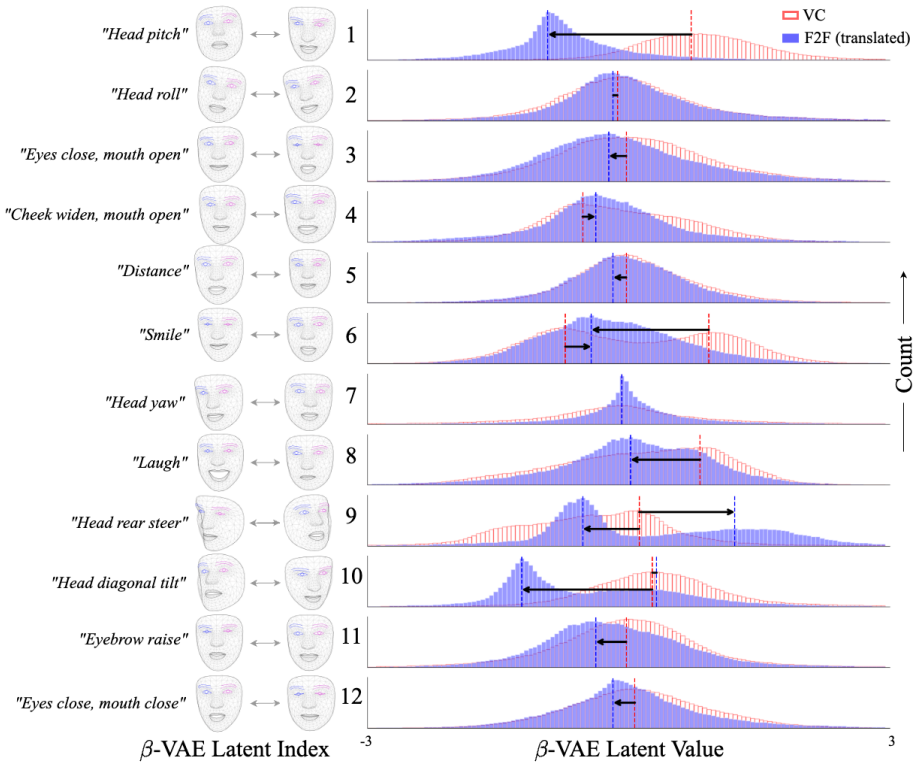

Fig. 3: Explanation of Disentangled Latents. We vary each dimension of the 12-dimensional latent obtained through β-VAE encoding by keeping other dimensions fixed (rows). **Left:** The faces corresponding to the extreme values of the perturbed latent, along with a description of the dominant change. **Right:** 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. Refer to supplementary for videos visualizing the latents.

as we will see in Fig. 6, our model G_{XY} finds consistent *translators* for clips with the same expressions such as "smiling" or "listening". We also parameterize the *translator* f to be a *shift and scale operation*, using a multiplicative factor $\omega \in \mathbb{R}^l$ and an additive factor $\phi \in \mathbb{R}^l$. Therefore the *translator* $f(x) = \omega \odot z + \phi$.

One issue with the above formulation is that predicting the same *translator* f for the entire clip can be too restrictive. The expressions and modes of conversation change within a clip. For example, a listener might start speaking or a speaker might exclaim in the middle of a clip. Therefore, instead of predicting a single *translator* f for the entire clip, we predict different *translators* for different chunks. This also makes the model more robust as it is forced to learn similar *translators* for similar segments. We achieve this by breaking down G_{XY} into two sub-modules: G_t and G_f. G_t first predicts $c-1$ temporal change-points

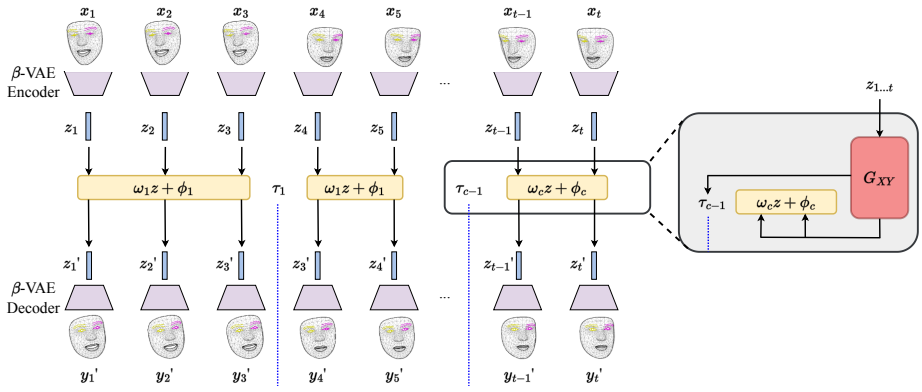

Fig. 4: Method Overview. Given a sequence of facial keypoints x, we use a pre-trained β-VAE encoder to obtain latents z. We then train a translation function G_{XY} that takes as input the latents z to produce a *translator*(ω, ϕ, τ). This *translator* when applied to z generates the transformed latent z', which can be decoded using the β-VAE decoder to obtain the transformed facial keypoints y' belonging in the new domain.

$\{\tau_1, \cdots \tau_{c-1}\}$ partition the clip into c chunks. G_f then predicts a *translator* for each chunk – e.g., for the k^{th} chunk, G_f predicts the *translator* f_k (ω_k, ϕ_k).

A key challenge here is that we do not have supervision to learn how to partition a clip into semantically meaningful chunks. In fact, discovering such meaningful chunks from the data is one of our goals. Therefore, we learn both the functions G_f and G_t together in an end-to-end fashion. Note that partitioning a clip using values from G_t and passing each chunk through G_f is a non-differentiable operation. To circumvent this, we approximate this partitioning operation with a continuous alternative. For each chunk k, we learn smooth weights over time $\bar{w}_k \in [0,1]^t$, where t is the total clip length. If T is the vector of time indexes $[1, 2, \cdots t]$ for a clip, then we define $\bar{w}_k$ as:

$$w_k = \begin{cases} \sigma(\tau_k - T, Q) & k = 1 \\ \min(\sigma(T - \tau_{k-1}, Q), \sigma(\tau_k - T, Q)) & k \in [2, c-1] \\ \sigma(T - \tau_{k-1}, Q) & k = c \end{cases} \qquad \bar{w}_k = \frac{w_k}{\sum_{k=1}^{c} w_k} \tag{4}$$

Here σ is the sigmoid function, and Q is the temperature. Essentially, w_k is a continuous alternative of a non-differentiable rectangular pulse function. $\bar{w}_k$ normalizes w_k, so that they sum up to 1 at every time step. With $\bar{w}_k$, a differentiable *translator* predictor can be written as $G_f(z^x \odot \bar{w}_k)$ for chunk k.

To summarize, our model minimizes the optimization problem defined in Eq. 3 to predict the parameters of the two networks G_t and G_f, which together represent G_{XY}. An overview of this model is illustrated in Fig. 4

3.3 Interpretability

Our translation function is interpretable by design. First, the model can partition a clip into semantically uniform chunks without supervision. Second, predicting a *translator* instead of predicting the translation function G_{XY}, allows us to understand how a particular expression or action changes between domains. For e.g., we can answer how "listening" changes between VC and F2F conversations. Finally, modeling a *translator* as a shift and scale operation, allows us to factor the translations into individual β-VAE latents. For e.g., we can answer whether people "smile more or less" when "listening" in VC versus F2F conversations.

Note that formulating the *translator* to be a shift and scale operation is a modeling choice, which can be flexibly changed depending on the task and sample representation. Also note that our framework closely resembles CycleGAN [87] with its adversarial objective, albeit without the cycle consistency loss. Cycle consistency is used to over-constrain the model and avoiding trivial solutions. Moreover, even with cycle consistency, a model with enough capacity can memorize a one-to-one mapping between examples from the two domains [87,88]. This is not a problem in cases where humans have prior knowledge and can qualitatively show that the outputs change in accordance with our prior knowledge. However, we do *not* have prior knowledge of what makes VC and F2F different. Our model is over-constrained by design to not lead to such trivial solutions. [89].

3.4 Domain Transfer

Besides systematically discovering differences across domains, `FacET` being generative in nature, can also transform inputs ($x \rightarrow y'$) from domain X to Y. To visualize the results in pixel-space, we leverage a video-to-video model, vid2vid [90], that learns keypoints to pixels using our dataset. Compared to existing methods that simply correct for gazes [91], we expect that style-transfer using `FacET` will lead to more faithful transfer that encompasses subtle expression changes as well.

4 Experiments

In this section, we show how our model can discover differences between domains. First, we describe the two datasets collected to study this task (Sec. 4.1). Then we evaluate `FacET` quantitatively (Sec. 4.2) on these datasets, and discuss its interpretable outcomes (Sec. 4.3). Refer supplementary for implementation details.

4.1 Datasets

To study the differences between VC and F2F, we collect the ZoomIn dataset from YouTube, containing 240 hours of conversations in both domains (See Fig. 1 for examples). Additionally, to study differences in communication styles across subjects, we collect a dataset of 20 hours of individual speaking styles of US presidents Obama and Trump. Each clip consists of detections on contiguous 176 frames from a participants. Refer to supplementary for more details.

4.2 Quantitative Results

Metrics. Evaluating discovered differences is hard in the absence of a paired dataset, since we cannot evaluate what will be discovered beforehand. Thus, our metric is the translation function G_{XY}'s ability to fool the discriminator D, while being interpretable. This is measured by discriminator accuracy, which we notice tends to stabilize after a period of training. We average it across the last 100 epochs to ensure robustness. Refer supplementary for more on metrics.

Table 1: Performance of our method by its ability to fool the discriminator whilst being interpretable (lower is better, 50% is optimal).

Model Type			**Discriminator Acc. (%) ↓**					
			ZoomIn			Presidents		
G_f	G_t	c	Avg	F2F→VC	VC→F2F	Avg	O→T	T→O
Fixed translator set	No Partitions	1	87.58	87.92	87.06	81.40	71.33	92.19
	Fixed-size Chunks	2	93.95	93.23	94.70	90.95	86.90	95.21
	Fixed-size Chunks	7	97.78	98.54	97.02	96.64	94.81	98.63
	Var. Chunks	2	92.26	91.91	92.62	83.80	75.14	92.63
	Var. Chunks	7	97.67	97.18	98.15	97.17	95.54	98.76
Predicted translator	No-partitions	1	78.54	79.71	77.71	79.25	69.88	88.97
	Fixed-size Chunks	2	82.99	82.10	84.12	88.67	87.63	89.42
	Fixed-size Chunks	7	81.84	81.06	82.95	89.86	89.73	89.66
	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

Baselines. To the best of our knowledge, ours is the first work that tackles this problem, so there are no comparable baselines. Thus, we validate our design choices through the following interpretable ablations of our model:

- **No Partitions:** a model where we predict a single *translator* for the entire duration of the clip i.e. $c = 1$. Note that this model does not need G_t.
- **Fixed-size Chunks:** a model with uniform, equal-sized chunks. This model also does not use G_t. However G_f predicts different *translator*s for each chunk.
- **Fixed translator-set:** instead of *predicting* the parameters of the *translator*, our model *retrieves* parameters from a set of p options. We also learn this fixed-size set of parameters P. G_f then becomes a p-way classifier that takes as input a chunk and predicts one of the p classes using the softmax function. Since the parameters are $\omega, \phi \in \mathbb{R}^l$, $P \in \mathbb{R}^{p \times 2l}$ is a matrix. We jointly optimize both the classifier and the parameters P. We use $p = 32$ for our experiments.

Is FacET better at translation than the baselines? Tab. 1 shows the performance of FacET and baselines using discriminator accuracy. We learn translation functions in both directions between the domains.

FacET performs better translations than baselines and consequently results in the lowest discriminator accuracy. With uniform equal-sized chunks, G_f cannot accurately predict good *translator*s. Facial expressions may change in a chunk, but this model would apply the same translation function throughout, resulting in poor translations. Similarly, if we do not partition the clip, the translations are also of poor quality. Swapping the model predicting the *translator* with a model selecting *translator*s from a set also results in poor translations. We argue that a fixed set of *translators* is too restrictive and cannot model all possible expressions even with a larger p. For FacET increasing the number of chunks from 2 to 7 does not affect performance by a lot. In the presidential data, since utterances are often monotonous and in a "speech"-like non-dyadic manner, more chunks are not beneficial to model performance. We believe, based on the nature of podcast conversations, expressions often do not change more than once in a 7-second clip. Thus, we use FacET with 2 chunks for qualitative analysis.

4.3 Qualitative Results

Now that we have demonstrated that FacET can faithfully translate samples, we look at what we can discover on our data.

How do F2F conversations differ from VC? We answer this by using FacET. We look at the distribution of latent codes for F2F and translated VC obtained from FacET. Fig. 3 shows differences between VC and F2F along the 12 latent codes. Some differences are quite apparent, for example, people look down with respect to camera in VC (because the cameras are generally above the screen during a VC). This is evident in latent code #1's distribution, where FacET rotates the head upwards when translating from VC to F2F.

Similarly, we observe big changes in head tilt (#10) and head steer (#9), because in VC the face often always stays towards the screen, whereas the orientation of the head can vary a lot in F2F. We also notice that the head tilt (#10) and head steer (#9) distributions become bimodal after translation (from unimodal). This can again be attributed to the fact that, during a VC, we only look at screens, however in a 2-person F2F conversation, we will see two modes with peaks at the orientation where the subjects are looking at one another.

Some other observations are not quite easily discernible but are evident from our method. We see that the the smile latent distribution (#6) translates from a bimodal distribution in VC to a unimodal distribution with a peak in the center in F2F. Similarly, the eyebrow raise latent (#11) becomes very muted in an F2F conversation. We conjecture that people tend to emote more during a VC as smaller reactions are harder to pick in a VC [92]. On the contrary, we notice that people laugh less during a VC. While it is easy to emote silently on Zoom, laughing is harder because VC systems usually allow only one speaker at a time.

How do specific expressions change between VC and F2F? We looked at how the overall behavior changes between the two domains. However, the key ability of our model is delving deeper – how do specific spatio-temporal patterns change across the two domains? We aim to answer questions such as 'How does a

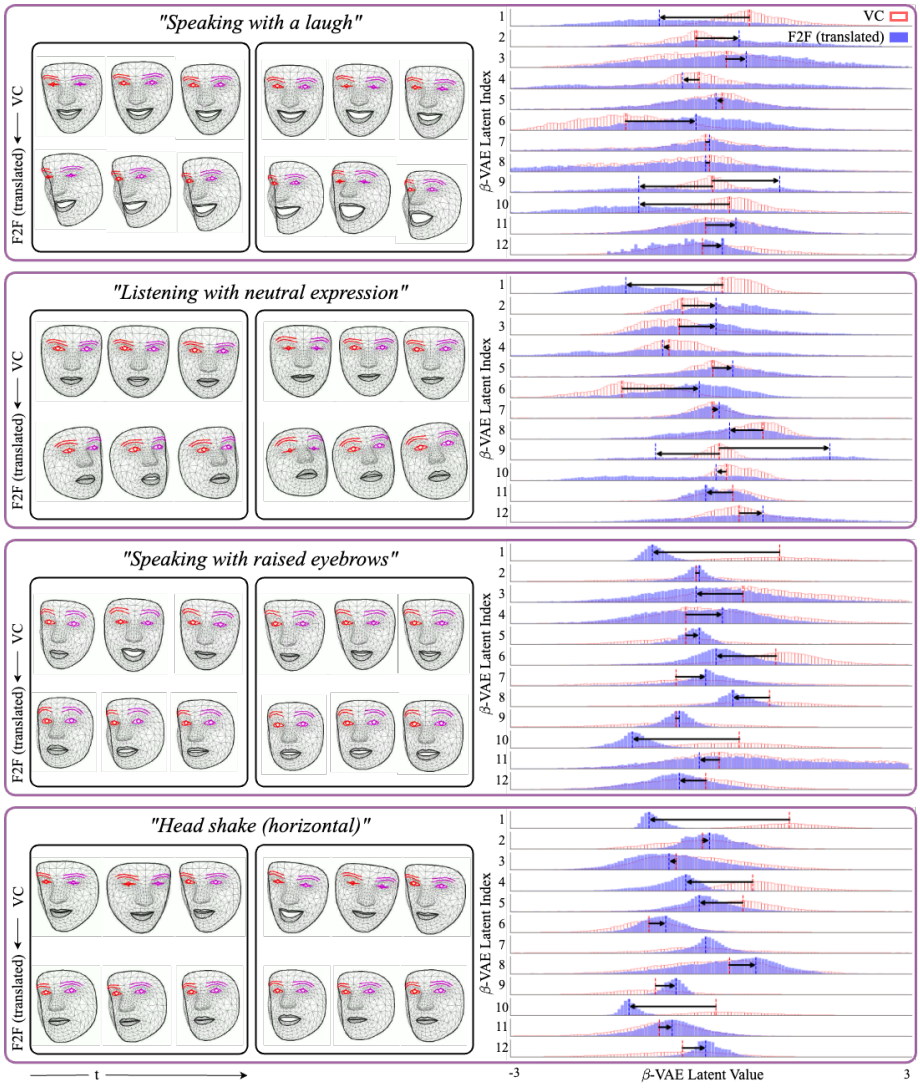

Fig. 5: Results. We showcase our key findings of domain differences through a detailed report. **Left:** Each row corresponds to two different examples belonging to a specific *translator* cluster (e.g., *"Speaking with a smile"*). The top row shows a video recorded in VC while the bottom row shows the transformed video as if it were F2F. **Right:** `FacET` generates a report for each corresponding cluster showing how each β-VAE latent varies across the domains. Refer to Fig. 3 to interpret each latent index. See Sec. 4.3 for detailed explanations. Refer supplementary for full videos.

person's laugh change between the domains?'. To answer this we extract spatio-temporal chunks with similar expressions. We posit that since our partitioning

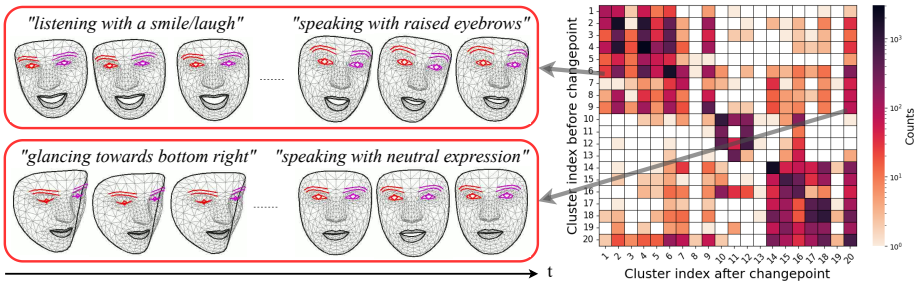

Fig. 6: ***Translator*** **cluster analysis.** We perform k-means clustering on all predicted *translators* of the dataset and perform change-point analysis using them. **Right:** An entry $[i, j]$ in the matrix shows the frequency of timestamp changes while going from cluster j to i. **Left:** We show two examples of frequently occurring cluster changes from #1 to #6 (**top**) and #20 to #9 (**bottom**). Note that the matrix would have been a diagonal if changepoint (τ) prediction were not needed.

function (G_t) segments clips into meaningful chunks and our translation predictor G_f predicts similar *translators* for similar chunks, clustering parameters of the *translator* for a chunk should give meaningful spatio-temporal clusters.

Therefore, we concatenate the ω and τ parameters for each chunk in our dataset and cluster them using k-means. We use the BIC-score [93] to find the optimal number of clusters. Many clusters are indeed meaningful. Fig. 5 (left) shows 2 examples each from 4 such clusters, with expressions such as "speaking with raised eyebrows" or "head shake". We translate all VC chunks from a cluster to F2F and observe the distribution of latents specific to the cluster. This analysis also results in many key insights. (see supplementary for more such insights)

– **Speakers tend to laugh bigger in F2F:** This can be observed with the "Laugh" latent code #8 in the first "speaking with a laugh" cluster. Even though the peaks of the histogram do not shift by much, we can clearly see a distribution shift towards a bigger laughter (to the left). Note that a "bigger speaker laugh in F2F" is a different observation than "more laughter in F2F" that we discussed in the previous section. We could not have made this observation without the analysis of clusters shown in Fig. 5
– **F2F Listeners do less head shakes:** This aligns with the previously reported results – people emote more in VC [92]. This is evident both from the translated faces and the distribution shift of the "Head read steer" code (#9).

How do expressions change temporally? Another important aspect is how people transition from one expression to the next. Since we are interested in temporally finer expression changes, we use the model with 7 **fixed-size chunks** and obtain *translators* for individual chunks. We cluster these *translators* into 20 clusters and visualize common expression transitions. We measure the number of times a transition happens from one cluster to another in consecutive chunks. Fig. 6 (right) shows the number of such transitions from one cluster to another.

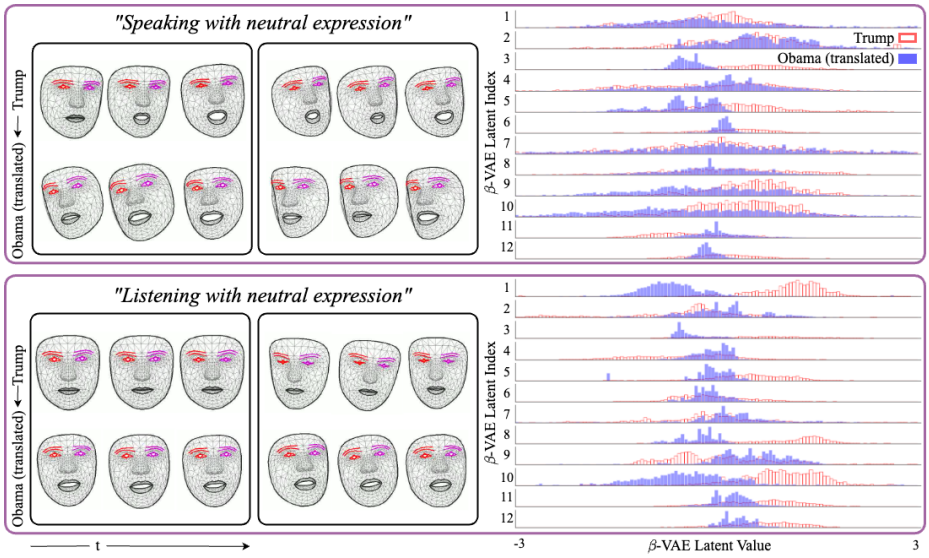

Fig. 7: Analysis of Presidents: 2 *translator* clusters (Left) and the corresponding FacET reports on the president (Obama and Trump) face dataset.

Using this we can infer the common expression transitions. We show examples of two frequently occurring transitions in Fig. 6 (left). The first transition is people switching from listening to speaking while keeping happy expressions. Similarly, people frequently switch from listening and looking downwards to speaking neutrally. Since a lot of values are off the diagonal, this also shows that our model can detect the changes due to the partitioning function G_t.

Is FacET generalizable to other tasks? We also apply FacET to a different task – finding differences between facial expressions of two subjects, presidents Obama (O) and Trump (T). Since this dataset is smaller, we initialize training weights of the β-VAE using the ZoomIn β-VAE. A favorable side-effect is that the meanings of the latent codes do not change between datasets. However, there are some small differences due to the dataset distribution (refer supplementary). Fig. 7 shows histograms of distributions of latent codes when translating from Trump to Obama, for two clusters. Several observations can be made:

- **Trump has a more circular mouth when speaking.** This is evident from the clips from the speaking clusters as well as the left shift of the latent code #6 distribution from Trump to Obama. Note that this shift is not observable in the listening cluster. Moreover, the distribution of latent code #6 is spread out for Trump in both clusters indicating a wider range of horizontal mouth motion. Both these observations have also been reported in the news [94, 95].
- **Trump raises eyebrows more while listening.** Looking at the latent code #11 we observe that Trump's eyebrows are more raised when listening in comparison to Obama and also in comparison to himself while speaking [96].

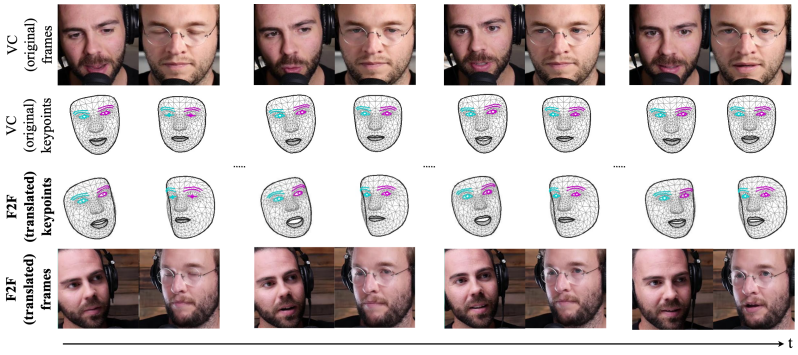

Fig. 8: De-zoomification Given facial keypoints of two people on a VC, FacET generates translated keypoints simulating a F2F conversation, which are then converted to pixel-space (Sec. 3.4). See supplementary for videos.

4.4 Applications (de-zoomification)

We showed in our results that VC → F2F is more than just changing the eye gaze or head pose. We call this process *de-zoomification*, and show that our method (explained in Sec. 3.4) can be used to convert an actual conversation that took place in a VC setting, into a F2F conversation as shown in Fig. 8. We observe that eye blinks (first frame) and smiles (last frame) in *de-zoomed* videos are consistent with the original. The subjects also look more toward each other in the translated video which makes it feel like a F2F conversation. This is an exciting application towards making virtual conversations feel life-like.

5 Discussion & Conclusion

Limitations & Future Work. While our method provides detailed and interpretable reports, there is still room for improvement. First, our method relies on β-VAE for disentangling, which can introduce some noise as true disentanglement without imposing priors is hard [97]. However, our method allows easy swap of β-VAE with better future alternatives. Second, we currently demonstrate results on human facial keypoints only. While our method is general to any keypoint data, it is non-trivial to extend our architecture to images.

Conclusion. We addressed the task of discovering spatio-temporal patterns from unpaired facial keypoint data by presenting FacET, a generative domain translation method that is interpretable-by-design. FacET pre-computes domain agnostic, disentangled features, and learns interpretable linear transformations between domains. FacET can predict temporal change-points that decouple expression changes from time. Extensive experiments showcase FacET's ability to generate informative reports of input-specific spatio-temporal patterns between domains. Lastly, we apply these learned differences to *de-zoom* VC videos.

Acknowledgements. This research is based on work partially supported by the DARPA CCU program under contract HR001122C0034 and the National Science Foundation AI Institute for Artificial and Natural Intelligence (ARNI). PT is supported by the Apple PhD fellowship.

References

1. N. Zhao, X. Zhang, J. A. Noah, M. Tiede, and J. Hirsch, "Separable processes for live "in-person" and live "zoom-like" faces," *Imaging Neuroscience*, 2023.
2. S. Balters, J. G. Miller, R. Li, G. Hawthorne, and A. L. Reiss, "Virtual (Zoom) Interactions Alter Conversational Behavior and Inter-Brain Coherence," *bioRxiv*, 2023.
3. S. Matz and G. Harari, "Personality–place transactions: Mapping the relationships between big five personality traits, states, and daily places," *Journal of Personality and Social Psychology*, 2020.
4. M. R. Khan, "A review of the effects of virtual communication on performance and satisfaction across the last ten years of research," *Journal of Applied Behavior Analysis*, 2021.
5. M. Archibald, R. Ambagtsheer, M. Casey, and M. Lawless, "Using zoom videoconferencing for qualitative data collection: Perceptions and experiences of researchers and participants," *International Journal of Qualitative Methods*, 2019.
6. H. Nesher Shoshan and W. Wehrt, "Understanding "zoom fatigue": A mixed-method approach," *Applied Psychology*, 2022.
7. G. Fauville, M. Luo, A. C. M. Queiroz, J. N. Bailenson, and J. Hancock, "Zoom Exhaustion & Fatigue Scale," *Computers in Human Behavior Reports*, 2021.
8. J. N. Bailenson, "Nonverbal Overload: A Theoretical Argument for the Causes of Zoom Fatigue," *Technology, Mind, and Behavior*, 2021.
9. J. Boland, P. Fonseca, I. Mermelstein, and M. Williamson, "Zoom disrupts the rhythm of conversation," *Journal of Experimental Psychology: General*, 2021.
10. G. Fauville, M. Luo, A. C. Queiroz, J. Bailenson, and J. Hancock, "Zoom exhaustion & fatigue scale," *SSRN Electronic Journal*, 2021.
11. M. Hoehe and F. Thibaut, "Going digital: how technology use may influence human brains and behavior," *Dialogues in clinical neuroscience*, 2020.
12. T. Numata, H. Sato, Y. Asa, T. Koike, K. Miyata, E. Nakagawa, M. Sumiya, and N. Sadato, "Achieving affective human–virtual agent communication by enabling virtual agents to imitate positive expressions," *Scientific Reports*, 2020.
13. H. J. Smith and M. Neff, "Communication behavior in embodied virtual reality," in *ACM CHI*, 2018.
14. S. Geng, R. Teotia, P. Tendulkar, S. Menon, and C. Vondrick, "Affective faces for goal-driven dyadic communication," *CoRR*, 2023.
15. R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *ICCV*, 2019.
16. V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," *CoRR*, 2018.
17. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017.
18. V. Shitole, F. Li, M. Kahng, P. Tadepalli, and A. Fern, "One explanation is not enough: structured attention graphs for image classification," *NeurIPS*, 2021.

19. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.
20. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.
21. D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI Magazine*, 2019.
22. Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *ICML*, 2019.
23. S. Vandenhende, D. Mahajan, F. Radenovic, and D. Ghadiyaram, "Making heads or tails: Towards semantically consistent visual counterfactuals," in *ECCV*, 2022.
24. P. Wang and N. Vasconcelos, "Scout: Self-aware discriminant counterfactual explanations," in *CVPR*, 2020.
25. M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *SIGKDD*, 2016.
26. P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *ICML*, 2017.
27. C.-K. Yeh, J. Kim, I. E.-H. Yen, and P. K. Ravikumar, "Representer point selection for explaining deep neural networks," *NeurIPS*, 2018.
28. C.-P. Tsai, C.-K. Yeh, and P. Ravikumar, "Sample based explanations via generalized representers," *CoRR*, 2023.
29. Y. Sui, G. Wu, and S. Sanner, "Representer point selection via local jacobian expansion for post-hoc classifier explanation of deep neural networks and ensemble models," in *NeurIPS*, 2021.
30. G. Pruthi, F. Liu, M. Sundararajan, and S. Kale, "Estimating training data influence by tracking gradient descent," *CoRR*, 2020.
31. A. Silva, R. Chopra, and M. C. Gombolay, "Cross-loss influence functions to explain deep network representations," in *AISTATS*, 2020.
32. H. Guo, N. Rajani, P. Hase, M. Bansal, and C. Xiong, "Fastif: Scalable influence functions for efficient model interpretation and debugging," *CoRR*, 2020.
33. W. Pan, S. Cui, J. Bian, C. Zhang, and F. Wang, "Explaining algorithmic fairness through fairness-aware causal path decomposition," in *SIGKDD*, 2021.
34. R. Pradhan, J. Zhu, B. Glavic, and B. Salimi, "Interpretable data-based explanations for fairness debugging," in *SIGMOD*, 2022.
35. C. Meng, L. Trinh, N. Xu, J. Enouen, and Y. Liu, "Interpretability and fairness evaluation of deep learning models on mimic-iv dataset," *Scientific Reports*, 2022.
36. S. Alelyani, "Detection and evaluation of machine learning bias," *Applied Sciences*, 2021.
37. L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *DSAA*, 2018.
38. S. S. Kim, N. Meister, V. V. Ramaswamy, R. Fong, and O. Russakovsky, "Hive: Evaluating the human interpretability of visual explanations," in *ECCV*, 2022.
39. R. R. Selvaraju, P. Tendulkar, D. Parikh, E. Horvitz, M. T. Ribeiro, B. Nushi, and E. Kamar, "Squinting at vqa models: Introspecting vqa models with sub-questions," in *CVPR*, 2020.
40. A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?," *Computer Vision and Image Understanding*, 2017.
41. W. Brendel and M. Bethge, "Approximating cnns with bag-of-local-features models works surprisingly well on imagenet," *CoRR*, 2019.
42. M. Bohle, M. Fritz, and B. Schiele, "Convolutional dynamic alignment networks for interpretable classifications," in *CVPR*, 2021.

43. M. Böhle, M. Fritz, and B. Schiele, "B-cos networks: Alignment is all we need for interpretability," in *CVPR*, 2022.
44. C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: deep learning for interpretable image recognition," *NeurIPS*, 2019.
45. J. Donnelly, A. J. Barnett, and C. Chen, "Deformable protopnet: An interpretable image classifier using deformable prototypes," in *CVPR*, 2022.
46. P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *ICML*, 2020.
47. T. Hastie and R. Tibshirani, "Generalized Additive Models," *Statistical Science*, 1986.
48. Y. Lou, R. Caruana, J. Gehrke, and G. Hooker, "Accurate intelligible models with pairwise interactions," *SIGKDD*, 2013.
49. A. Dubey, F. Radenovic, and D. Mahajan, "Scalable interpretability via polynomials," *NeurIPS*, 2022.
50. F. Radenovic, A. Dubey, and D. Mahajan, "Neural basis models for interpretability," *NeurIPS*, 2022.
51. C.-H. Chang, R. Caruana, and A. Goldenberg, "Node-gam: Neural generalized additive model for interpretable deep learning," in *ICLR*, 2022.
52. R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton, "Neural additive models: Interpretable machine learning with neural nets," *NeurIPS*, 2021.
53. C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-vae," *CoRR*, 2018.
54. Z. Zhu, P. Luo, X. Wang, and X. Tang, "Multi-view perceptron: a deep model for learning face identity and view representations," in *NeurIPS*, 2014.
55. S. E. Reed, K. Sohn, Y. Zhang, and H. Lee, "Learning to disentangle factors of variation with manifold interaction," in *ICML*, 2014.
56. W. F. Whitney, M. Chang, T. D. Kulkarni, and J. B. Tenenbaum, "Understanding visual concepts with continuation learning," *CoRR*, 2016.
57. B. Cheung, J. A. Livezey, A. K. Bansal, and B. A. Olshausen, "Discovering hidden factors of variation in deep networks," *CoRR*, 2014.
58. Z. Lin, K. K. Thekumparampil, G. C. Fanti, and S. Oh, "Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers," *CoRR*, 2019.
59. I. Jeon, W. Lee, M. Pyeon, and G. Kim, "IB-GAN: Disentangled Representation Learning with Information Bottleneck Generative Adversarial Networks," *AAAI*, 2021.
60. X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *NeurIPS*, 2016.
61. A. Ramesh, Y. Choi, and Y. LeCun, "A spectral regularizer for unsupervised disentanglement," *CoRR*, 2018.
62. Y. Dalva, S. F. Altındiş, and A. Dundar, "Vecgan: Image-to-image translation with interpretable latent directions," in *ECCV*, 2022.
63. Y. Dalva, H. Pehlivan, C. Moran, Ö. I. Hatipoğlu, and A. Dündar, "Face attribute editing with disentangled latent vectors," *CoRR*, 2023.
64. I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.
65. H. Kim and A. Mnih, "Disentangling by factorising," in *ICML*, 2018.
66. T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," *CoRR*, 2018.

67. Y. Jeong and H. O. Song, "Learning discrete and continuous factors of data via alternating disentanglement," in *ICML*, 2019.
68. A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," *CoRR*, 2017.
69. K. Yesu, S. Shandilya, N. Rekharaj, K. Ankit, and P. S. Sairam, "Big five personality traits inference from five facial shapes using cnn," in *International Conference on Computing, Power and Communication Technologies (GUCON)*, 2021.
70. G. G. Knyazev, A. V. Bocharov, H. R. Slobodskaya, and T. I. Ryabichenko, "Personality-linked biases in perception of emotional facial expressions," *Personality and Individual Differences*, 2008.
71. A. Kachur, E. Osin, D. Davydov, K. Shutilov, and A. Novokshonov, "Assessing the big five personality traits using real-life static facial images," *Scientific Reports*, 2020.
72. B. Büdenbender, T. T. A. Höfling, A. B. M. Gerdes, and G. W. Alpers, "Training machine learning algorithms for automatic facial coding: the role of emotional facial expressions prototypicality," *PLOS one*, 2023.
73. A. Stahelski, A. Anderson, N. Browitt, and M. Radeke, "Facial expressions and emotion labels are separate initiators of trait inferences from the face," *Frontiers in Psychology*, 2021.
74. L. Snoek, R. Jack, P. Schyns, O. Garrod, M. Mittenbühler, C. Chen, S. Oosterwijk, and H. Scholte, "Testing, explaining, and exploring models of facial expressions of emotions," *Science advances*, 2023.
75. E. Straulino, C. Scarpazza, and L. Sartori, "What is missing in the study of emotion expression?," *Frontiers in Psychology*, 2023.
76. S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *PNAS*, 2014.
77. K. Minetaki, "Facial expression and description of personality," in *ACM MISNC*, 2023.
78. P. Jonell, T. Kucherenko, G. E. Henter, and J. Beskow, "Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings," in *ACM IVA*, 2020.
79. E. Ng, S. Subramanian, D. Klein, A. Kanazawa, T. Darrell, and S. Ginosar, "Can language models learn to listen?," in *ICCV*, 2023.
80. E. Ng, H. Joo, L. Hu, H. Li, , T. Darrell, A. Kanazawa, and S. Ginosar, "Learning to listen: Modeling non-deterministic dyadic facial motion," *CVPR*, 2022.
81. C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-vae," 2018.
82. Z. Li and H. Liu, "Beta-vae has 2 behaviors: Pca or ica?," 2023.
83. P. García de Herreros García, "Towards latent space disentanglement of variational autoencoders for language," 2022.
84. R. Pastrana, "Disentangling variational autoencoders," *CoRR*, 2022.
85. I. Higgins, L. Chang, V. Langston, D. Hassabis, C. Summerfield, D. Tsao, and M. Botvinick, "Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons," *Nature Communications*, 2021.
86. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, 2014.
87. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.
88. A. Chakrabarty and S. Das, "On Translation and Reconstruction Guarantees of the Cycle-Consistent Generative Adversarial Networks," in *Advances in Neural*

Information Processing Systems (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 23607–23620, Curran Associates, Inc., 2022.
89. Z. Shen, S. K. Zhou, Y. Chen, B. Georgescu, X. Liu, and T. Huang, "One-to-one mapping for unpaired image-to-image translation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1170–1179, 2020.
90. T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *NeurIPS*, 2018.
91. C. Kuster, T. Popa, J.-C. Bazin, C. Gotsman, and M. Gross, "Gaze correction for home video conferencing," *ACM TOG*, 2012.
92. F. Hill, "The gesture that encapsulates remote-work life," *The Atlantic*, July 20 2023.
93. G. Schwarz, "Estimating the dimension of a model," *The Annals of statistics*, 1978.
94. D. Denby, "The three faces of trump," *The New Yorker*, August 2015.
95. P. Collett, "The seven faces of donald trump – a psychologist's view," *The Guardian*, January 2017.
96. T. Golshan, "Donald trump's unique speaking style, explained by linguists," *Vox*, January 2017.
97. F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *ICML*, 2019.