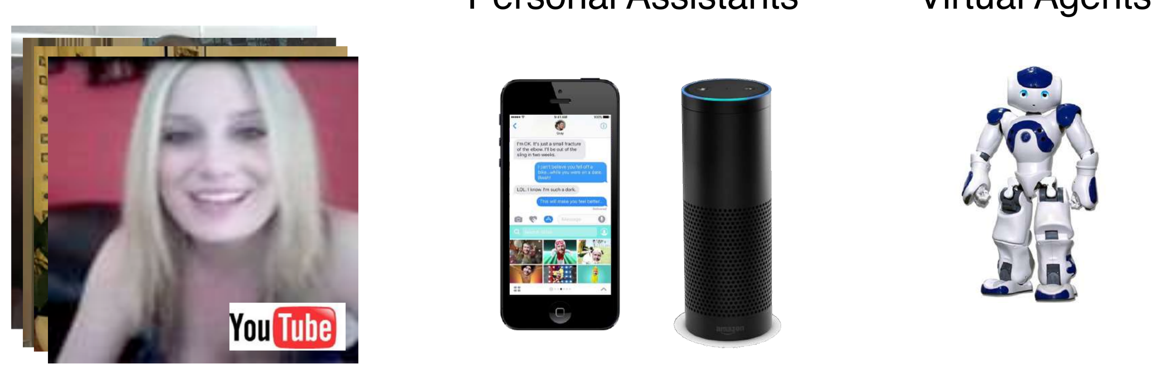
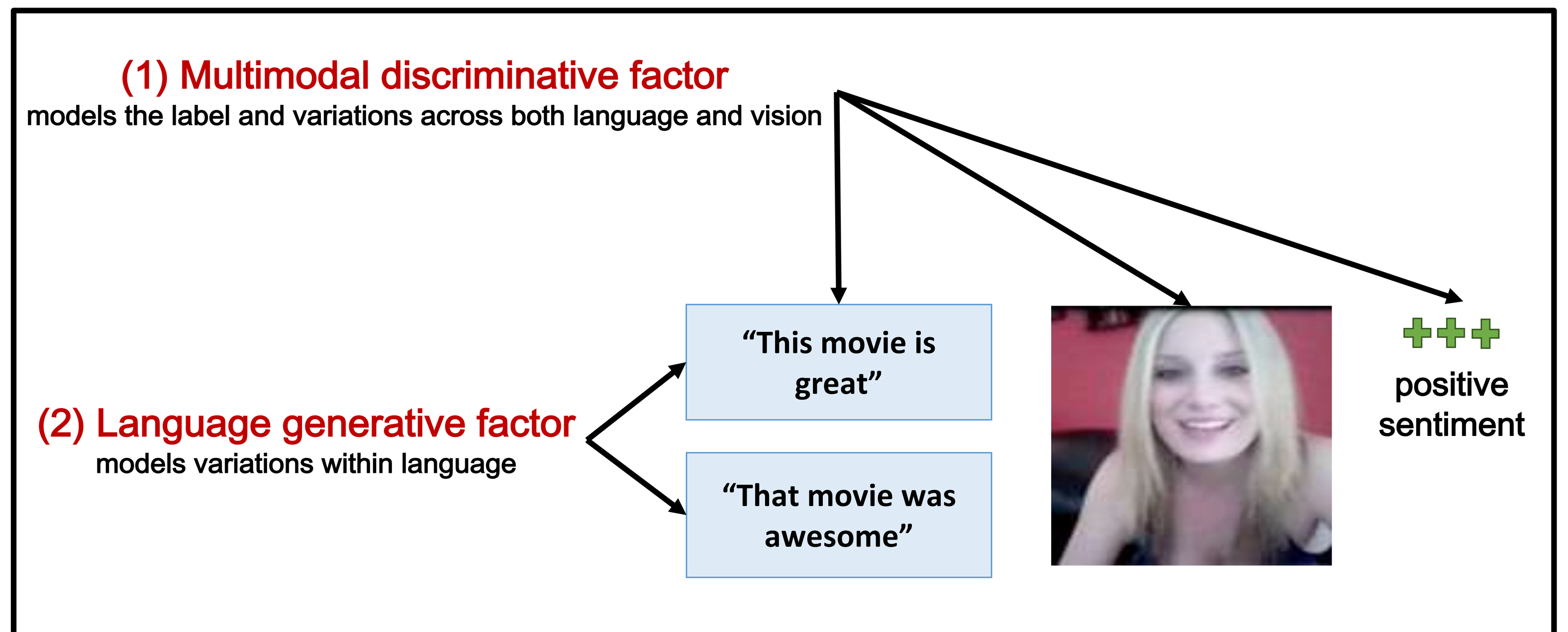
Learning Factorized Multimodal Representations

Yao-Hung Hubert Tsai*, Paul Pu Liang*, Amir Zadeh, Louis-Philippe Morency, Ruslan Salakhutdinov

Multimodal Representation Learning

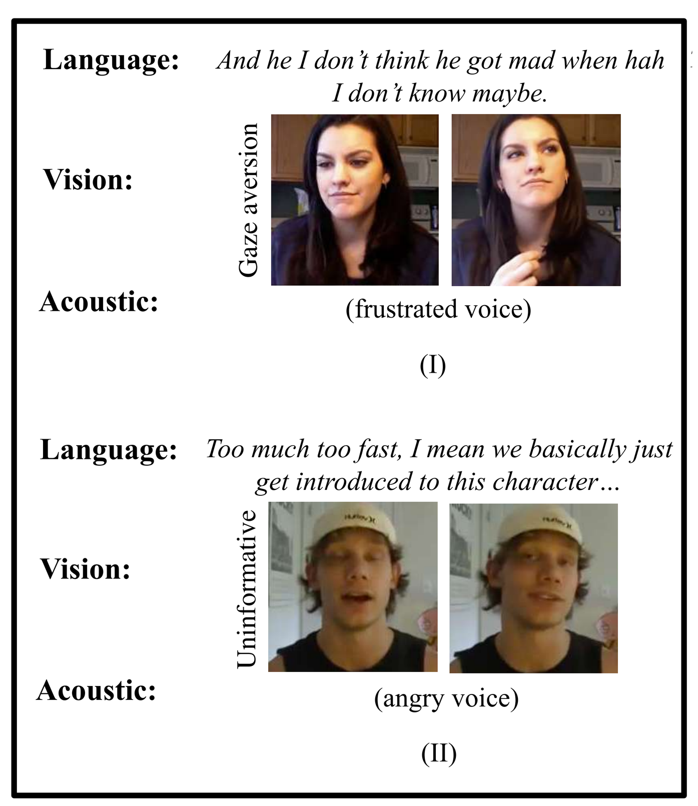


Factorize representation into independent sets of factors



Experimental Setup: Datasets and Features

- **Datasets in Human Multimodal Language**
  - **Multimodal Personal Trait Recognition**
    - Movie Reviews (POM)
  - **Multimodal Sentiment Analysis**
    - Monologue Opinion Videos (CMU-MOSI)
    - Online Social Reviews (ICT-MMMO)
    - Product Reviews (MOUD and YouTube)
  - **Multimodal Emotion Recognition**
    - Recorded Dyadic Dialogues (IEMOCAP)

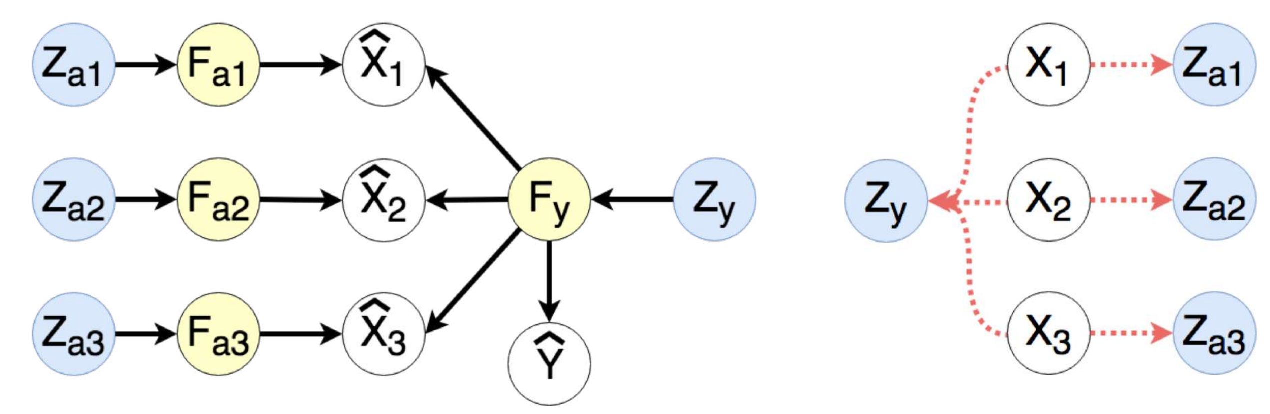


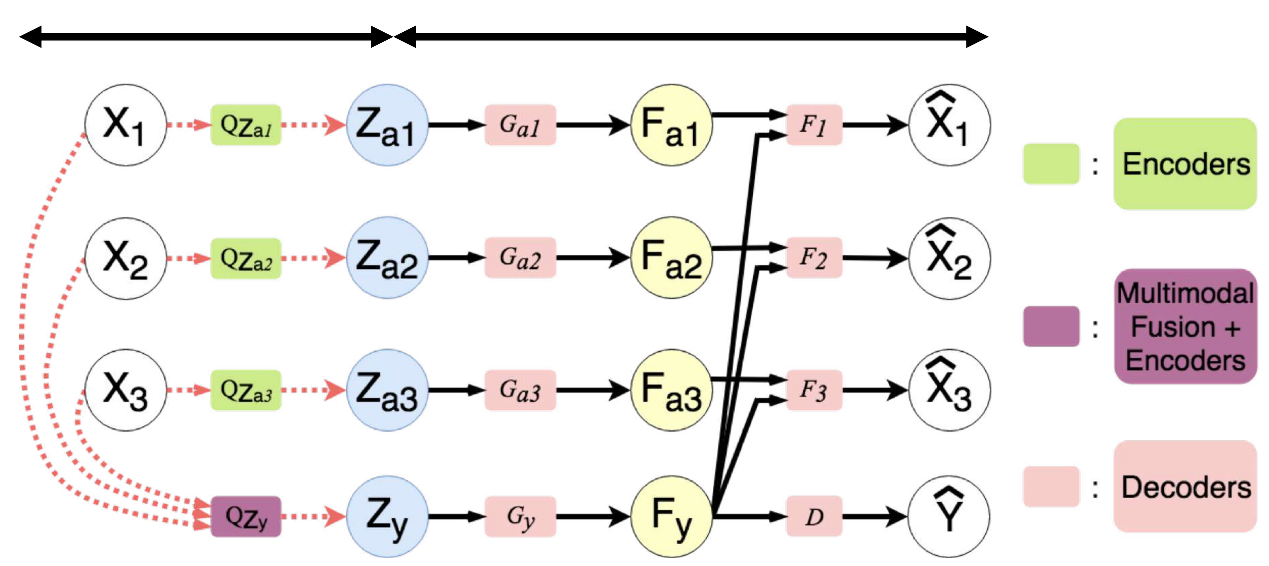
- **Multimodal Features**
  - **Language**: pre-trained Glove word embeddings
  - **Visual**: facial action units from Facet
  - **Acoustic**: MFCCs from COVAREP
  - Aligned by P2FA

Dataset	CMU-MOSI	ICT-MMMO	YouTube	MOUD	IEMOCAP	POM
Level	Segment	Video	Segment	Segment	Segment	Video
# Speakers	98	200	50	101	10	903
# Train	52→1284	220	30→169	49→243	5→6373	600
# Valid	10→229	40	5→41	10→37	1→1775	100
# Test	31→686	80	11→59	20→106	1→1807	203
Label	Sentiment	Sentiment	Sentiment	Sentiment	Emotions	Personalities
Features	$\{\ell, v, a\}$	$\{\ell, v, a\}$	$\{\ell, v, a\}$	$\{\ell, v, a\}$	$\{\ell, v, a\}$	$\{\ell, v, a\}$

Multimodal Factorization Model (MFM)

- **Bayesian Network**





- **Notations**
  - $\mathbf{X}_{1:M}$: multimodal data from M modalities, $\mathbf{Y}$: labels
  - $\hat{\mathbf{X}}_{1:M}$: generated multimodal data, $\hat{\mathbf{Y}}$: generated labels
  - $\mathbf{Z}_{\mathbf{a}\cdot}$: modality-specific latent variables, $\mathbf{F}_{\cdot}$: factors

Generation: Factorization over Joint Distribution

$$P(\hat{\mathbf{X}}_{1:M}, \hat{\mathbf{Y}}) = \int_{\mathbf{F},\mathbf{Z}} P(\hat{\mathbf{X}}_{1:M}, \hat{\mathbf{Y}}|\mathbf{F}) P(\mathbf{F}|\mathbf{Z}) P(\mathbf{Z}) d\mathbf{F} d\mathbf{Z}$$
$$= \int_{\substack{\mathbf{F_y},\mathbf{F_a}_{\{1:M\}}\\ \mathbf{Z_y},\mathbf{Z_a}_{\{1:M\}}}} \Big(P(\hat{\mathbf{Y}}|\mathbf{F_y}) \prod_{i=1}^{M} P(\hat{\mathbf{X}}_i|\mathbf{F_a}_i, \mathbf{F_y})\Big)\Big(P(\mathbf{F_y}|\mathbf{Z_y}) \prod_{i=1}^{M} P(\mathbf{F_a}_i|\mathbf{Z_a}_i)\Big)\Big(P(\mathbf{Z_y}) \prod_{i=1}^{M} P(\mathbf{Z_a}_i)\Big) \, d\mathbf{F}\, d\mathbf{Z},$$

with $d\mathbf{F} = d\mathbf{F_y} \prod_{i=1}^{M} d\mathbf{F_a}_i$ and $d\mathbf{Z} = d\mathbf{Z_y} \prod_{i=1}^{M} d\mathbf{Z_a}_i$

Inference: Joint-Distribution Wasserstein Distance

- **Proposition 1**: For any functions $G_y : \mathbf{Z_y} \to \mathbf{F_y}$, $G_{a\{1:M\}} : \mathbf{Z_a}_{\{1:M\}} \to \mathbf{F_a}_{\{1:M\}}$, $D : \mathbf{F_y} \to \hat{\mathbf{Y}}$, and $F_{1:M} : \mathbf{F_a}_{\{1:M\}}, \mathbf{F_y} \to \hat{\mathbf{X}}_{1:M}$, we have Joint-Distribution Wasserstein distance $W_c(P_{\mathbf{X}_{1:M},\mathbf{Y}}, P_{\hat{\mathbf{X}}_{1:M},\hat{\mathbf{Y}}}) =$

$$\inf_{Q_{\mathbf{Z}}=P_{\mathbf{Z}}} \mathbf{E}_{P_{\mathbf{X}_{1:M},\mathbf{Y}}} \mathbf{E}_{Q(\mathbf{Z}|\mathbf{X}_{1:M},\mathbf{Y})} \Big[\sum_{i=1}^{M} c_{X_i}\Big(\mathbf{X}_i, F_i\big(G_{ai}(\mathbf{Z_a}_i), G_y(\mathbf{Z_y})\big)\Big) + c_Y\Big(\mathbf{Y}, D\big(G_y(\mathbf{Z_y})\big)\Big)\Big],$$

where $P_{\mathbf{Z}}$ is the prior over $\mathbf{Z} = [\mathbf{Z_y}, \mathbf{Z_a}_{\{1,M\}}]$ and $Q_{\mathbf{Z}}$ is the aggregated posterior of the proposed approximate inference distribution $Q(\mathbf{Z}|\mathbf{X}_{1:M}, \mathbf{Y})$.

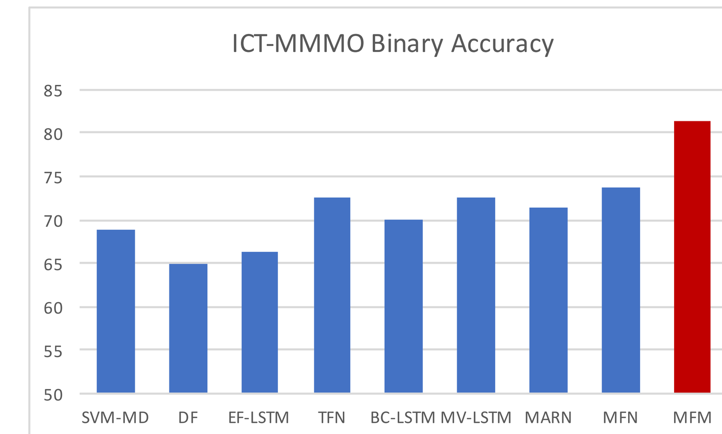
Relaxed Generative-Discriminative Objective

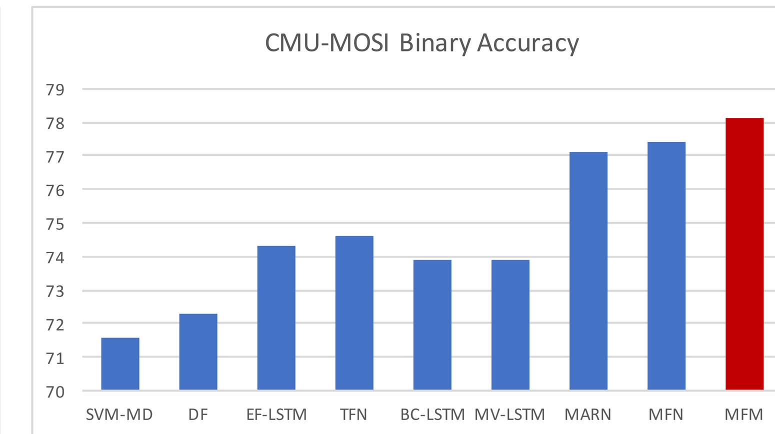
Mean-field assumption: $Q(\mathbf{Z}|\mathbf{X}_{1:M}, \mathbf{Y}) := Q(\mathbf{Z}|\mathbf{X}_{1:M}) := Q(\mathbf{Z_y}|\mathbf{X}_{1:M}) \prod_{i=1}^{M} Q(\mathbf{Z_a}_i|\mathbf{X}_i).$

Objective:

$$\min_{F, G_{a\{1:M\}}, G_y, D} \inf_{Q(\mathbf{Z}|\cdot)\in\mathcal{Q}} \mathbf{E}_{P_{\mathbf{X}_{1:M},\mathbf{Y}}} \mathbf{E}_{Q(\mathbf{Z_a}_1|\mathbf{X}_1)} \cdots \mathbf{E}_{Q(\mathbf{Z_a}_M|\mathbf{X}_M)} \mathbf{E}_{Q(\mathbf{Z_y}|\mathbf{X}_{1:M})}$$
$$\Big[\underbrace{\sum_{i=1}^{M} c_{X_i}\Big(\mathbf{X}_i, F\big(G_{ai}(\mathbf{Z_a}_i), G_y(\mathbf{Z_y})\big)\Big)}_{\text{Generative}} + \underbrace{c_Y\Big(\mathbf{Y}, D\big(G_y(\mathbf{Z_y})\big)\Big)}_{\text{Discriminative}} \Big] + \underbrace{\lambda \mathcal{MMD}(Q_{\mathbf{Z}}, P_{\mathbf{Z}})}_{\text{Regularizer}},$$

Results on Multimodal Time Series Datasets

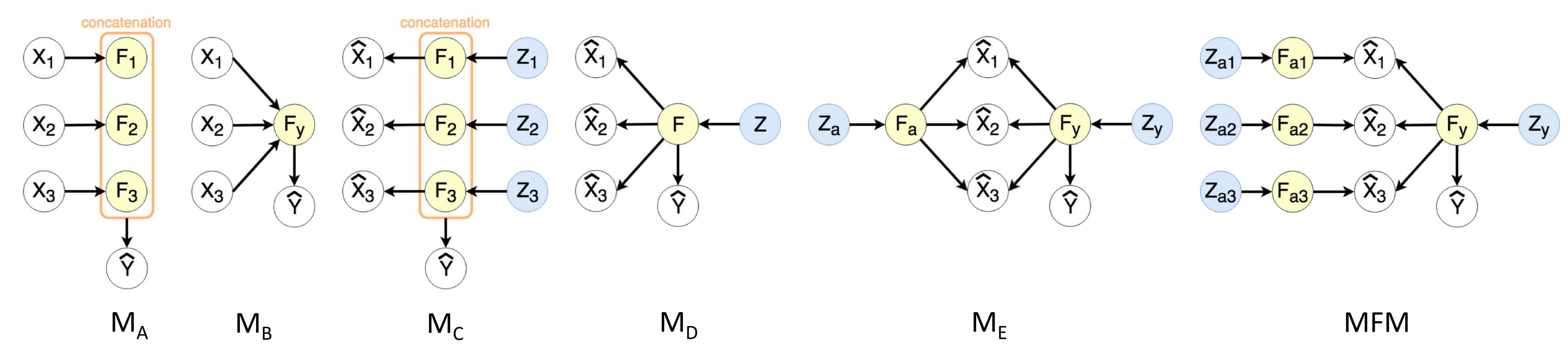




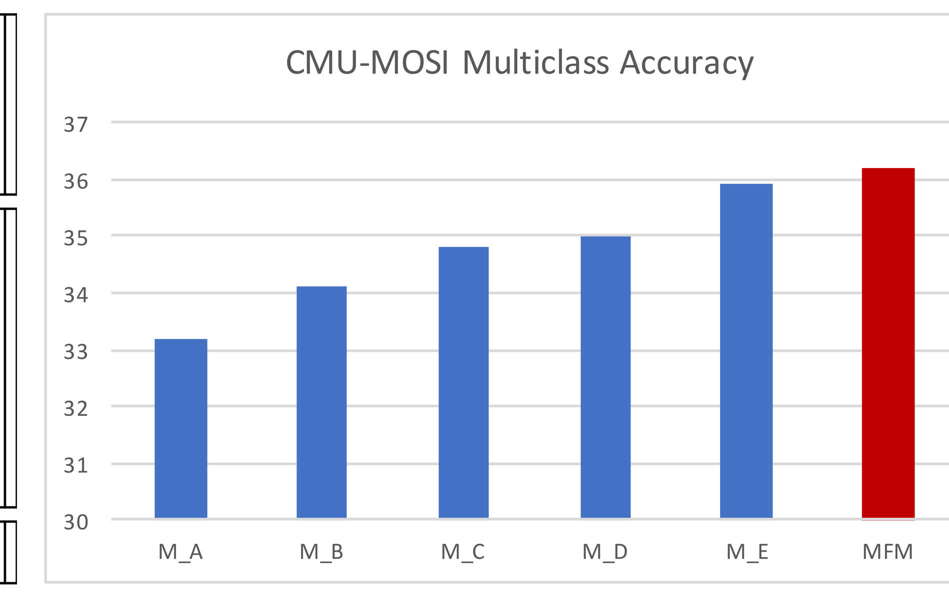
MFM achieves strong performance on 6 multimodal time-series datasets

MFM can be applied on any multimodal fusion encoder

Ablation Studies

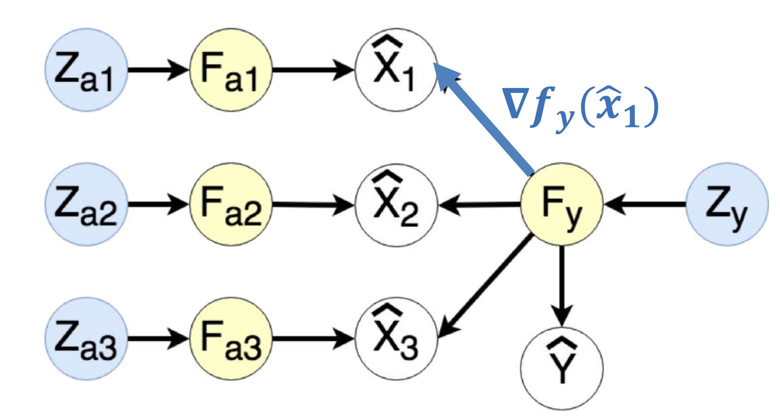


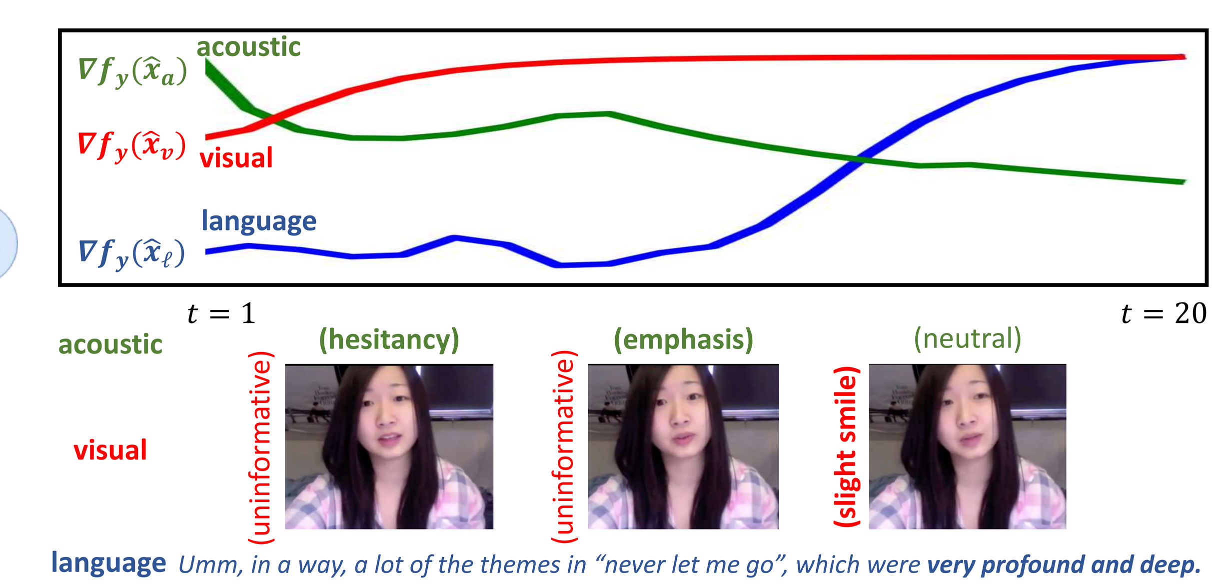
Model	Multimodal Disc. Factor	Hybrid Gen.-Disc. Objective	Factorized Gen.-Disc. Factors	Mod.-Spec. Gen. Factors
M_A	no	no	-	-
M_B	yes	no	-	-
M_C	no	yes	no	-
M_D	yes	yes	no	-
M_E	yes	yes	yes	no
MFM	yes	yes	yes	yes



IMPORTANT:
(1) Multimodal discriminative factor (2) Modality-specific generative factors (3) Hybrid generative-discriminative objective

Analyzing Multimodal Representations





Code and Models:
http://github.com/pliang279/factorized