ICML
International Conference
On Machine Learning

UNIVERSITÄT DES SAARLANDES

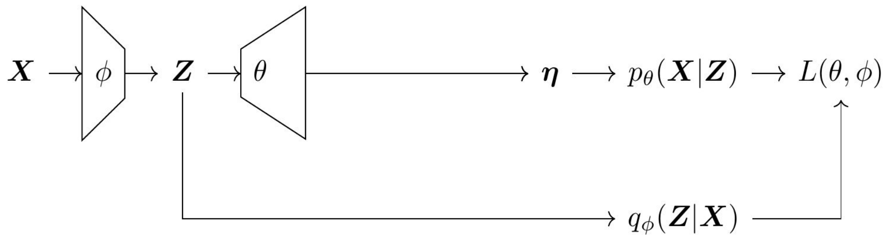
# Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization



Adrián Javaloy Maryam Meghdadi Isabel Valera

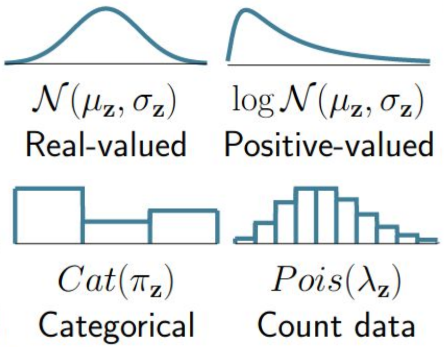
# Problem motivation

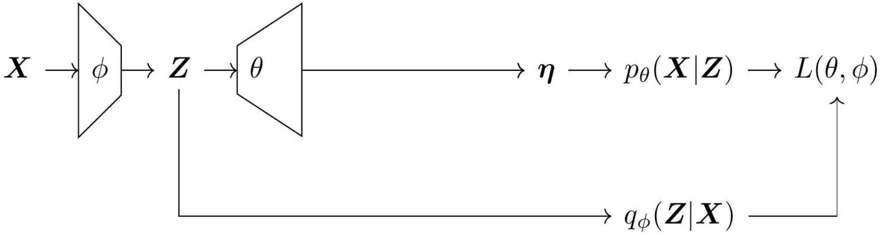
# Problem motivation



# Problem motivation

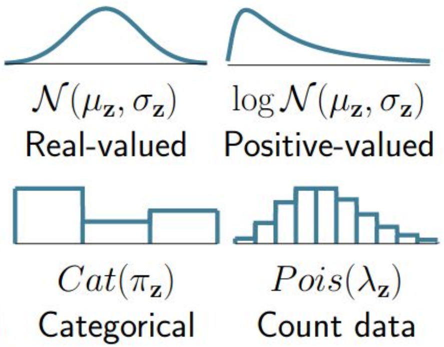
Each modality is of a different type:

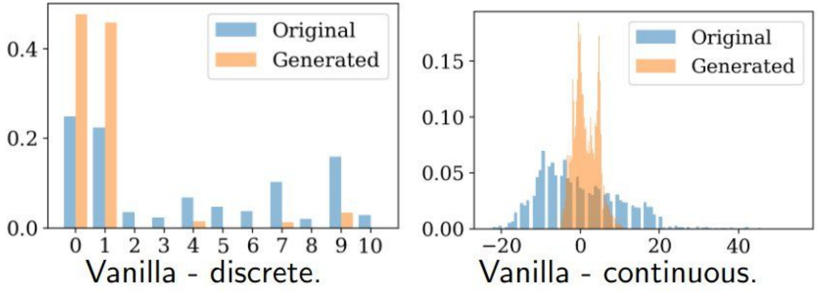


# Problem motivation

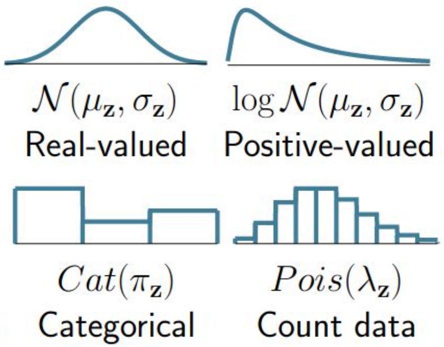
Each modality is of a different type:



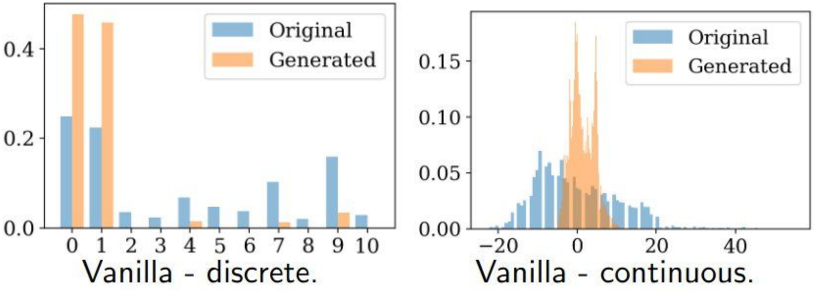
# Problem motivation

Each modality is of a different type:



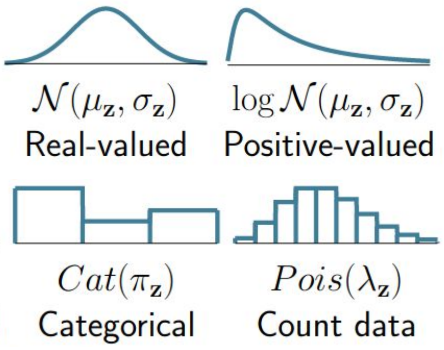

**Likelihood impartiality**

We aim for a learning process that does not prioritize learning any of the different likelihood modalities.



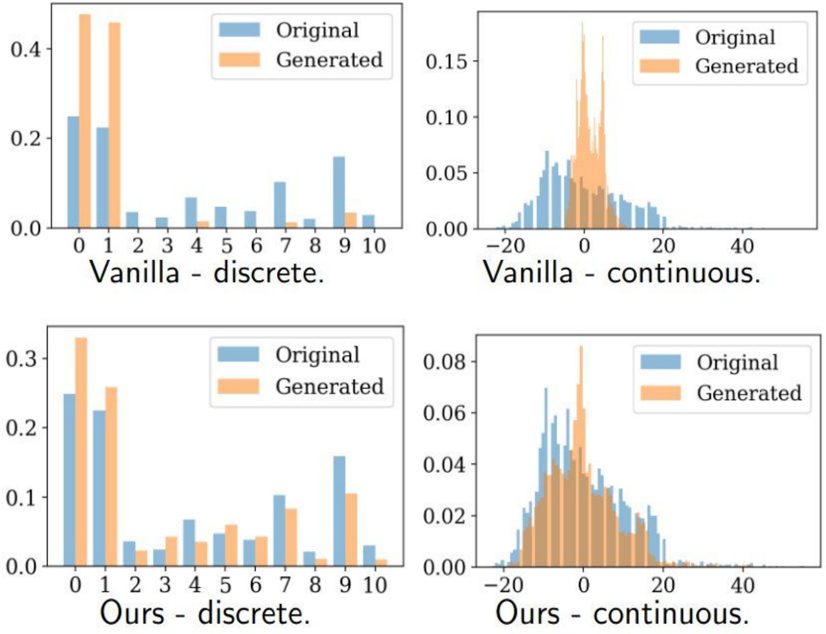
# Problem motivation

Each modality is of a different type:



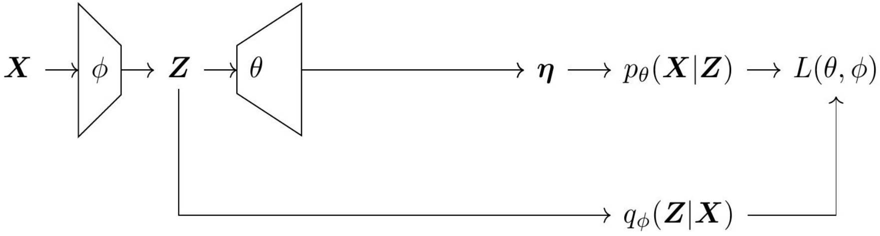

**Likelihood impartiality**

We aim for a learning process that does not prioritize learning any of the different likelihood modalities.

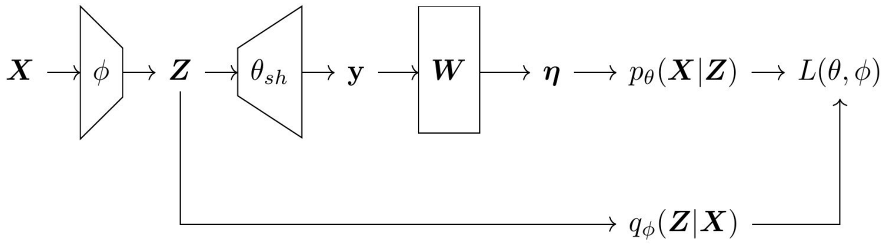


# Analyzing modality collapse

# Analyzing modality collapse



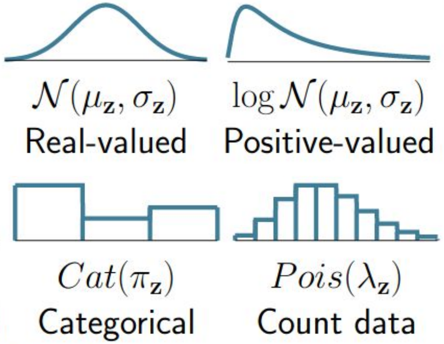
# Analyzing modality collapse

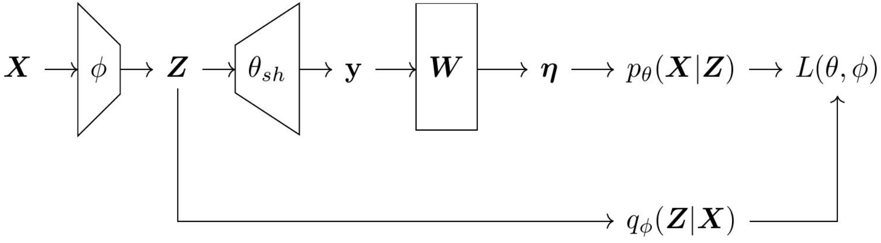


# Analyzing modality collapse



Each modality is of a different type:



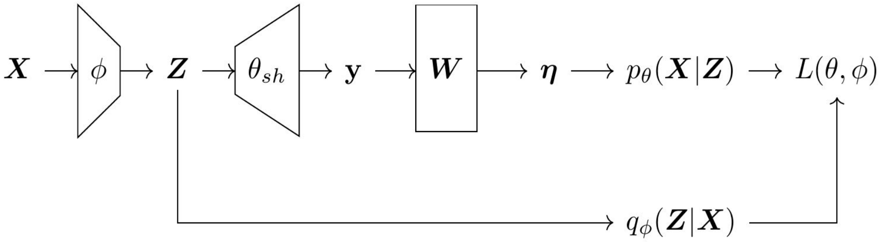



# Analyzing modality collapse

UNIVERSITÄT DES SAARLANDES

Each modality is of a different type:

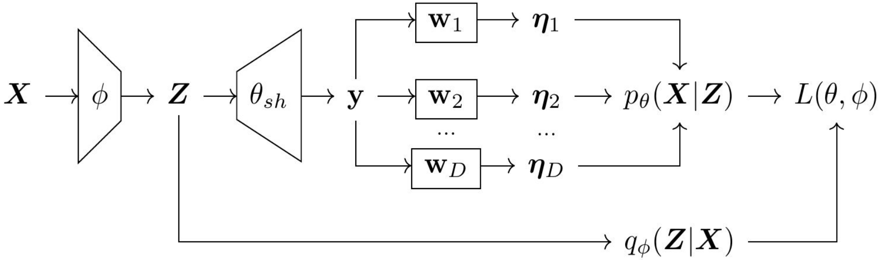
$$p_\theta(\boldsymbol{X}|\mathbf{z}) = \prod_{d=1}^{D} p_d(\mathbf{x}_d|\mathbf{z}).$$



# Analyzing modality collapse

Each modality is of a different type:

$$p_\theta(\boldsymbol{X}|\mathbf{z}) = \prod_{d=1}^{D} p_d(\mathbf{x}_d|\mathbf{z}).$$

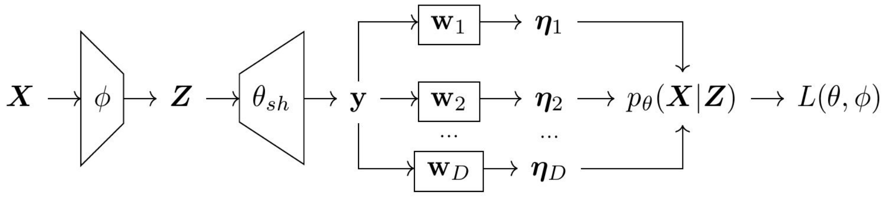


# Analyzing modality collapse



Each modality is of a different type:

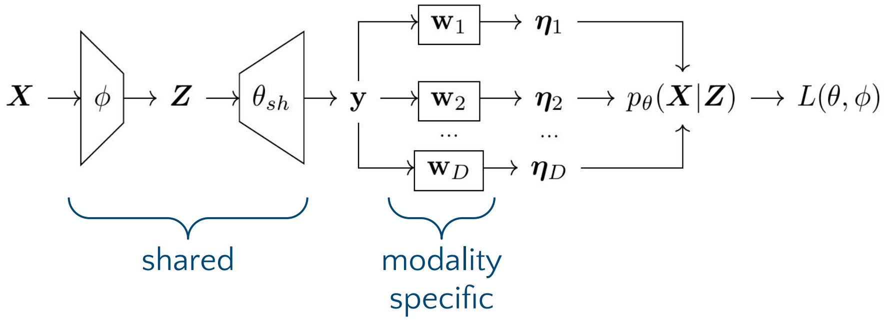
$$p_\theta(\boldsymbol{X}|\mathbf{z}) = \prod_{d=1}^{D} p_d(\mathbf{x}_d|\mathbf{z}).$$



# Analyzing modality collapse

Each modality is of a different type:

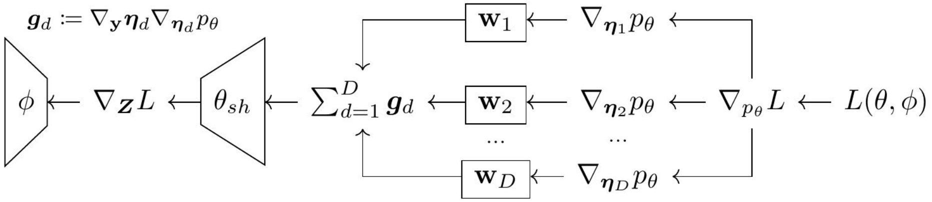
$$p_\theta(\boldsymbol{X}|\mathbf{z}) = \prod_{d=1}^{D} p_d(\mathbf{x}_d|\mathbf{z}).$$



# Analyzing modality collapse

Each modality is of a different type:

$$p_\theta(\boldsymbol{X}|\mathbf{z}) = \prod_{d=1}^{D} p_d(\mathbf{x}_d|\mathbf{z}).$$

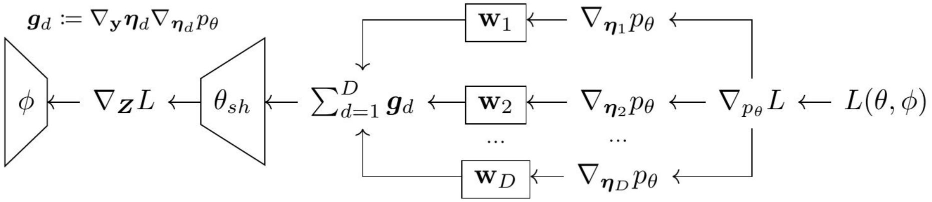


# Analyzing modality collapse



Each modality is of a different type:

$$p_\theta(\boldsymbol{X}|\mathbf{z}) = \prod_{d=1}^{D} p_d(\mathbf{x}_d|\mathbf{z}).$$

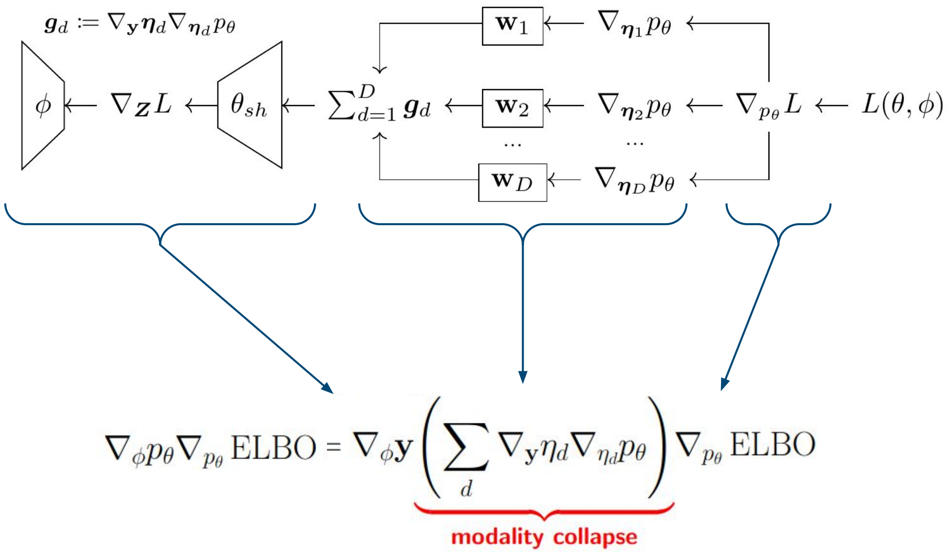



$$\nabla_\phi p_\theta \nabla_{p_\theta} \mathrm{ELBO}$$

# Analyzing modality collapse

Each modality is of a different type:

$$p_\theta(\boldsymbol{X}|\mathbf{z}) = \prod_{d=1}^{D} p_d(\mathbf{x}_d|\mathbf{z}).$$



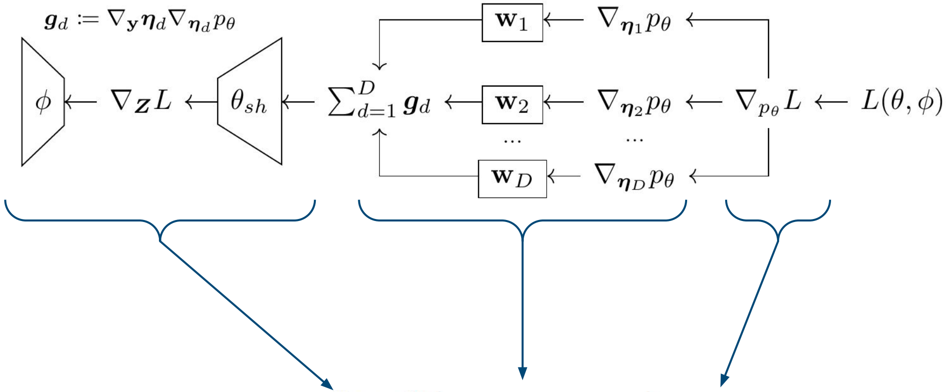
# Analyzing modality collapse

UNIVERSITÄT DES SAARLANDES

Each modality is of a different type:

$$p_\theta(\boldsymbol{X}|\mathbf{z}) = \prod_{d=1}^{D} p_d(\mathbf{x}_d|\mathbf{z}).$$

Modality collapse




$$\nabla_\phi p_\theta \nabla_{p_\theta} \text{ELBO} = \nabla_\phi \mathbf{y} \left( \sum_d \nabla_{\mathbf{y}} \eta_d \nabla_{\eta_d} p_\theta \right) \nabla_{p_\theta} \text{ELBO}$$

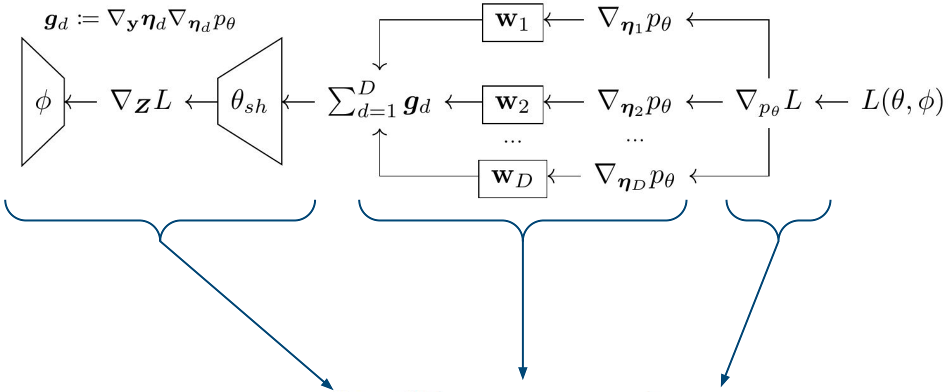
modality collapse

# Analyzing modality collapse

Each modality is of a different type:

$$p_\theta(\boldsymbol{X}|\mathbf{z}) = \prod_{d=1}^{D} p_d(\mathbf{x}_d|\mathbf{z}).$$

Modality collapse ← Gradient conflict



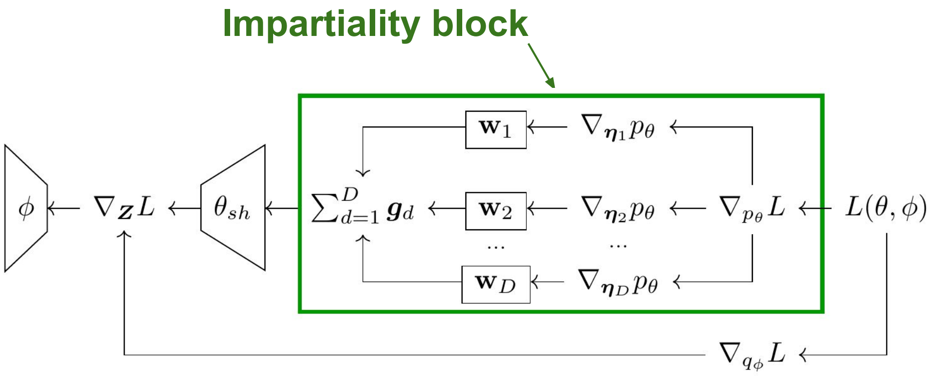

$$\nabla_\phi p_\theta \nabla_{p_\theta} \text{ELBO} = \nabla_\phi \mathbf{y} \underbrace{\left( \sum_d \nabla_{\mathbf{y}} \eta_d \nabla_{\eta_d} p_\theta \right)}_{\text{modality collapse}} \nabla_{p_\theta} \text{ELBO}$$

# Analyzing modality collapse



Each modality is of a different type:

$$p_\theta(\boldsymbol{X}|\mathbf{z}) = \prod_{d=1}^{D} p_d(\mathbf{x}_d|\mathbf{z}).$$




Modality collapse ← Gradient conflict

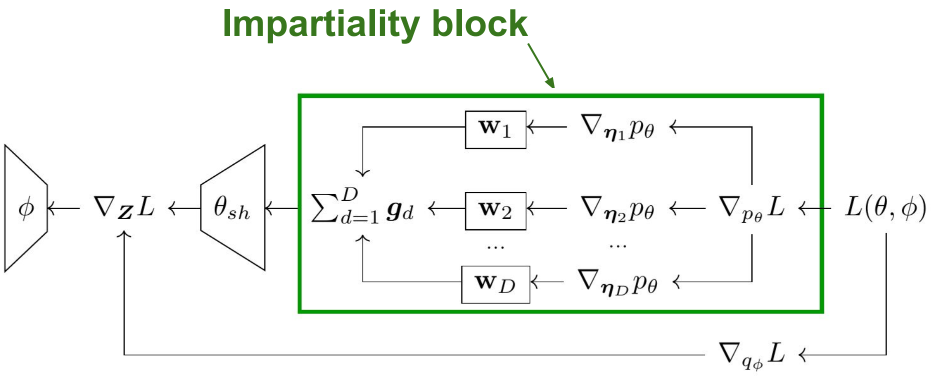
$\subset$ ↓

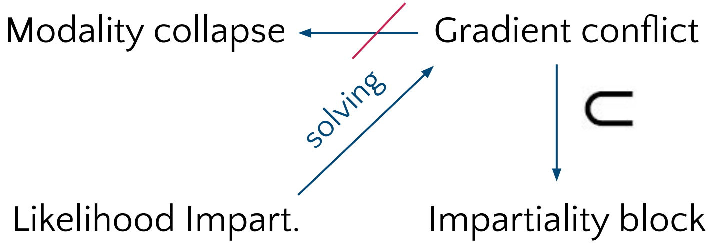
Impartiality block

# Analyzing modality collapse

Each modality is of a different type:

$$p_\theta(\boldsymbol{X}|\mathbf{z}) = \prod_{d=1}^{D} p_d(\mathbf{x}_d|\mathbf{z}).$$

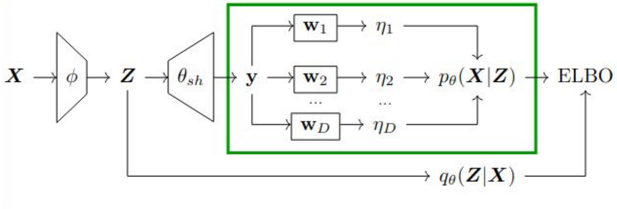


# A familiar face



## Multimodal VAEs

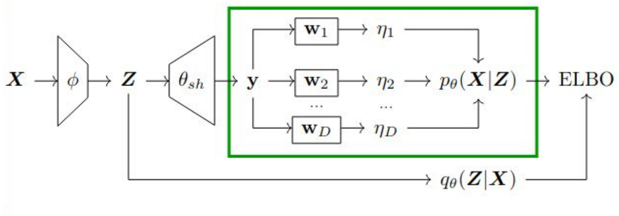



**Want:** Model all modalities equally well.
**Problem:** Modality collapse.
**Shared params:** $\phi$ and $\theta_{sh}$.
**Excl. params:** $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$.
**Updating** $\phi$**:**

$$\nabla_\phi p_\theta \nabla_{p_\theta} \mathrm{ELBO} =$$
$$= \nabla_\phi \mathbf{y} \underbrace{\left(\sum_d \nabla_{\mathbf{y}} \eta_d \nabla_{\eta_d} p_\theta\right)}_{\text{modality collapse}} \nabla_{p_\theta} \mathrm{ELBO}$$

# A familiar face

**Multimodal VAEs**




**Want:** Model all modalities equally well.

**Problem:** Modality collapse.

**Shared params:** $\phi$ and $\theta_{sh}$.

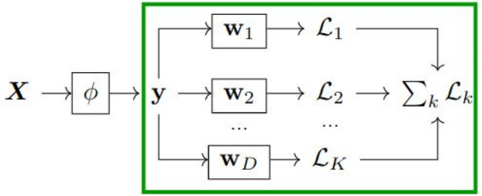
**Excl. params:** $\mathbf{w_1}, \mathbf{w_2}, \dots, \mathbf{w_K}$.

**Updating** $\phi$**:**

$$\nabla_\phi p_\theta \nabla_{p_\theta} \mathrm{ELBO} = $$

$$= \nabla_\phi \mathbf{y} \underbrace{\left( \sum_d \nabla_\mathbf{y} \eta_d \nabla_{\eta_d} p_\theta \right)}_{\text{modality collapse}} \nabla_{p_\theta} \mathrm{ELBO}$$

**Multitask Learning**




**Want:** Learn all tasks equally well.

**Problem:** Negative transfer.

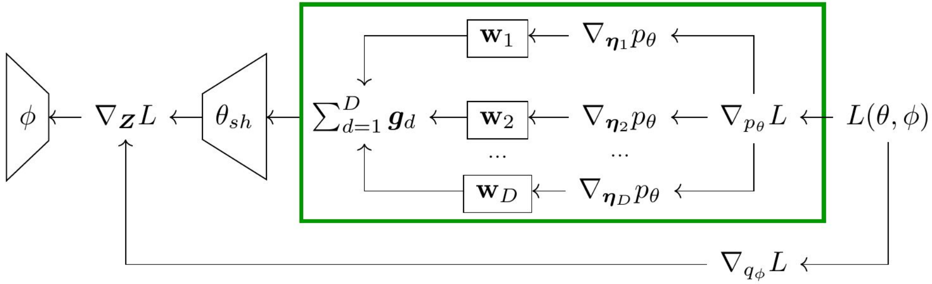
**Shared params:** $\phi$.

**Excl. params:** $\mathbf{w_1}, \mathbf{w_2}, \dots, \mathbf{w_K}$.

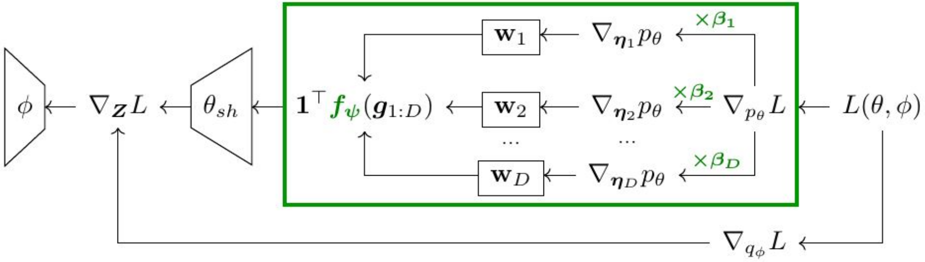
**Updating** $\phi$**:**

$$\nabla_\phi \mathcal{L} = \nabla_\phi \mathbf{y} \underbrace{\left( \sum_k \nabla_\mathbf{y} \mathcal{L}_k \right)}_{\text{negative transfer}}$$

# How to achieve impartiality



# How to achieve impartiality

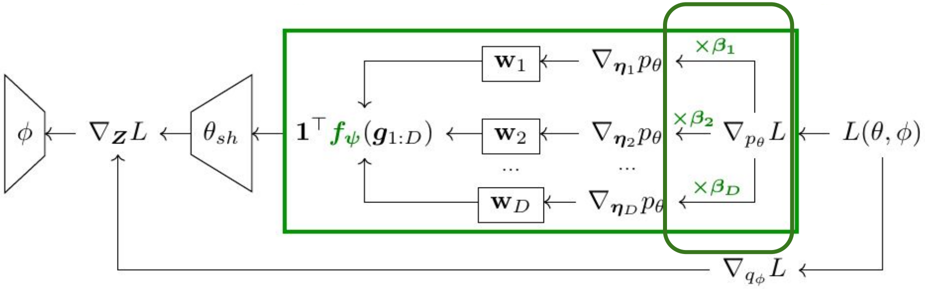



**Algorithm 1** Backward pass within the impartiality block.

**Input:** Output gradient, $\nabla_{p_\theta} L$.
**for** $d = 1$ **to** $D$ **do**
  $\boldsymbol{h}_d \leftarrow \boldsymbol{\beta_d} \nabla_{\eta_d} p_\theta \nabla_{p_\theta} L$
  $\nabla_{\omega_d} L \leftarrow \nabla_{\omega_d} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
  $\boldsymbol{g}_d \leftarrow \nabla_{\boldsymbol{y}} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
**end for**
$\tilde{\boldsymbol{g}}_{1:D} \leftarrow \boldsymbol{f}_\psi(\boldsymbol{g}_{1:D})$
**return** $\sum_d \tilde{\boldsymbol{g}}_d$

Two-step solution:

# How to achieve impartiality



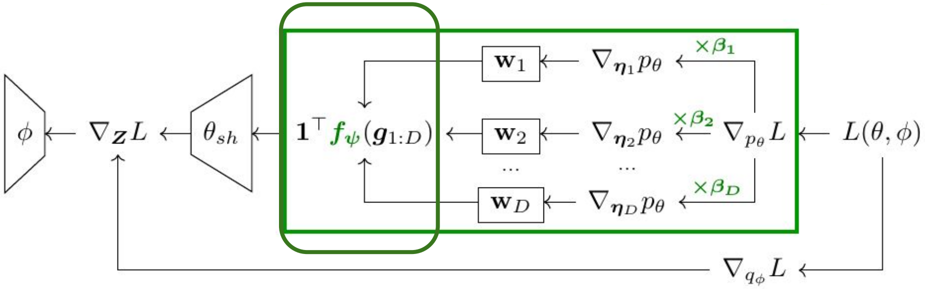

**Algorithm 1** Backward pass within the impartiality block.

**Input:** Output gradient, $\nabla_{p_\theta} L$.
**for** $d = 1$ **to** $D$ **do**
  $\boldsymbol{h}_d \leftarrow \boldsymbol{\beta_d} \nabla_{\boldsymbol{\eta}_d} p_\theta \nabla_{p_\theta} L$
  $\nabla_{\omega_d} L \leftarrow \nabla_{\omega_d} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
  $\boldsymbol{g}_d \leftarrow \nabla_{\boldsymbol{y}} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
**end for**
$\tilde{\boldsymbol{g}}_{1:D} \leftarrow \boldsymbol{f}_\psi(\boldsymbol{g}_{1:D})$
**return** $\sum_d \tilde{\boldsymbol{g}}_d$

Two-step solution:

Local step, β: re-scale gradients to make them comparable.

# How to achieve impartiality




**Algorithm 1** Backward pass within the impartiality block.

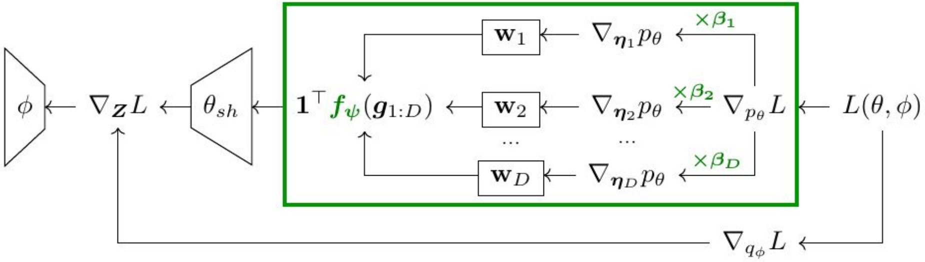
**Input:** Output gradient, $\nabla_{p_\theta} L$.
**for** $d = 1$ **to** $D$ **do**
  $\boldsymbol{h}_d \leftarrow \boldsymbol{\beta_d} \nabla_{\boldsymbol{\eta}_d} p_\theta \, \nabla_{p_\theta} L$
  $\nabla_{\boldsymbol{\omega}_d} L \leftarrow \nabla_{\boldsymbol{\omega}_d} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
  $\boldsymbol{g}_d \leftarrow \nabla_{\boldsymbol{y}} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
**end for**
$\tilde{\boldsymbol{g}}_{1:D} \leftarrow \boldsymbol{f}_\psi(\boldsymbol{g}_{1:D})$
**return** $\sum_d \tilde{\boldsymbol{g}}_d$

Two-step solution:

Local step, β: re-scale gradients to make them comparable.

Global step, f: apply an MTL algorithm to alleviate gradient conflict.

# How to achieve impartiality



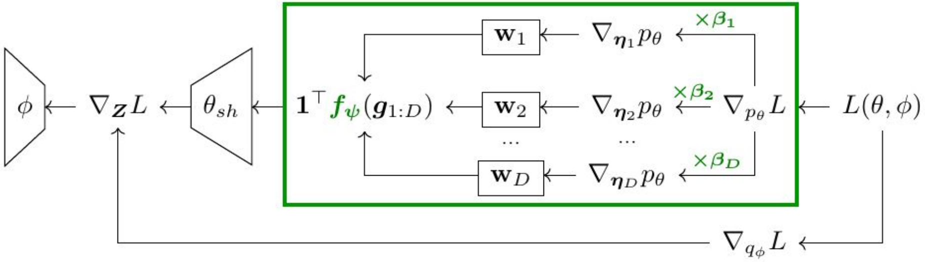

**Algorithm 1** Backward pass within the impartiality block.

**Input:** Output gradient, $\nabla_{p_\theta} L$.
**for** $d = 1$ **to** $D$ **do**
  $\boldsymbol{h}_d \leftarrow \boldsymbol{\beta_d} \nabla_{\eta_d} p_\theta \nabla_{p_\theta} L$
  $\nabla_{\omega_d} L \leftarrow \nabla_{\omega_d} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
  $\boldsymbol{g}_d \leftarrow \nabla_{\boldsymbol{y}} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
**end for**
$\tilde{\boldsymbol{g}}_{1:D} \leftarrow \boldsymbol{f}_\psi(\boldsymbol{g}_{1:D})$
**return** $\sum_d \tilde{\boldsymbol{g}}_d$

The impartiality block is local!

# How to achieve impartiality





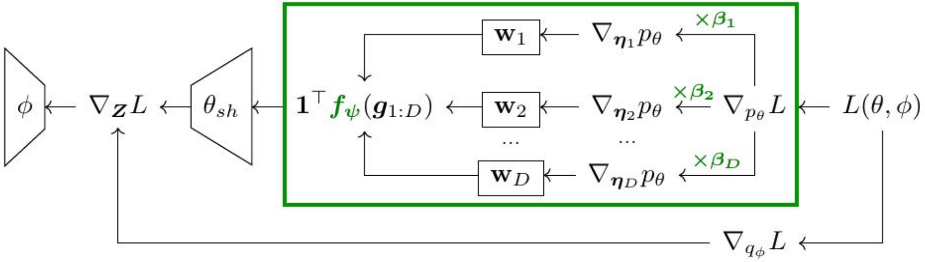

**Algorithm 1** Backward pass within the impartiality block.

**Input:** Output gradient, $\nabla_{p_\theta} L$.
**for** $d = 1$ **to** $D$ **do**
  $\boldsymbol{h}_d \leftarrow \boldsymbol{\beta_d} \nabla_{\eta_d} p_\theta \, \nabla_{p_\theta} L$
  $\nabla_{\omega_d} L \leftarrow \nabla_{\omega_d} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
  $\boldsymbol{g}_d \leftarrow \nabla_{\boldsymbol{y}} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
**end for**
$\tilde{\boldsymbol{g}}_{1:D} \leftarrow \boldsymbol{f}_\psi(\boldsymbol{g}_{1:D})$
**return** $\sum_d \tilde{\boldsymbol{g}}_d$

The impartiality block is local!

It can appear anywhere.

# How to achieve impartiality




**Algorithm 1** Backward pass within the impartiality block.

1: **Input:** Output gradient, $\nabla_{p_\theta} L$.
2: **for** $d = 1$ **to** $D$ **do**
3: $\quad \boldsymbol{h}_d \leftarrow \boldsymbol{\beta_d} \nabla_{\eta_d} p_\theta \nabla_{p_\theta} L$
4: $\quad \nabla_{\omega_d} L \leftarrow \nabla_{\omega_d} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
5: $\quad \boldsymbol{g}_d \leftarrow \nabla_{\boldsymbol{y}} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
6: **end for**
7: $\tilde{\boldsymbol{g}}_{1:D} \leftarrow \boldsymbol{f}_\psi(\boldsymbol{g}_{1:D})$
8: **return** $\sum_d \tilde{\boldsymbol{g}}_d$

The impartiality block is local!

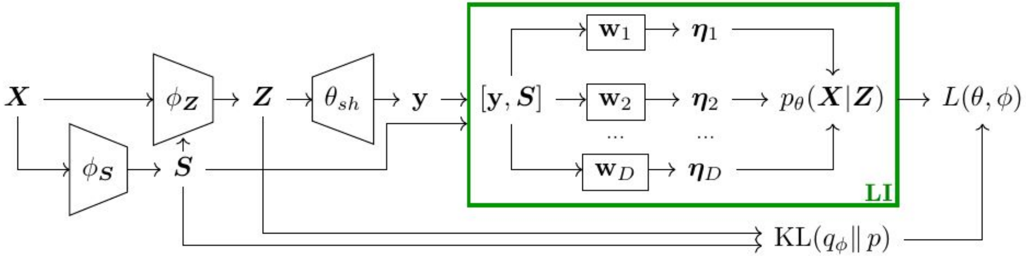
It can appear anywhere.

It can appear more than once.

# VAE extensions

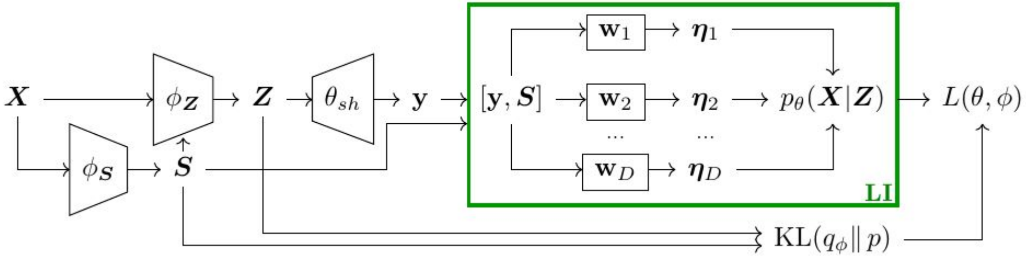
# VAE extensions

HI-VAE



# VAE extensions

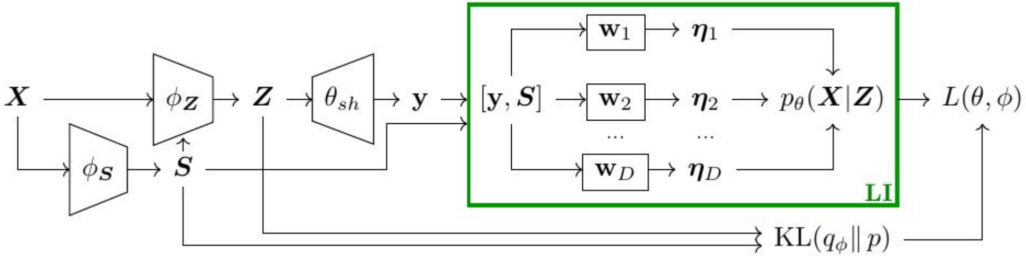
HI-VAE




Two impartiality blocks:
One for y
One for S

# VAE extensions

HI-VAE

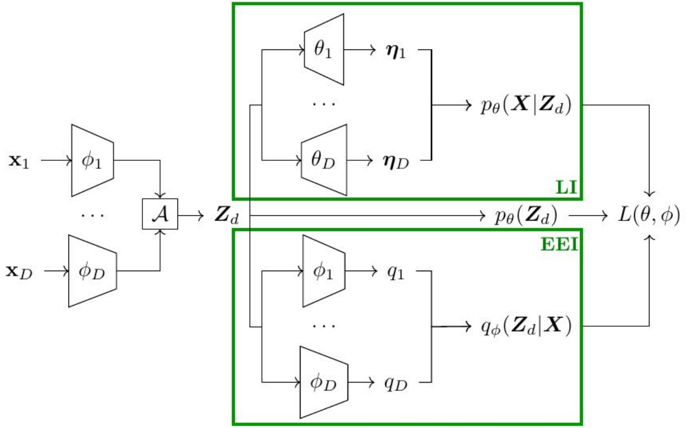



Two impartiality blocks:
One for y
One for S

MVAE

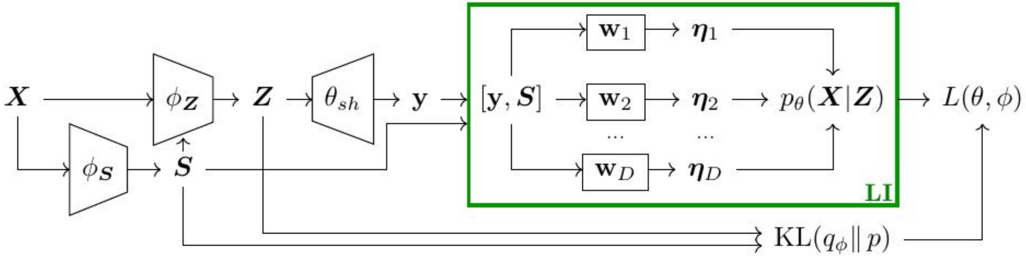
MMVAE

MoPoE



# VAE extensions

HI-VAE

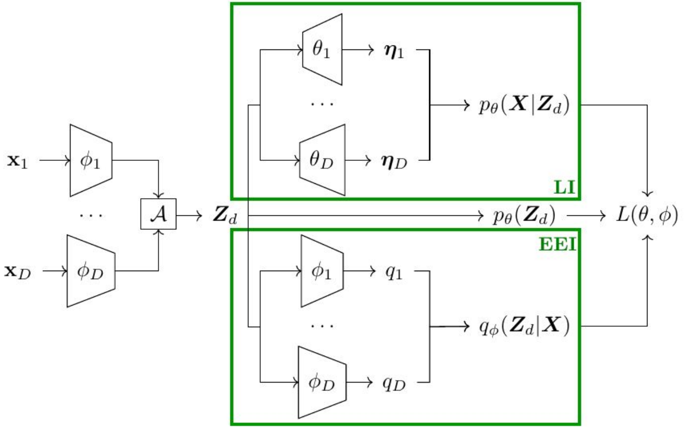



Two impartiality blocks:
One for y
One for S

MVAE

MMVAE

MoPoE




Additional goals:
Encoder Expert-Impartiality
Decoder Expert-Impartiality

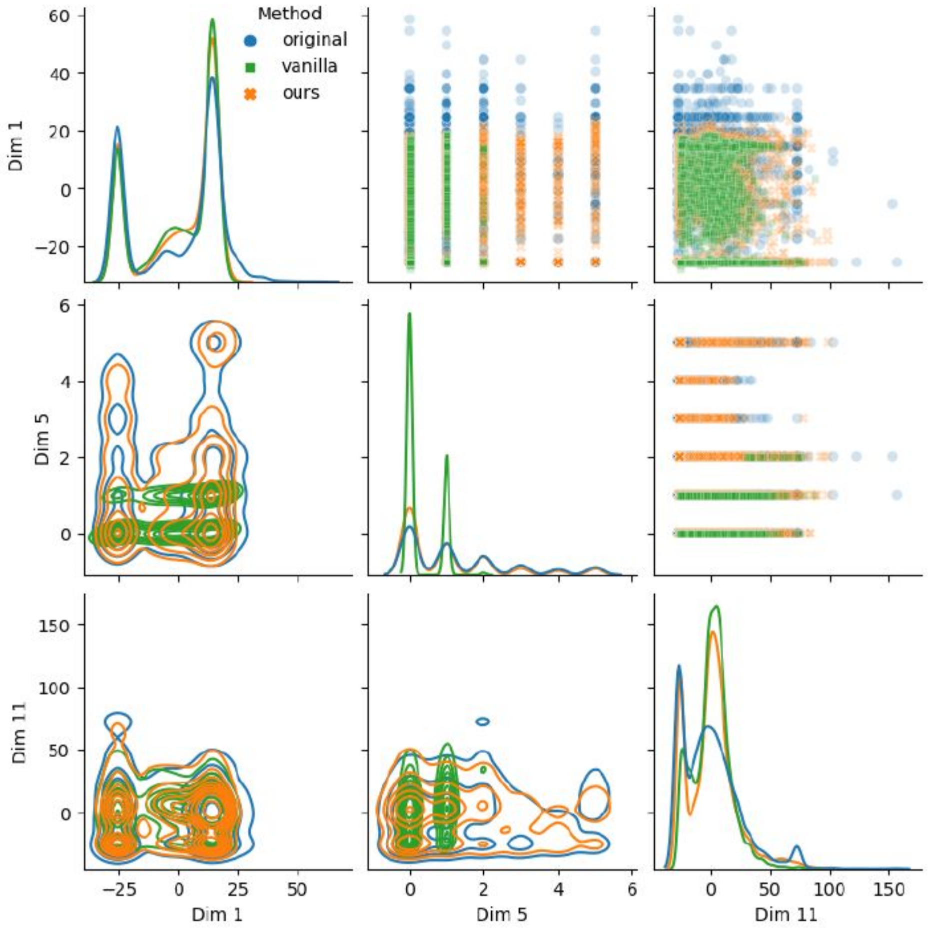
As many impartiality blocks as modalities.

# Results – heterogeneous data

UNIVERSITÄT DES SAARLANDES

*Table 1.* Test reconstruction errors (median over five seeds) for different datasets and VAE models. Statistically different values according to a corrected paired t-test ($\alpha = 0.1$) are shown in bold. Models trained with our approach outperforms the baseline in most cases.

| | | | Heterogeneous | | | | | | | | | Homogeneous | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *Adult* | *Credit* | *Wine* | *Diam.* | *Bank* | *IMDB* | *HI* | *rwm5yr* | *labour* | *El Nino* | *Magic* | *BooNE* |
| Standard VAE | ELBO | vanilla | 0.213 | 0.128 | 0.086 | 0.187 | 0.203 | 0.082 | 0.170 | 0.105 | 0.109 | 0.109 | 0.064 | 0.042 |
| Standard VAE | ELBO | ours | **0.104** | **0.041** | **0.071** | **0.139** | **0.043** | **0.032** | **0.041** | **0.026** | **0.063** | **0.068** | 0.058 | 0.039 |
| Standard VAE | IWAE | vanilla | 0.226 | 0.134 | 0.075 | 0.185 | 0.199 | 0.090 | 0.155 | 0.094 | 0.098 | 0.086 | 0.053 | 0.037 |
| Standard VAE | IWAE | ours | **0.129** | **0.051** | **0.066** | **0.125** | 0.076 | **0.035** | **0.042** | **0.032** | **0.066** | **0.061** | 0.048 | **0.035** |
| Standard VAE | DReG | vanilla | 0.234 | 0.132 | 0.077 | 0.176 | 0.191 | 0.088 | 0.153 | 0.094 | 0.096 | 0.085 | 0.050 | 0.037 |
| Standard VAE | DReG | ours | 0.168 | **0.075** | **0.065** | **0.139** | 0.103 | **0.055** | **0.042** | **0.026** | **0.076** | **0.069** | 0.046 | 0.036 |
| | HI-VAE | vanilla | 0.127 | 0.107 | 0.126 | 0.114 | 0.141 | 0.079 | 0.105 | 0.044 | 0.100 | 0.098 | 0.062 | 0.039 |
| | HI-VAE | ours | **0.081** | 0.060 | **0.117** | **0.011** | **0.095** | 0.049 | 0.109 | **0.024** | **0.069** | **0.015** | **0.033** | 0.038 |


Method
original
vanilla
ours
Dim 1
Dim 5
Dim 11
60
40
20
0
−20
6
4
2
0
150
100
50
0
−25
0
25
50
0
2
4
6
0
50
100
150

# Results – multimodal data

UNIVERSITÄT DES SAARLANDES

*Table 5.* Self and cross latent classification accuracy (%) for different models and losses on MNIST-SVHN-Text.

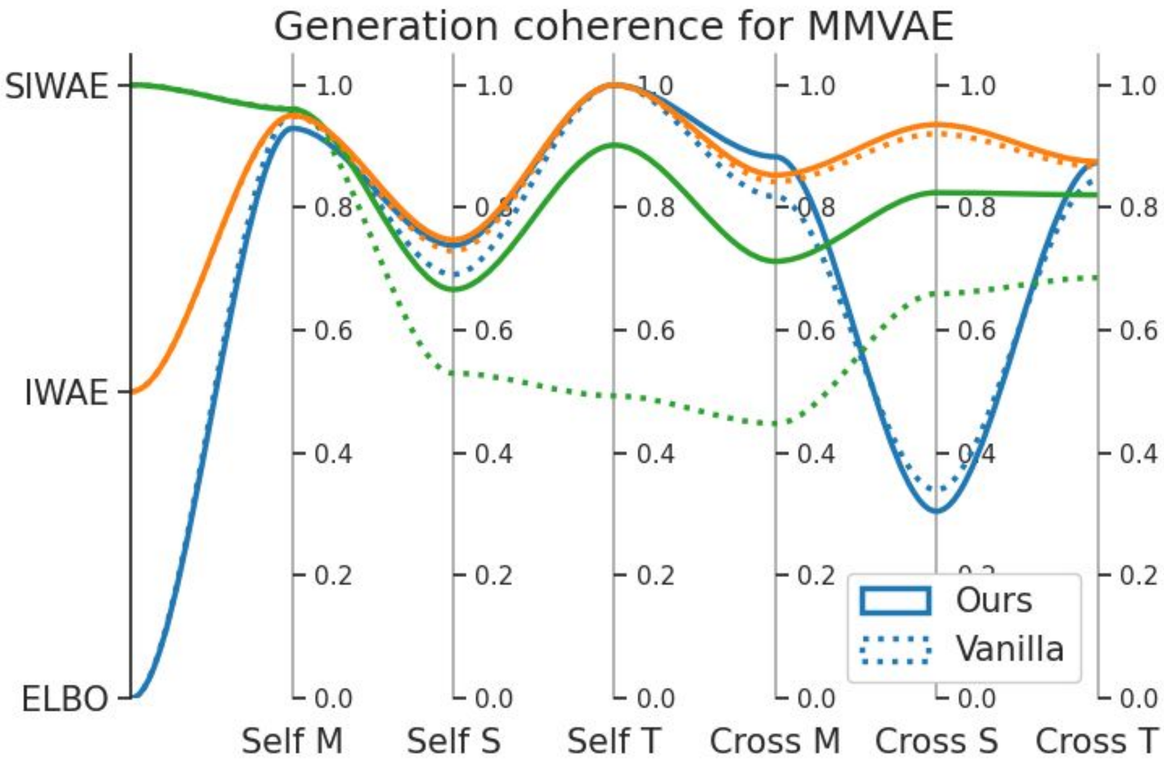
| | | ELBO | IWAE | SIWAE |
|---|---|---|---|---|
| | | Self latent classification | | |
| MVAE | vanilla | 69.68 | 69.14 | 68.58 |
| | ours | 69.95 | 69.06 | **69.75** |
| MMVAE | vanilla | 71.81 | 87.55 | 71.30 |
| | ours | **87.83** | 90.78 | **85.55** |
| MoPoE | vanilla | 89.85 | 87.23 | 67.58 |
| | ours | **91.47** | **90.74** | **69.26** |
| | | Cross latent classification | | |
| MVAE | vanilla | 33.60 | 39.15 | 38.36 |
| | ours | 35.25 | **49.73** | **46.23** |
| MMVAE | vanilla | 44.25 | 76.81 | 40.60 |
| | ours | **71.42** | **84.80** | **60.50** |
| MoPoE | vanilla | 66.14 | 83.71 | 40.36 |
| | ours | **84.52** | **90.48** | **53.24** |

# Results – multimodal data

*Table 9.* Reconstruction coherence ($A = \{M, S, T\}$) for each modality, model, and dataset.

| | | ELBO | | | IWAE | | | SIWAE | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathbf{x}_d$ | M | S | T | M | S | T | M | S | T |
| MVAE | vanilla | 97.53 | 88.26 | 99.30 | 97.27 | 87.19 | 98.76 | 97.37 | 87.47 | 98.83 |
| | ours | **97.85** | **89.65** | **99.64** | **98.28** | **89.01** | **99.93** | 97.42 | 87.63 | **99.20** |
| MMVAE | vanilla | 86.01 | 45.59 | 89.17 | 85.25 | 84.03 | 88.66 | 58.95 | 61.27 | 63.27 |
| | ours | **89.42** | 45.83 | **91.54** | 87.55 | **86.87** | 90.93 | **74.85** | **73.89** | **81.09** |
| MoPoE | vanilla | 95.72 | 85.86 | 98.01 | 95.82 | 87.55 | 97.93 | 75.10 | 67.16 | 76.61 |
| | ours | **96.50** | **93.60** | **99.14** | **97.29** | **92.93** | **99.00** | 96.91 | 89.01 | 99.28 |

# Results – multimodal data



# Make sure to follow us!

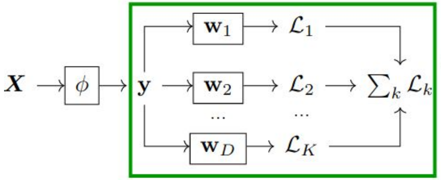
@javaloyML

@maryam_meghdadi

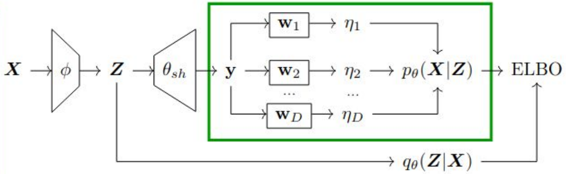
@IValeraM

# Gradient conflict in multimodal VAEs

**Multitask Learning**




**Heterogeneous VAEs**




**Want:** Learn all tasks equally well.
**Problem:** Negative transfer.
**Shared params:** $\phi$.
**Excl. params:** $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$.
**Updating** $\phi$:

$$\nabla_\phi \mathcal{L} = \nabla_\phi \mathbf{y} \underbrace{\left( \sum_k \nabla_\mathbf{y} \mathcal{L}_k \right)}_{\text{negative transfer}}$$

**Want:** Model all features equally well.
**Problem:** Feature overlooking.
**Shared params:** $\phi$ and $\theta_{sh}$.
**Excl. params:** $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$.
**Updating** $\phi$:

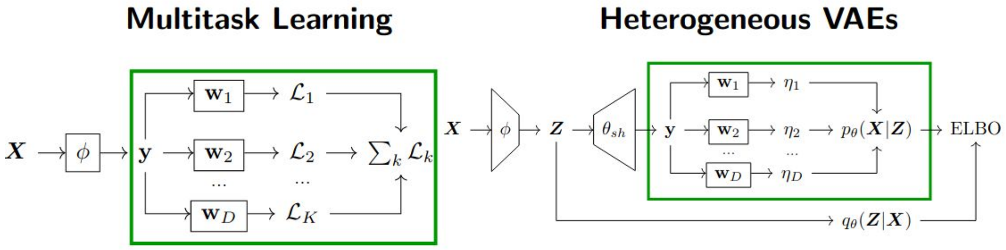
$$\nabla_\phi p_\theta \nabla_{p_\theta} \text{ELBO} = = \nabla_\phi \mathbf{y} \underbrace{\left( \sum_d \nabla_\mathbf{y} \eta_d \nabla_{\eta_d} p_\theta \right)}_{\text{feature overlooking}} \nabla_{p_\theta} \text{ELBO}$$

*Table 3.* Reconstruction coherence ($A = \{M, S, T\}$) for each modality and model, trained using SIWAE.

| | $\mathbf{x}_d$ | M | S | T |
|---|---|---|---|---|
| MVAE | vanilla | 97.37 | 87.47 | 98.83 |
| | ours | 97.42 | 87.63 | **99.20** |
| MMVAE | vanilla | 58.95 | 61.27 | 63.27 |
| | ours | **74.16** | 68.93 | **78.17** |
| MoPoE | vanilla | 75.10 | 67.16 | 76.61 |
| | ours | 96.91 | 89.01 | 99.28 |

| | | | *Adult* | *Credit* | *Wine* |
|---|---|---|---|---|---|
| Standard VAE | ELBO | vanilla | 0.213 | 0.128 | 0.086 |
| | | ours | **0.104** | **0.041** | **0.071** |
| | IWAE | vanilla | 0.226 | 0.134 | 0.075 |
| | | ours | **0.129** | **0.051** | **0.066** |
| | DReG | vanilla | 0.234 | 0.132 | 0.077 |
| | | ours | 0.168 | **0.075** | **0.065** |
| | HI-VAE | vanilla | 0.127 | 0.107 | 0.126 |
| | | ours | **0.081** | 0.060 | **0.117** |

# Gradient conflict in multimodal VAEs




**Want:** Learn all <u>tasks</u> equally well.
**Problem:** Negative transfer.
**Shared params:** $\phi$.
**Excl. params:** $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$.
**Updating** $\phi$:

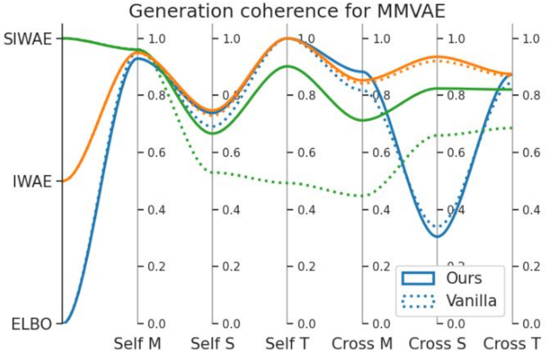
$$\nabla_\phi \mathcal{L} = \nabla_\phi \mathbf{y} \underbrace{\left( \sum_k \nabla_{\mathbf{y}} \mathcal{L}_k \right)}_{\text{negative transfer}}$$

**Want:** Model all <u>features</u> equally well.
**Problem:** Feature overlooking.
**Shared params:** $\phi$ and $\theta_{sh}$.
**Excl. params:** $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$.
**Updating** $\phi$:

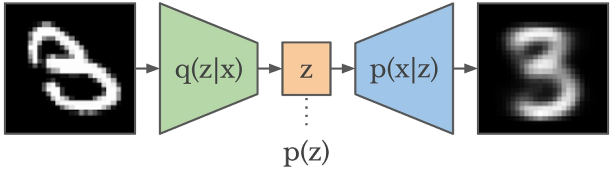
$$\nabla_\phi p_\theta \nabla_{p_\theta} \text{ELBO} = $$
$$= \nabla_\phi \mathbf{y} \underbrace{\left( \sum_d \nabla_{\mathbf{y}} \eta_d \nabla_{\eta_d} p_\theta \right)}_{\text{feature overlooking}} \nabla_{p_\theta} \text{ELBO}$$

*Table 3.* Reconstruction coherence ($A = \{M, S, T\}$) for each modality and model, trained using SIWAE.

| | $\mathbf{x}_d$ | M | S | T |
|---|---|---|---|---|
| MVAE | vanilla | 97.37 | 87.47 | 98.83 |
| | ours | 97.42 | 87.63 | **99.20** |
| MMVAE | vanilla | 58.95 | 61.27 | 63.27 |
| | ours | **74.16** | 68.93 | **78.17** |
| MoPoE | vanilla | 75.10 | 67.16 | 76.61 |
| | ours | 96.91 | 89.01 | 99.28 |



# Promise 2




$$\underset{\theta,\ \phi}{\text{maximize}}\, \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))$$

- **Heterogeneous data.** Each feature $\mathbf{x}_d$ is of a different type:

| $\mathcal{N}(\mu_\mathbf{z}, \sigma_\mathbf{z})$ | $\log \mathcal{N}(\mu_\mathbf{z}, \sigma_\mathbf{z})$ | $Cat(\pi_\mathbf{z})$ | $Pois(\lambda_\mathbf{z})$ |
|---|---|---|---|
| Real-valued | Positive-valued | Categorical | Count data |

Thus, a heterogeneous likelihood is of the form

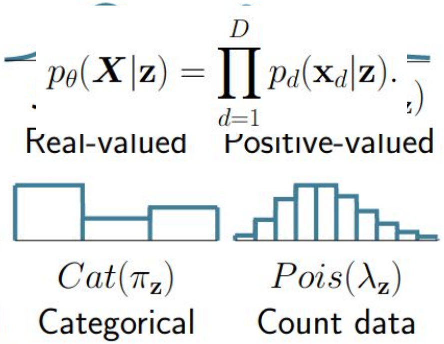
$$p_\theta(\boldsymbol{X}|\mathbf{z}) = \prod_{d=1}^{D} p_d(\mathbf{x}_d|\mathbf{z}).$$

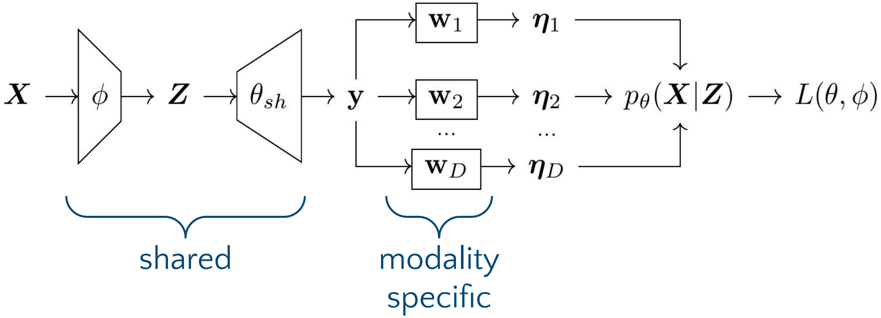
**Likelihood impartiality**

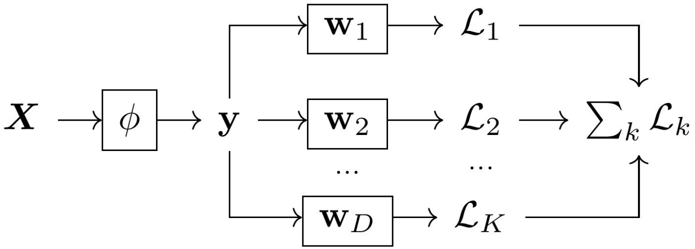
All dimensions are equally important. We want a learning process impartial to the likelihood of each of the dimensions.

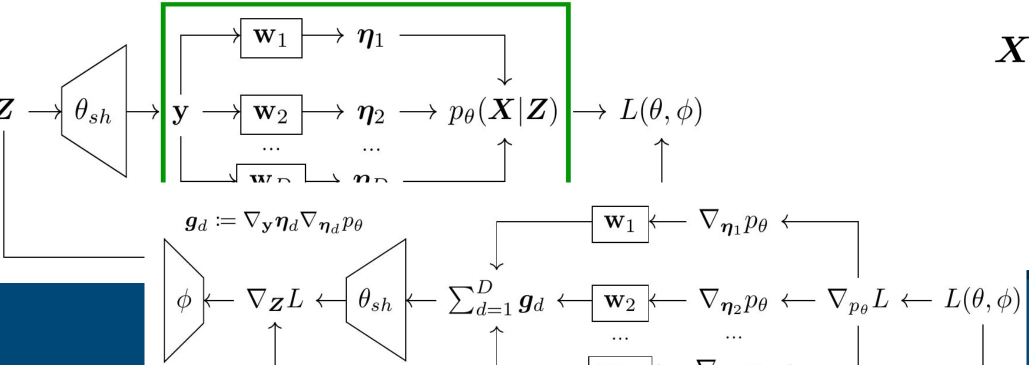
# We noticed

Each modality is of a different type:

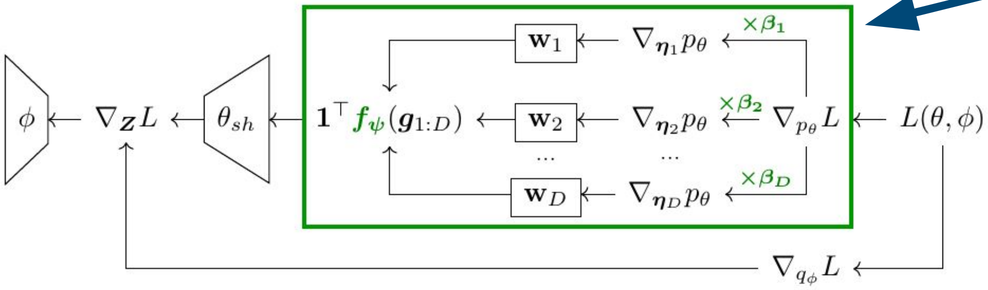




# Solution algorithm



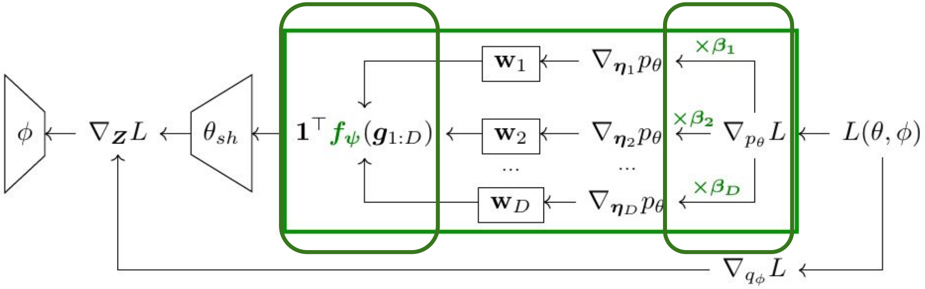

**Algorithm 1** Backward pass within the impartiality block.

**Input:** Output gradient, $\nabla_{p_\theta} L$.
**for** $d = 1$ **to** $D$ **do**
  $\boldsymbol{h}_d \leftarrow \boldsymbol{\beta_d} \nabla_{\eta_d} p_\theta \nabla_{p_\theta} L$
  $\nabla_{\omega_d} L \leftarrow \nabla_{\omega_d} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
  $\boldsymbol{g}_d \leftarrow \nabla_{\boldsymbol{y}} \boldsymbol{\eta}_d \cdot \boldsymbol{h}_d$
**end for**
$\tilde{\boldsymbol{g}}_{1:D} \leftarrow \boldsymbol{f}_\psi(\boldsymbol{g}_{1:D})$
**return** $\sum_d \tilde{\boldsymbol{g}}_d$

**Impartiality block**

- Input: shared
- Split-and-merge structure
- Suffers gradient conflict
- Two-step solution:
  - Local step, β: re-scale grads to make them comparable.
  - Global step, f: apply an MTL algorithm to alleviate gradient conflict.
- *Local* character
  - Anywhere
  - More than once

# How to achieve impartiality




**Algorithm 1** Backward pass within the impartiality block.

1: **Input:** Output gradient, $\nabla_{p_\theta} L$.
2: **for** $d = 1$ **to** $D$ **do**
3: $\quad h_d \leftarrow \beta_d \nabla_{\eta_d} p_\theta \nabla_{p_\theta} L$
4: $\quad \nabla_{\omega_d} L \leftarrow \nabla_{\omega_d} \eta_d \cdot h_d$
5: $\quad g_d \leftarrow \nabla_y \eta_d \cdot h_d$
6: **end for**
7: $\tilde{g}_{1:D} \leftarrow f_\psi(g_{1:D})$
8: **return** $\sum_d \tilde{g}_d$

Two-step solution:

The impartiality block is local!

Local step: rescale gradients to make them comparable.

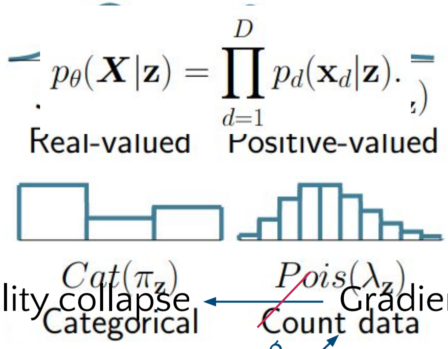
β [illegible] ideally (other solutions possible)

Global step: apply an MTL algorithm.

[illegible] different MTL algorithms to alleviate gradient conflict.

# Analyzing modality collapse

Each modality is of a different type:




Modality collapse ← Gradient conflict → Negative transfer

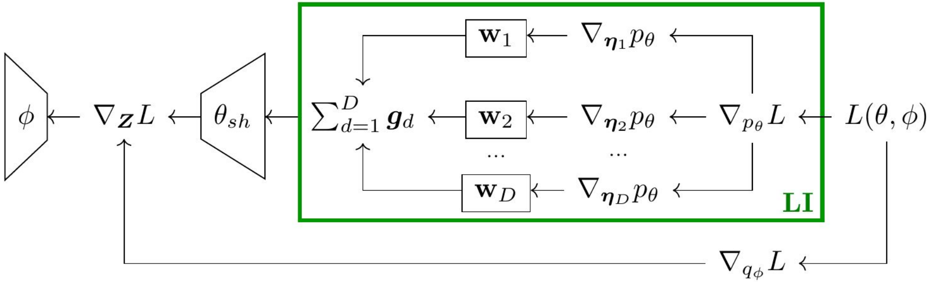
solving

$\subset$

Likelihood Impart.

Impartiality block

Impartiality block




specific