Combining Deep Generative Models and Multi-lingual Pretraining for Semi-supervised Document Classification

Yi Zhu♡, Ehsan Shareghi♡♥, Yingzhen Li♢, Roi Reichart♣, Anna Korhonen♡

♡Language Technology Lab, University of Cambridge, ♥Department of Data Science & AI, Monash University

♢Department of Computing, Imperial College London, ♣Faculty of Industrial Engineering and Management, Technion, IIT

Code: https://github.com/cambridgeltl/mling_sdgms

EACL 2021

Deep generative models+Multi-lingual Pretraining

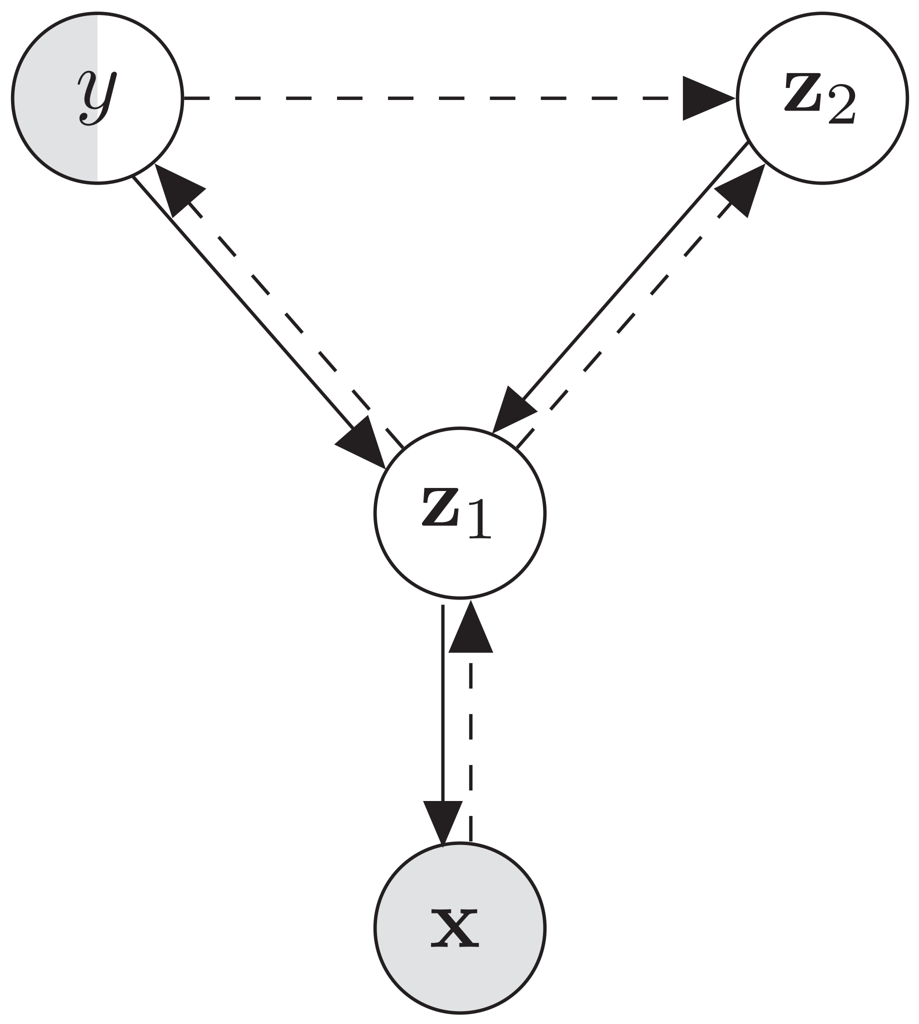
- Deep generative models (**DGMs**)
  - capable of capturing complex data distributions at scale with rich latent representations
- Semi-supervised deep generative models (**SDGMs**)
  - leverage both *labelled* and *unlabelled* data by modelling input and output (label) together

Deep generative models+Multi-lingual Pretraining

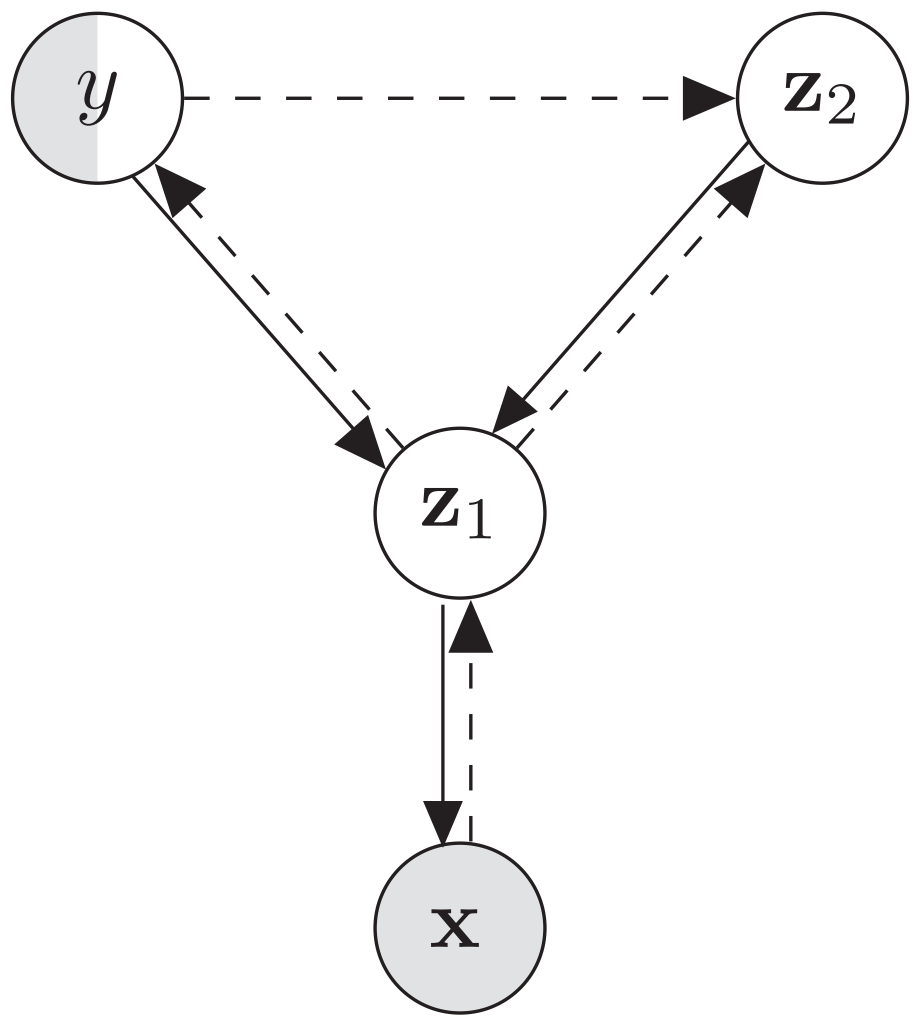
- Deep generative models (**DGMs**)
  - capable of capturing complex data distributions at scale with rich latent representations
- Semi-supervised deep generative models (**SDGMs**)
  - leverage both *labelled* and *unlabelled* data by modelling input and output (label) together

- Effectively use *unlabelled* data through learning shared representations across languages that can be transferred to downstream tasks
- The lack of *labelled* data still leads to inferior performance of the same model compared to those trained in languages with more labelled data such as English

Deep generative models+Multi-lingual Pretraining



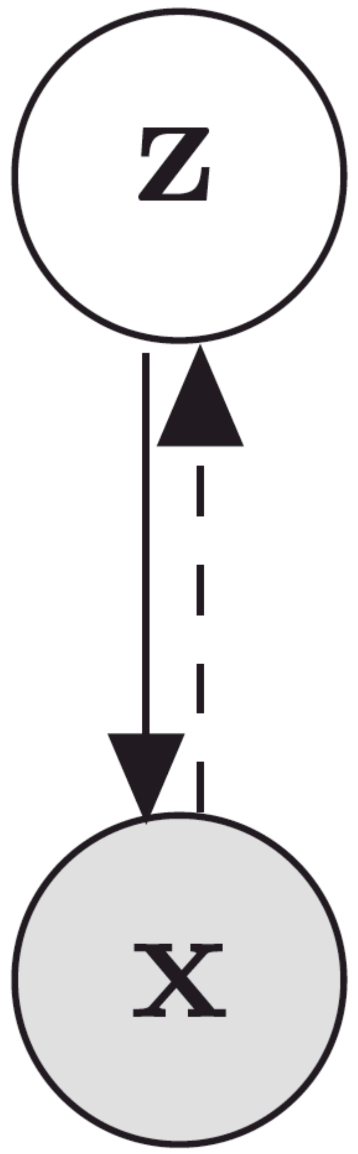
Deep generative models+Multi-lingual Pretraining



- We bridge the gap to form a **pipeline framework** by combining both for *multi-lingual document classification*
  - The multi-lingual pretrained model serves as multi-lingual encoder
  - SDGMs can operate on top of it independently of encoding architecture

DGMs

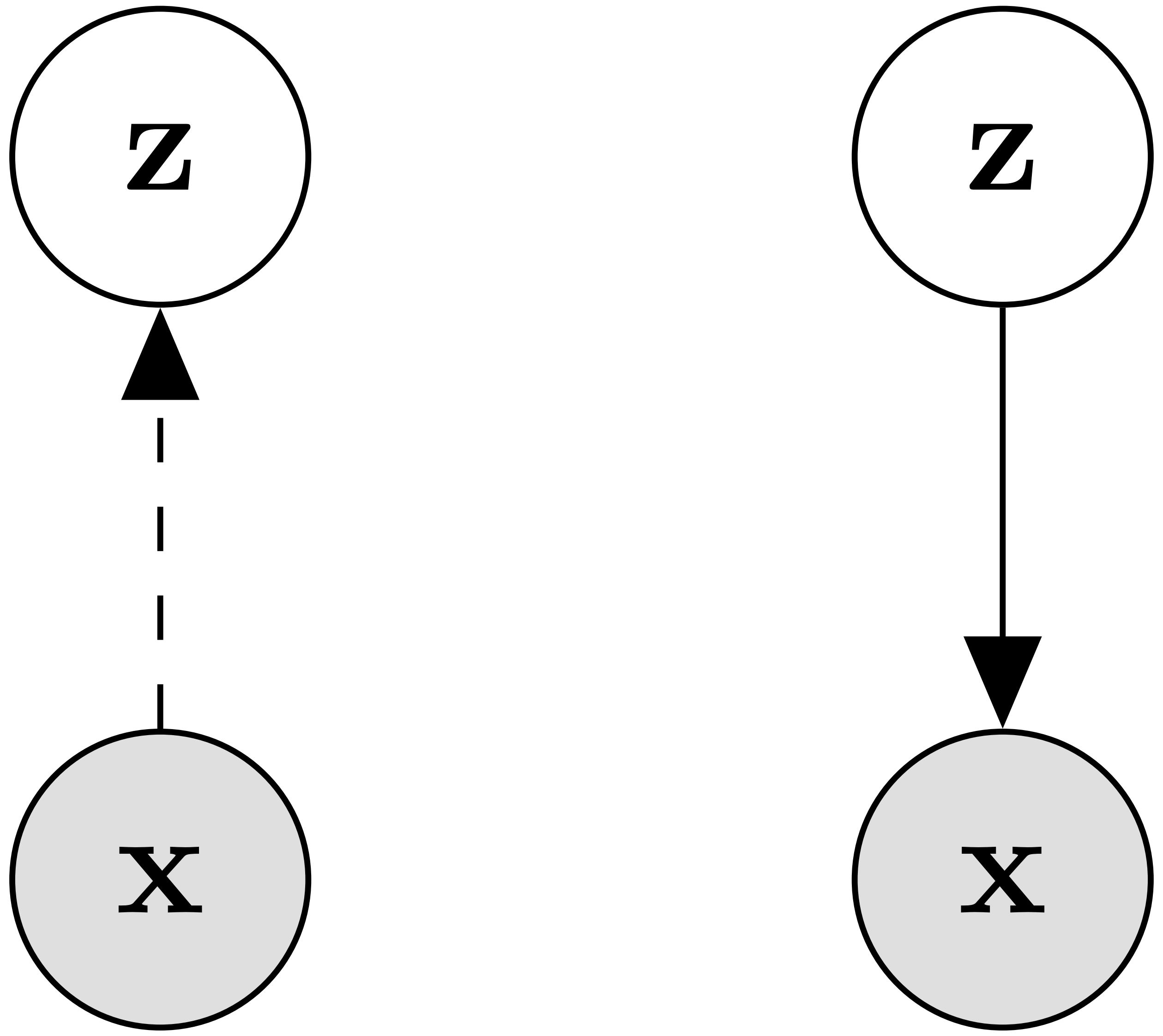
VAE [Kingma et al. 2014]



$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \leq \log p(\mathbf{x})$$

DGMs

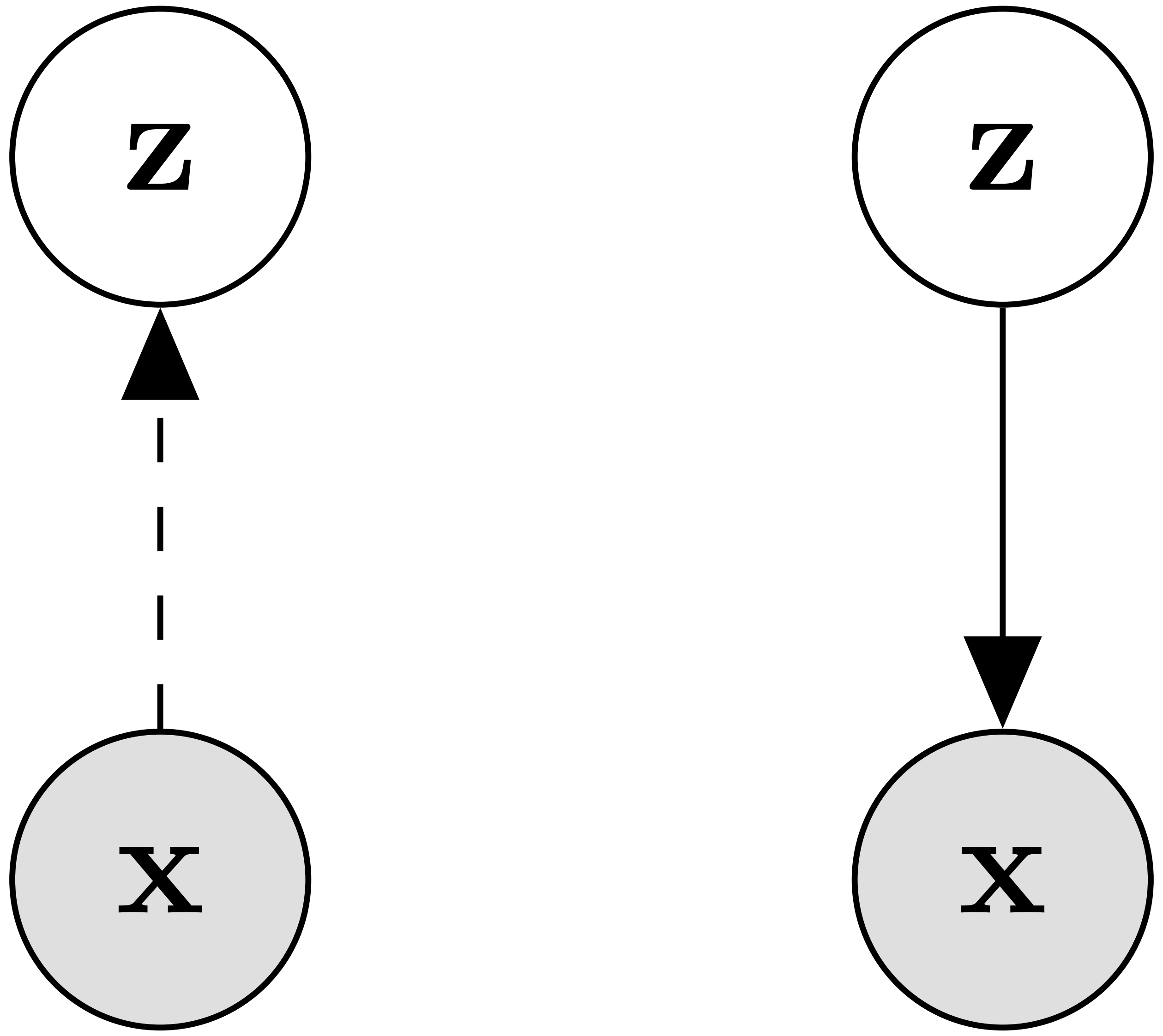
VAE [Kingma et al. 2014]



$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \leq \log p(\mathbf{x})$$

DGMs

VAE [Kingma et al. 2014]

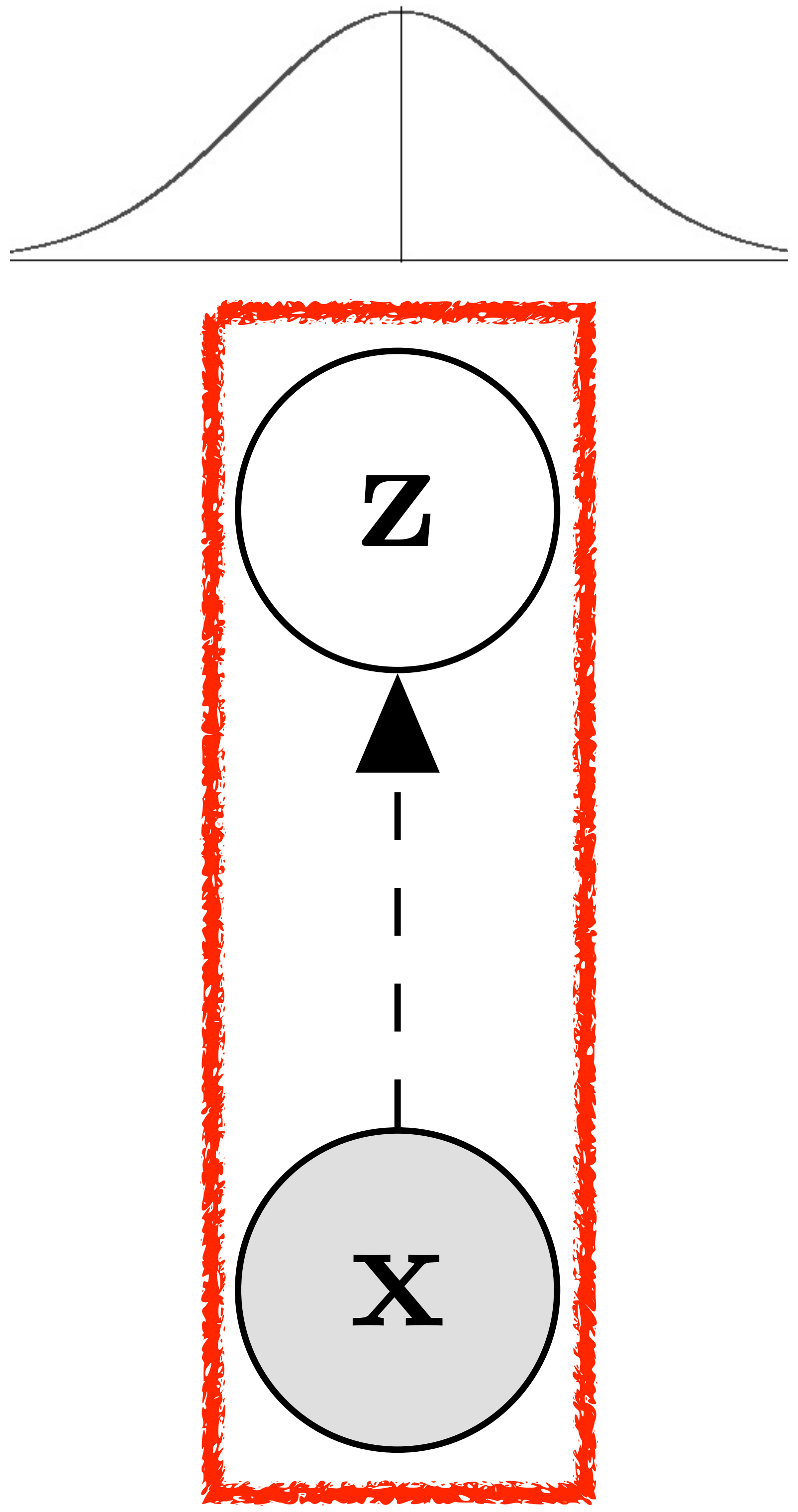


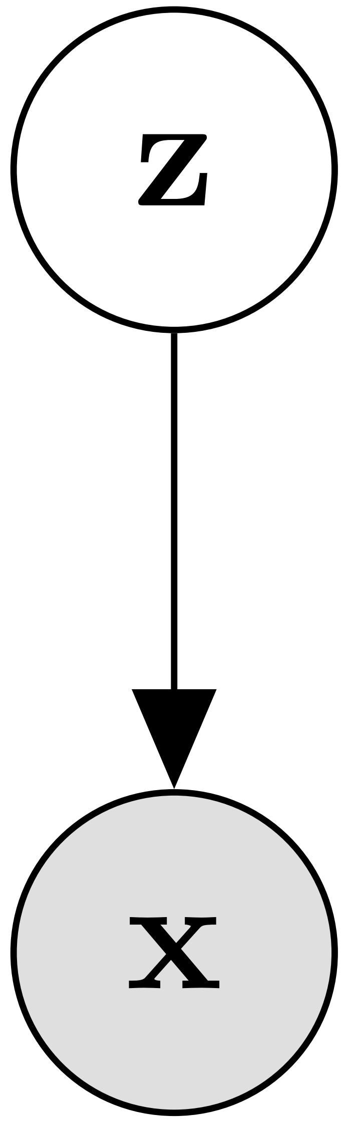
$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \leq \log p(\mathbf{x})$$

$$\underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big]}_{\text{Reconstruction}} - \underbrace{\mathrm{KL}\big(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})\big)}_{\text{KL}}$$

DGMs

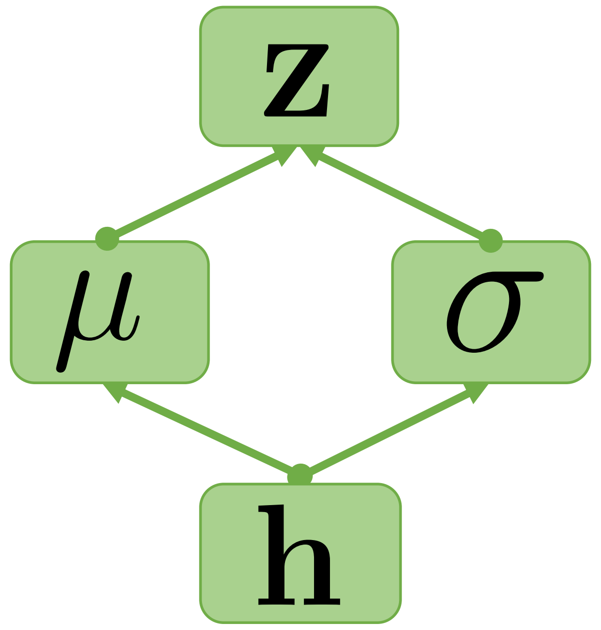
VAE [Kingma et al. 2014]





$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{\boxed{q_\phi(\mathbf{z}|\mathbf{x})}}\right] \leq \log p(\mathbf{x})$$

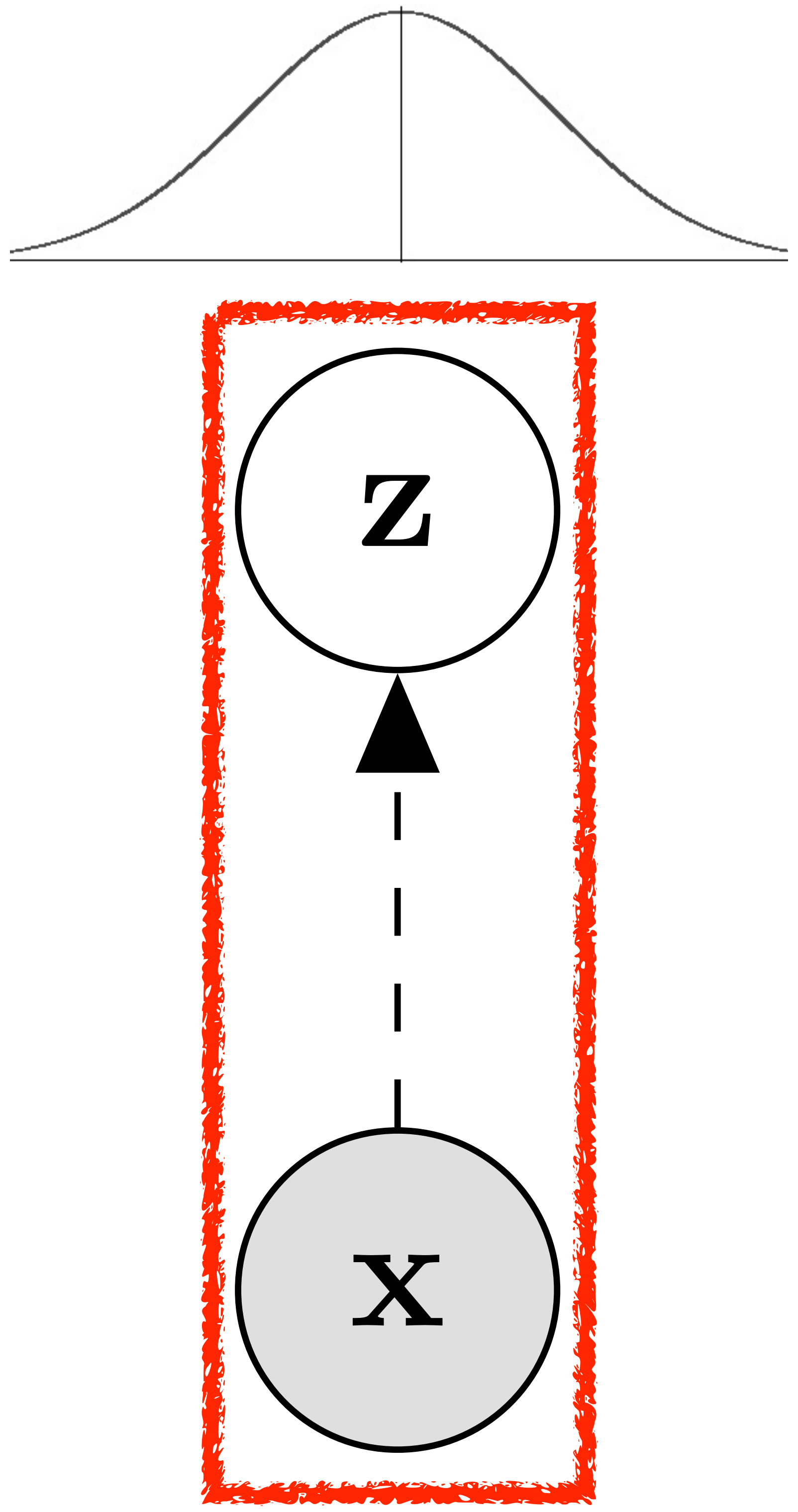
$$\underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{Reconstruction}} - \underbrace{\mathrm{KL}\left(\boxed{q_\phi(\mathbf{z}|\mathbf{x})}\|p(\mathbf{z})\right)}_{\text{KL}}$$

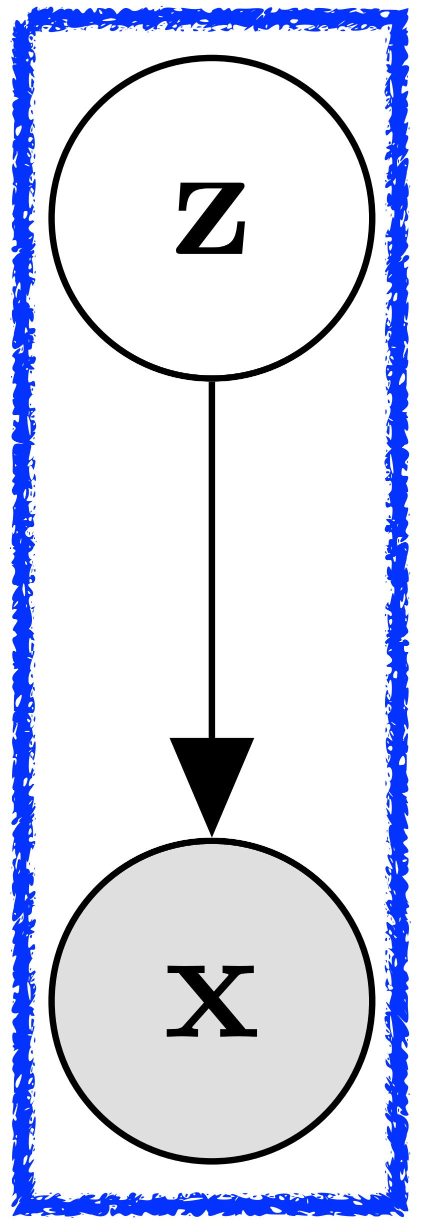


$$\boxed{q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\boldsymbol{\mu}_\phi(\mathbf{h}), \mathrm{diag}\left(\boldsymbol{\sigma}^2_\phi(\mathbf{h})\right)\right)}$$

DGMs

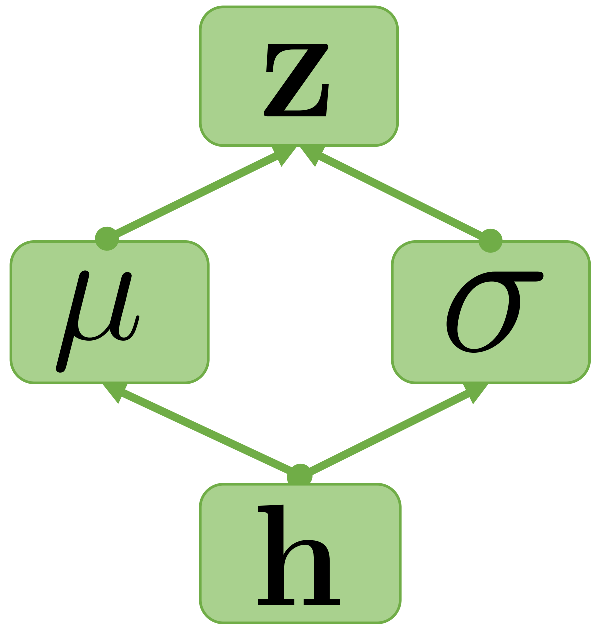
VAE [Kingma et al. 2014]





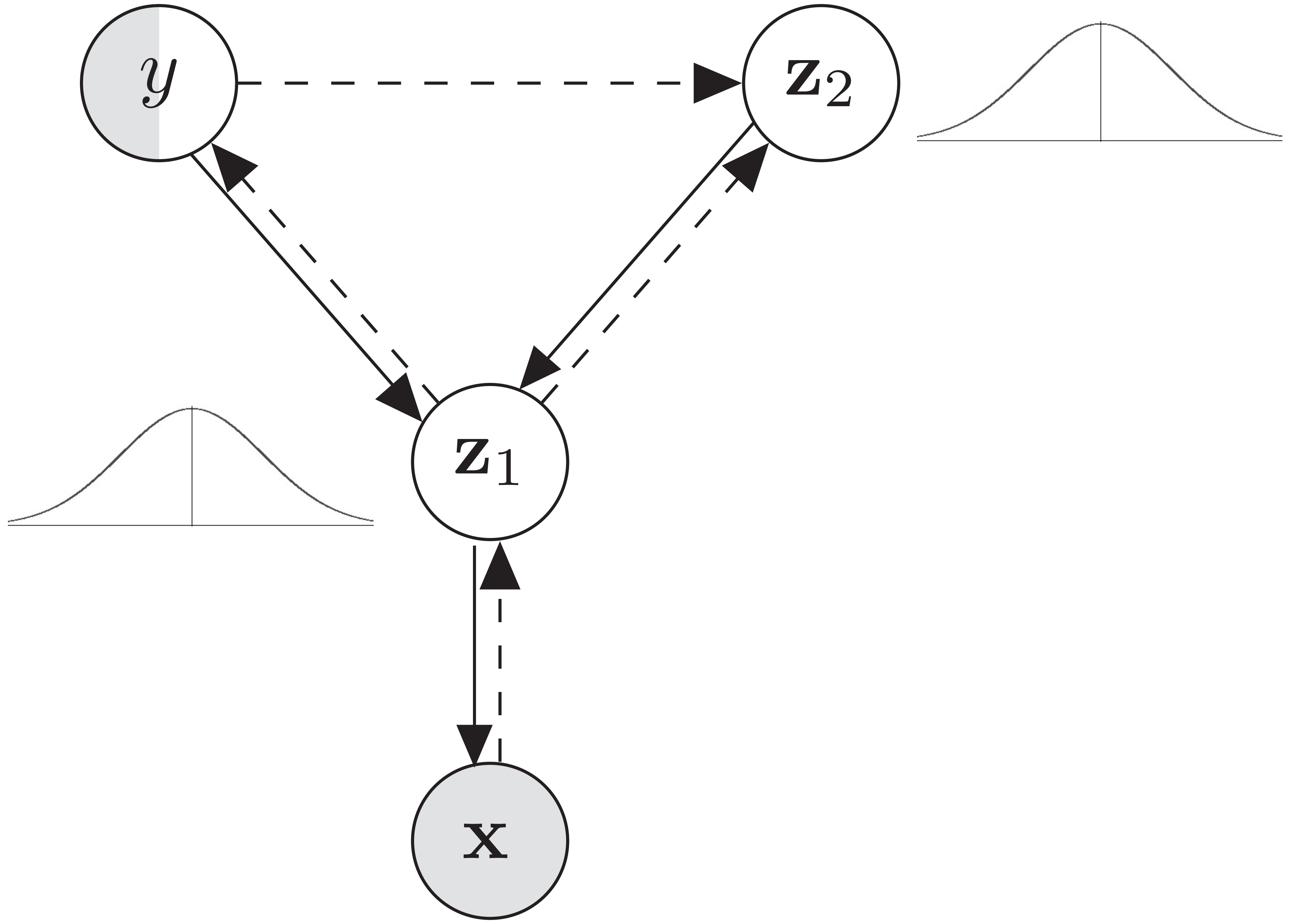
$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{\boxed{p_\theta(\mathbf{x},\mathbf{z})}}{\boxed{q_\phi(\mathbf{z}|\mathbf{x})}}\right] \le \log p(\mathbf{x})$$

$$\underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \boxed{p_\theta(\mathbf{x}|\mathbf{z})}\right]}_{\text{Reconstruction}} - \underbrace{\mathrm{KL}\left(\boxed{q_\phi(\mathbf{z}|\mathbf{x})}\middle\|\boxed{p(\mathbf{z})}\right)}_{\text{KL}}$$

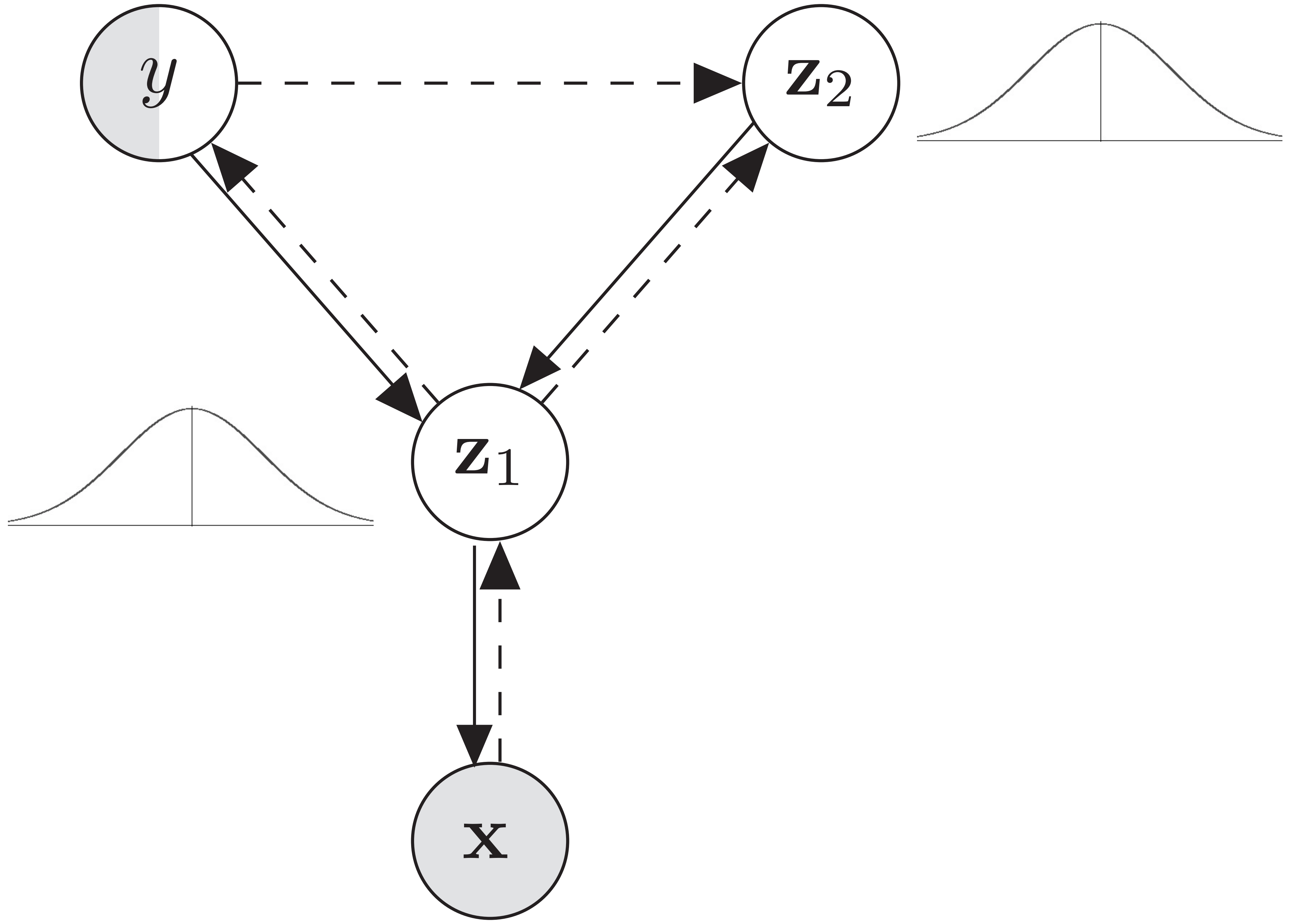


$$\boxed{q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\boldsymbol{\mu}_\phi(\mathbf{h}), \mathrm{diag}\left(\boldsymbol{\sigma}^2_\phi(\mathbf{h})\right)\right)}$$

Our SDGM



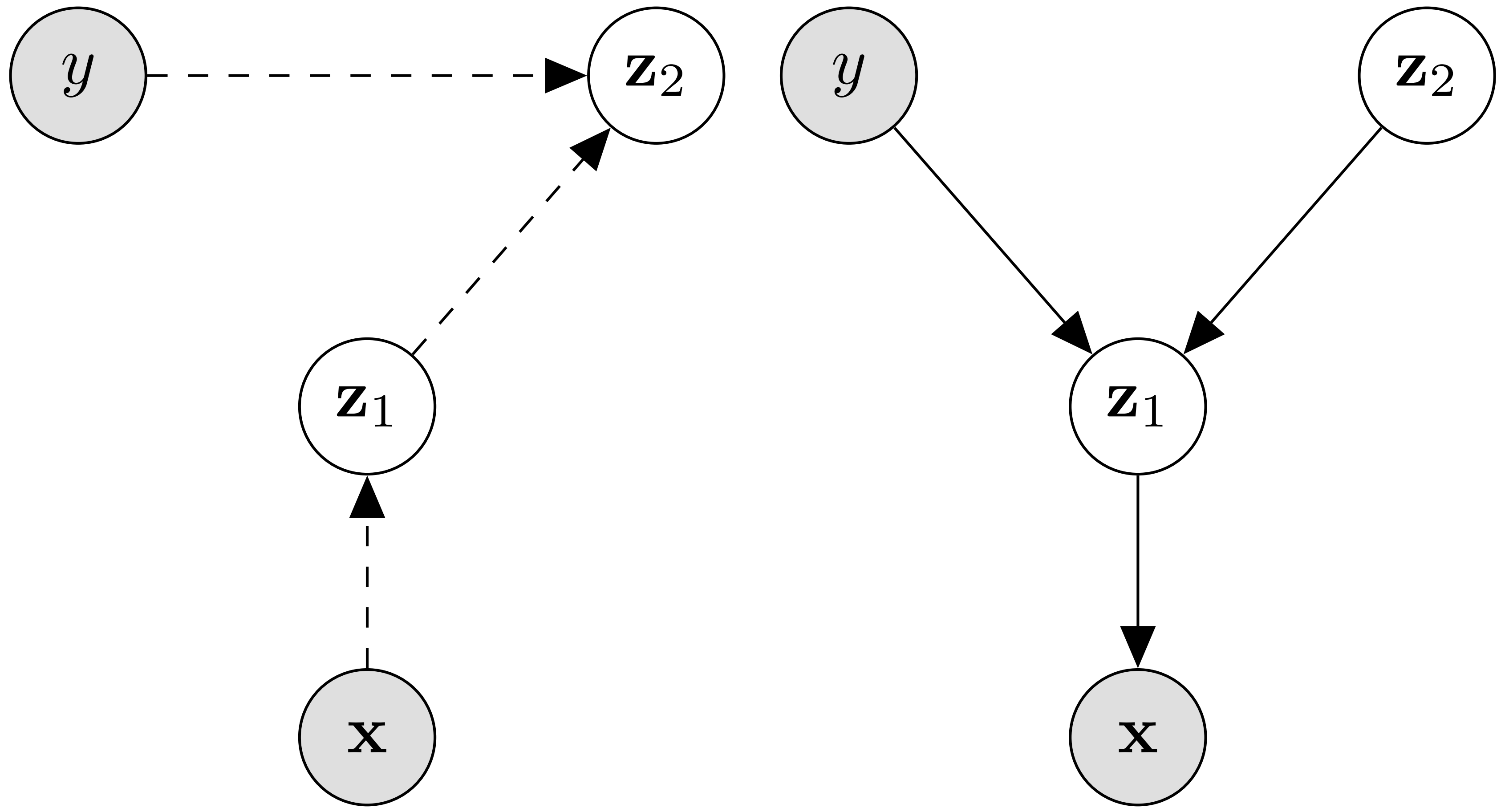
Our SDGM



$$\mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x},y)}\left[\log \frac{p_\theta(\mathbf{x}, y, \mathbf{z}_1, \mathbf{z}_2)}{q_\phi(\mathbf{z}_1, \mathbf{z}_2|\mathbf{x}, y)}\right] \leq \log p(\mathbf{x}, y)$$

$$\mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2,y|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, y, \mathbf{z}_1, \mathbf{z}_2)}{q_\phi(\mathbf{z}_1, \mathbf{z}_2, y|\mathbf{x})}\right] \leq \log p(\mathbf{x})$$

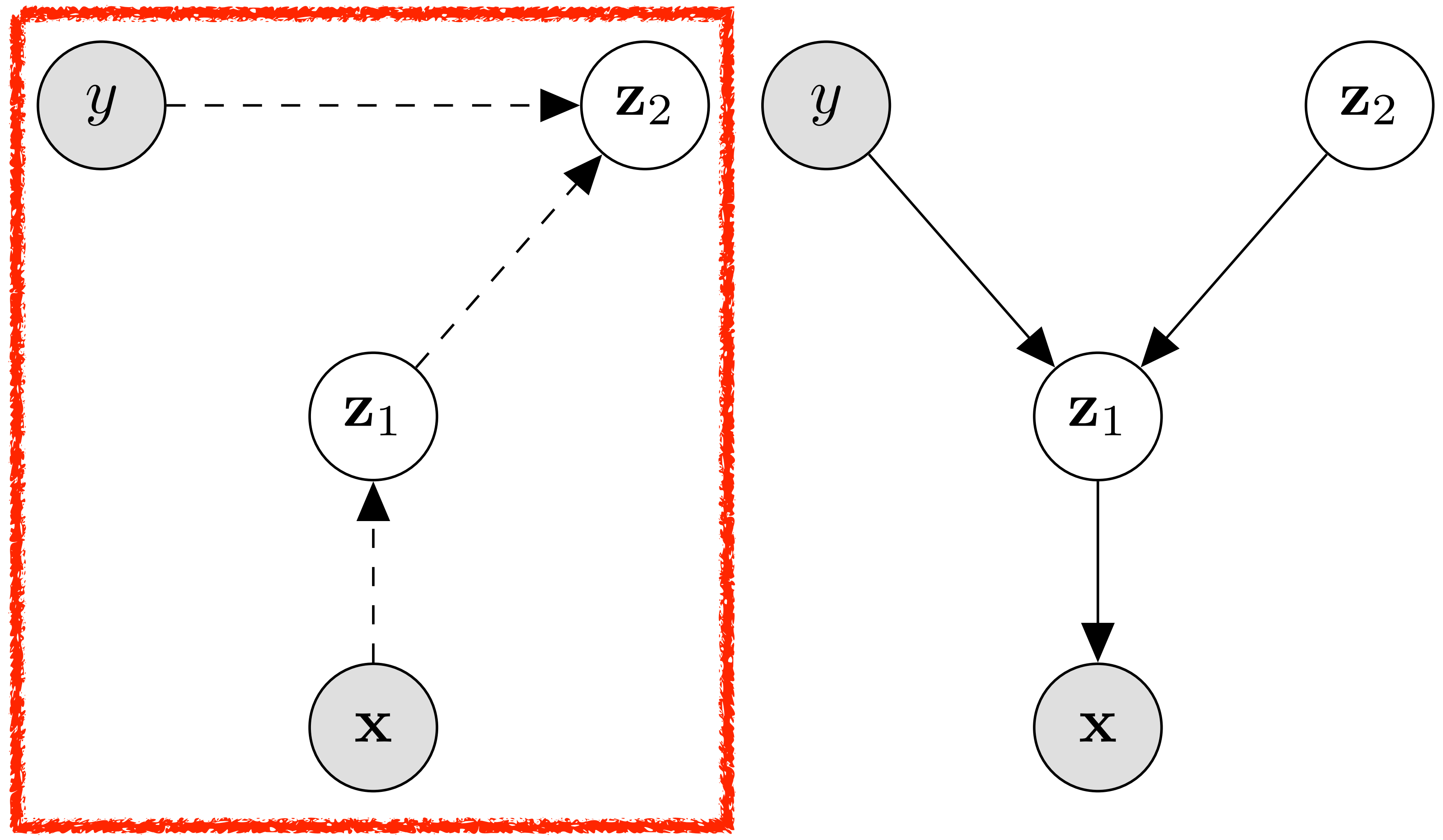
Our SDGM



$$\mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x},y)}\left[\log \frac{p_\theta(\mathbf{x},y,\mathbf{z}_1,\mathbf{z}_2)}{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x},y)}\right] \leq \log p(\mathbf{x},y)$$

$$\mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2,y|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},y,\mathbf{z}_1,\mathbf{z}_2)}{q_\phi(\mathbf{z}_1,\mathbf{z}_2,y|\mathbf{x})}\right] \leq \log p(\mathbf{x})$$

Our SDGM

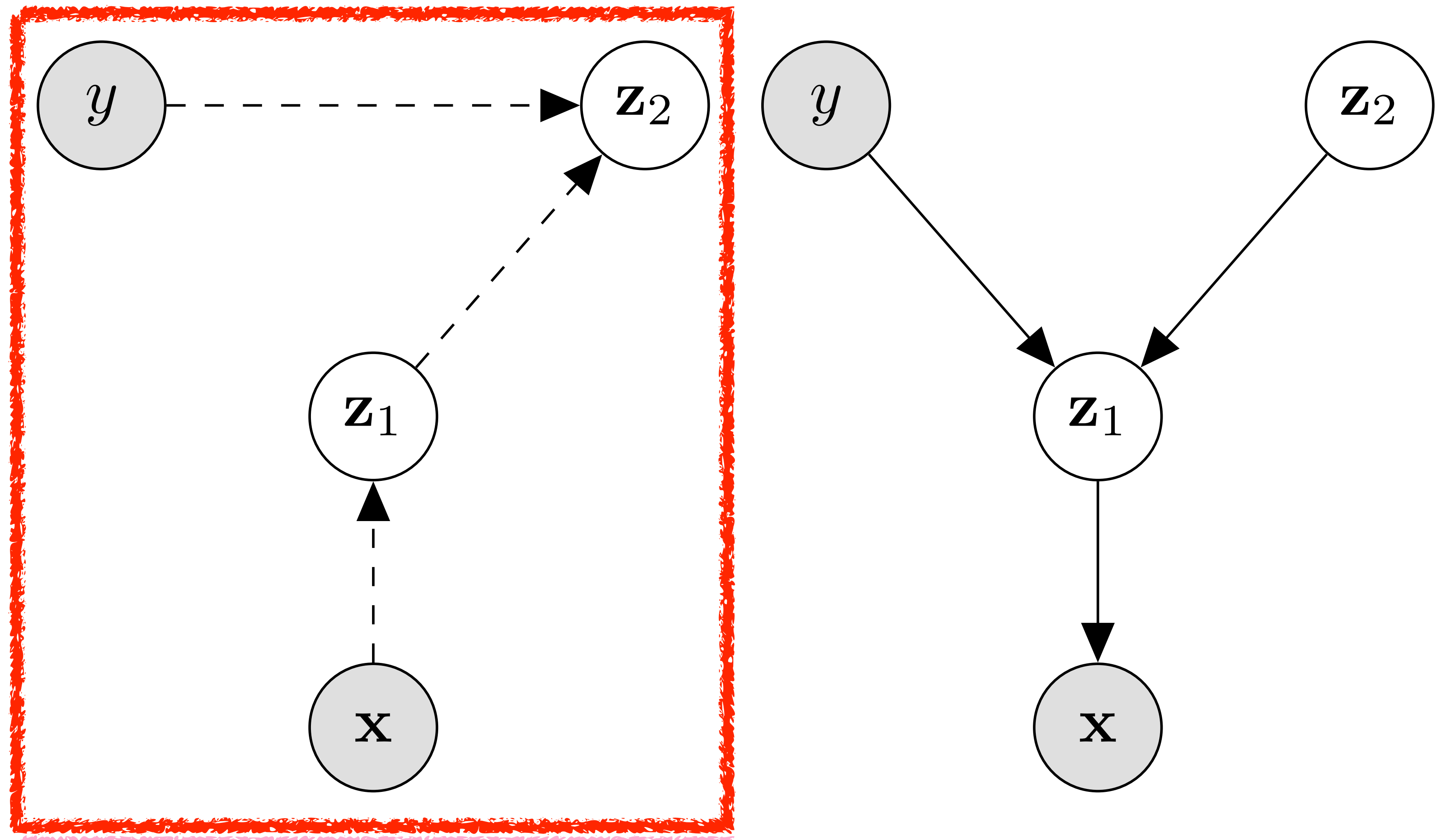


$$\mathbb{E}_{\boxed{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x},y)}}\left[\log \frac{p_\theta(\mathbf{x},y,\mathbf{z}_1,\mathbf{z}_2)}{\boxed{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x},y)}}\right] \leq \log p(\mathbf{x},y)$$

$$\mathcal{L}(\mathbf{x},y) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_1)]}_{\text{Reconstruction}} - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})q_\phi(\mathbf{z}_2|\mathbf{z}_1,y)}[\log \frac{q_\phi(\mathbf{z}_2|\mathbf{z}_1,y)}{p(\mathbf{z}_2)} + \log \frac{q_\phi(\mathbf{z}_1|\mathbf{x})}{p_\theta(\mathbf{z}_1|\mathbf{z}_2,y)}]}_{\text{KL}} + \underbrace{\log p(y)}_{\text{Constant}}$$

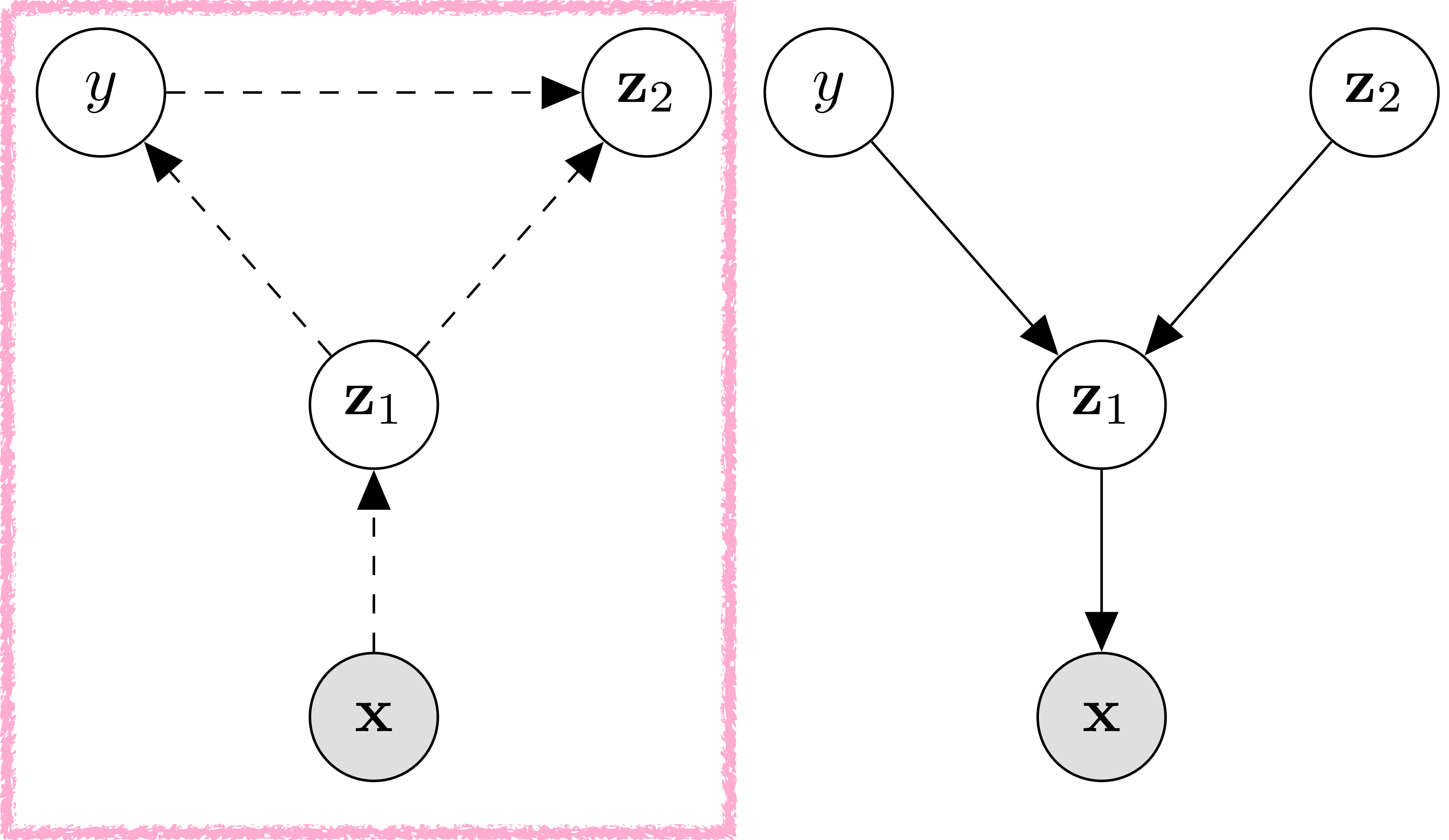
$$\mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2,y|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},y,\mathbf{z}_1,\mathbf{z}_2)}{q_\phi(\mathbf{z}_1,\mathbf{z}_2,y|\mathbf{x})}\right] \leq \log p(\mathbf{x})$$

Our SDGM



$$\mathbb{E}_{\boxed{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x},y)}}\left[\log \frac{p_\theta(\mathbf{x}, y, \mathbf{z}_1, \mathbf{z}_2)}{\boxed{q_\phi(\mathbf{z}_1, \mathbf{z}_2|\mathbf{x}, y)}}\right] \leq \log p(\mathbf{x}, y)$$

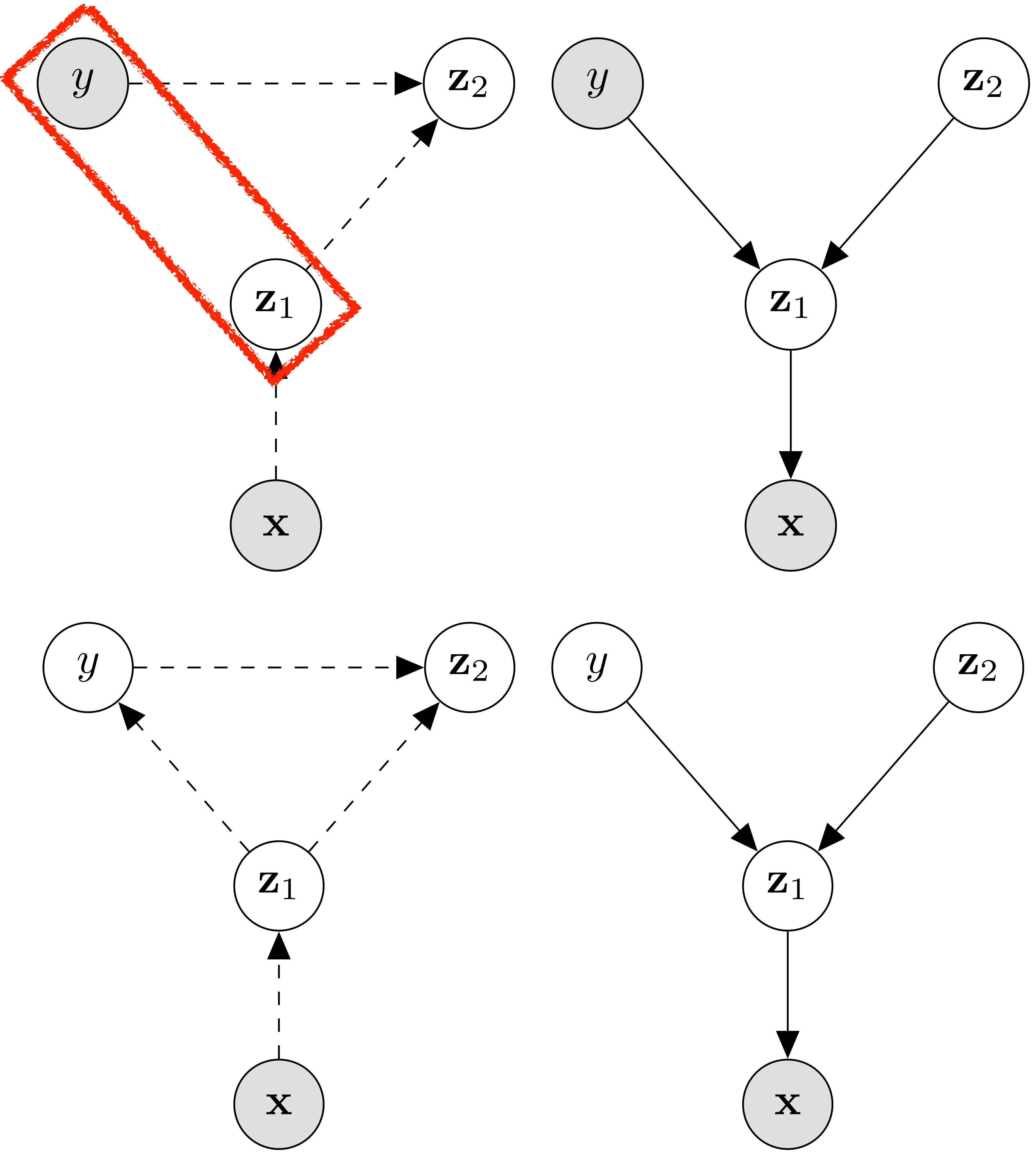
$$\mathcal{L}(\mathbf{x}, y) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_1)]}_{\text{Reconstruction}} - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})q_\phi(\mathbf{z}_2|\mathbf{z}_1,y)}[\log \frac{q_\phi(\mathbf{z}_2|\mathbf{z}_1, y)}{p(\mathbf{z}_2)} + \log \frac{q_\phi(\mathbf{z}_1|\mathbf{x})}{p_\theta(\mathbf{z}_1|\mathbf{z}_2, y)}]}_{\text{KL}} + \underbrace{\log p(y)}_{\text{Constant}}$$



$$\mathbb{E}_{\boxed{q_\phi(\mathbf{z}_1,\mathbf{z}_2,y|\mathbf{x})}}\left[\log \frac{p_\theta(\mathbf{x}, y, \mathbf{z}_1, \mathbf{z}_2)}{\boxed{q_\phi(\mathbf{z}_1, \mathbf{z}_2, y|\mathbf{x})}}\right] \leq \log p(\mathbf{x})$$

$$\mathcal{U}(\mathbf{x}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_1)]}_{\text{Reconstruction}} - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})q_\phi(y|\mathbf{z}_1)q_\phi(\mathbf{z}_2|\mathbf{z}_1,y)}[\log \frac{q_\phi(\mathbf{z}_2|\mathbf{z}_1, y)}{p(\mathbf{z}_2)} + \log \frac{q_\phi(\mathbf{z}_1|\mathbf{x})}{p_\theta(\mathbf{z}_1|\mathbf{z}_2, y)} + \log \frac{q_\phi(y|\mathbf{z}_1)}{p(y)}]}_{\text{KL}}$$

Our SDGM



$$\mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x},y)}\left[\log \frac{p_\theta(\mathbf{x},y,\mathbf{z}_1,\mathbf{z}_2)}{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x},y)}\right] \leq \log p(\mathbf{x},y)$$

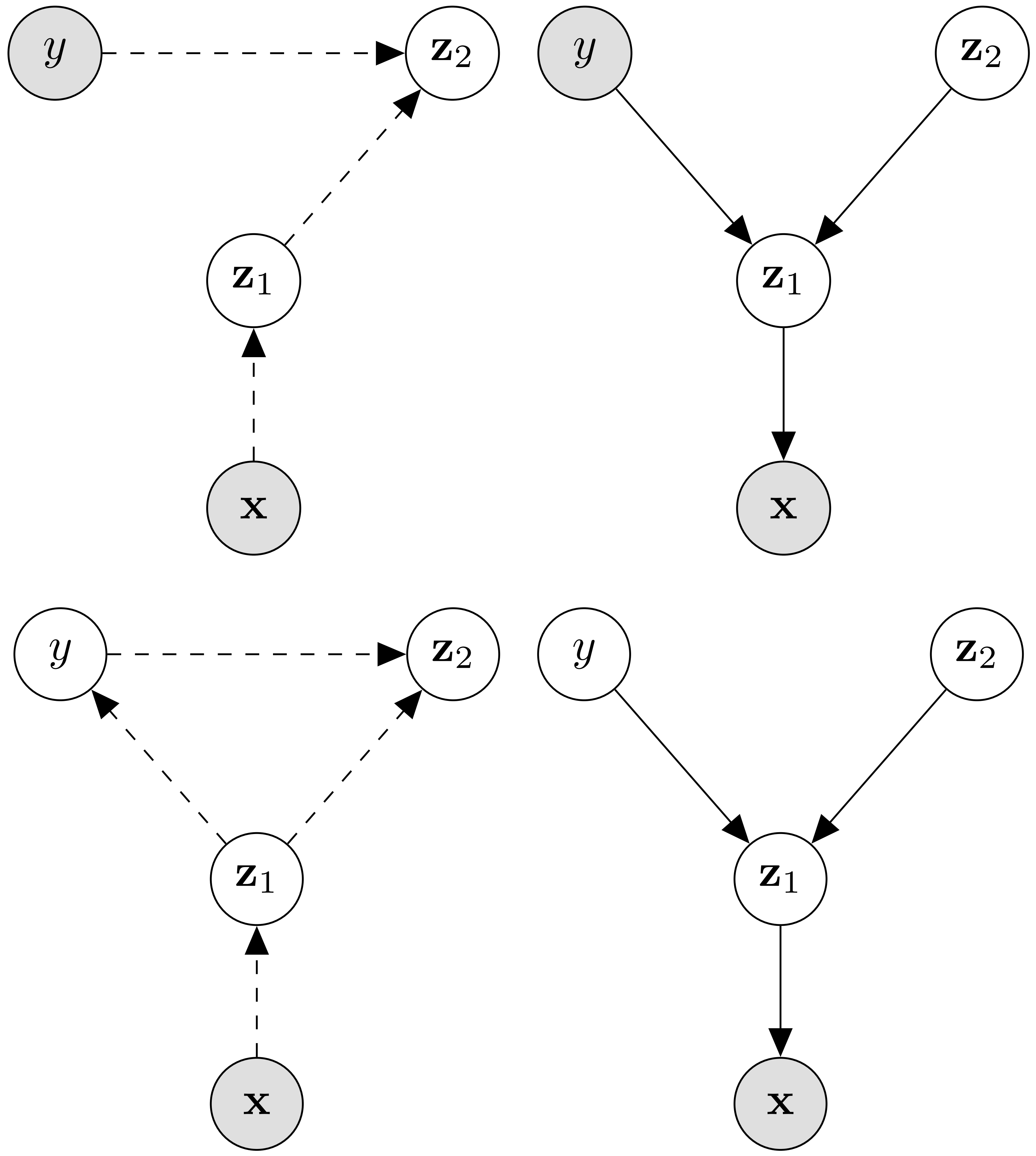
$$\mathcal{L}(\mathbf{x},y) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_1)]}_{\text{Reconstruction}} - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})q_\phi(\mathbf{z}_2|\mathbf{z}_1,y)}[\log \frac{q_\phi(\mathbf{z}_2|\mathbf{z}_1,y)}{p(\mathbf{z}_2)} + \log \frac{q_\phi(\mathbf{z}_1|\mathbf{x})}{p_\theta(\mathbf{z}_1|\mathbf{z}_2,y)}]}_{\text{KL}} + \underbrace{\log p(y)}_{\text{Constant}}$$

$$\mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2,y|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},y,\mathbf{z}_1,\mathbf{z}_2)}{q_\phi(\mathbf{z}_1,\mathbf{z}_2,y|\mathbf{x})}\right] \leq \log p(\mathbf{x})$$

$$\mathcal{U}(\mathbf{x}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_1)]}_{\text{Reconstruction}} - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})q_\phi(y|\mathbf{z}_1)q_\phi(\mathbf{z}_2|\mathbf{z}_1,y)}[\log \frac{q_\phi(\mathbf{z}_2|\mathbf{z}_1,y)}{p(\mathbf{z}_2)} + \log \frac{q_\phi(\mathbf{z}_1|\mathbf{x})}{p_\theta(\mathbf{z}_1|\mathbf{z}_2,y)} + \log \frac{q_\phi(y|\mathbf{z}_1)}{p(y)}]}_{\text{KL}}$$

$$\mathcal{J}_{cls}(\mathbf{x},y) = \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})}[q_\phi(y|\mathbf{z}_1)]$$

Our SDGM



$$\mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x},y)}\left[\log \frac{p_\theta(\mathbf{x},y,\mathbf{z}_1,\mathbf{z}_2)}{q_\phi(\mathbf{z}_1,\mathbf{z}_2|\mathbf{x},y)}\right] \leq \log p(\mathbf{x},y)$$

$$\mathcal{L}(\mathbf{x},y) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_1)]}_{\text{Reconstruction}} - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})q_\phi(\mathbf{z}_2|\mathbf{z}_1,y)}[\log \frac{q_\phi(\mathbf{z}_2|\mathbf{z}_1,y)}{p(\mathbf{z}_2)} + \log \frac{q_\phi(\mathbf{z}_1|\mathbf{x})}{p_\theta(\mathbf{z}_1|\mathbf{z}_2,y)}]}_{\text{KL}} + \underbrace{\log p(y)}_{\text{Constant}}$$

$$\mathbb{E}_{q_\phi(\mathbf{z}_1,\mathbf{z}_2,y|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},y,\mathbf{z}_1,\mathbf{z}_2)}{q_\phi(\mathbf{z}_1,\mathbf{z}_2,y|\mathbf{x})}\right] \leq \log p(\mathbf{x})$$

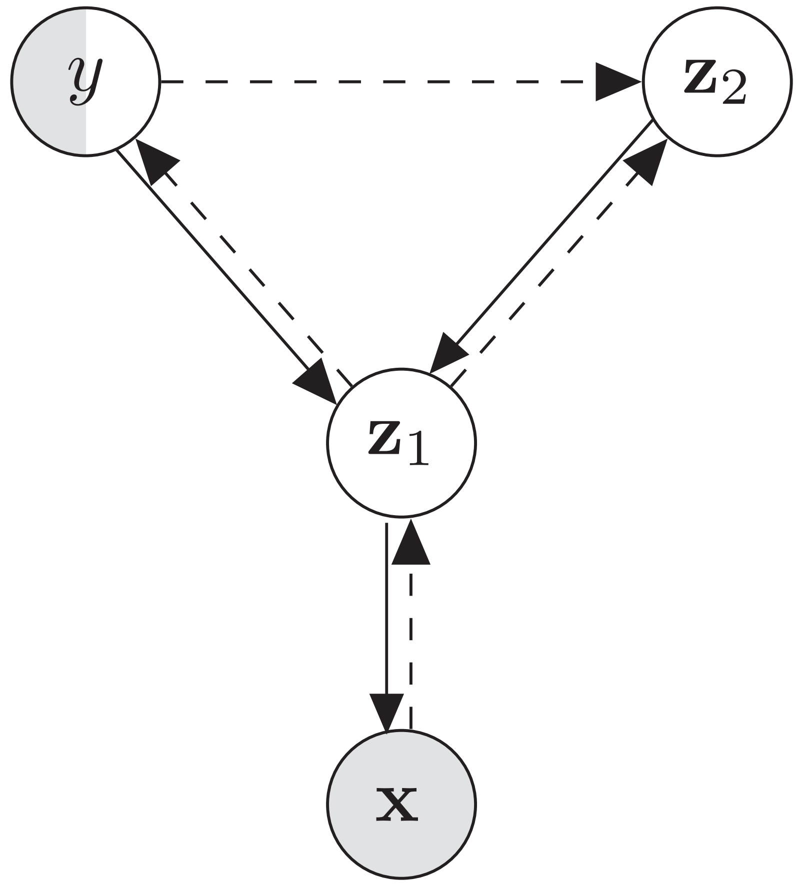
$$\mathcal{U}(\mathbf{x}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}_1)]}_{\text{Reconstruction}} - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})q_\phi(y|\mathbf{z}_1)q_\phi(\mathbf{z}_2|\mathbf{z}_1,y)}[\log \frac{q_\phi(\mathbf{z}_2|\mathbf{z}_1,y)}{p(\mathbf{z}_2)} + \log \frac{q_\phi(\mathbf{z}_1|\mathbf{x})}{p_\theta(\mathbf{z}_1|\mathbf{z}_2,y)} + \log \frac{q_\phi(y|\mathbf{z}_1)}{p(y)}]}_{\text{KL}}$$

$$\mathcal{J}_{cls}(\mathbf{x},y) = \mathbb{E}_{q_\phi(\mathbf{z}_1|\mathbf{x})}[q_\phi(y|\mathbf{z}_1)]$$

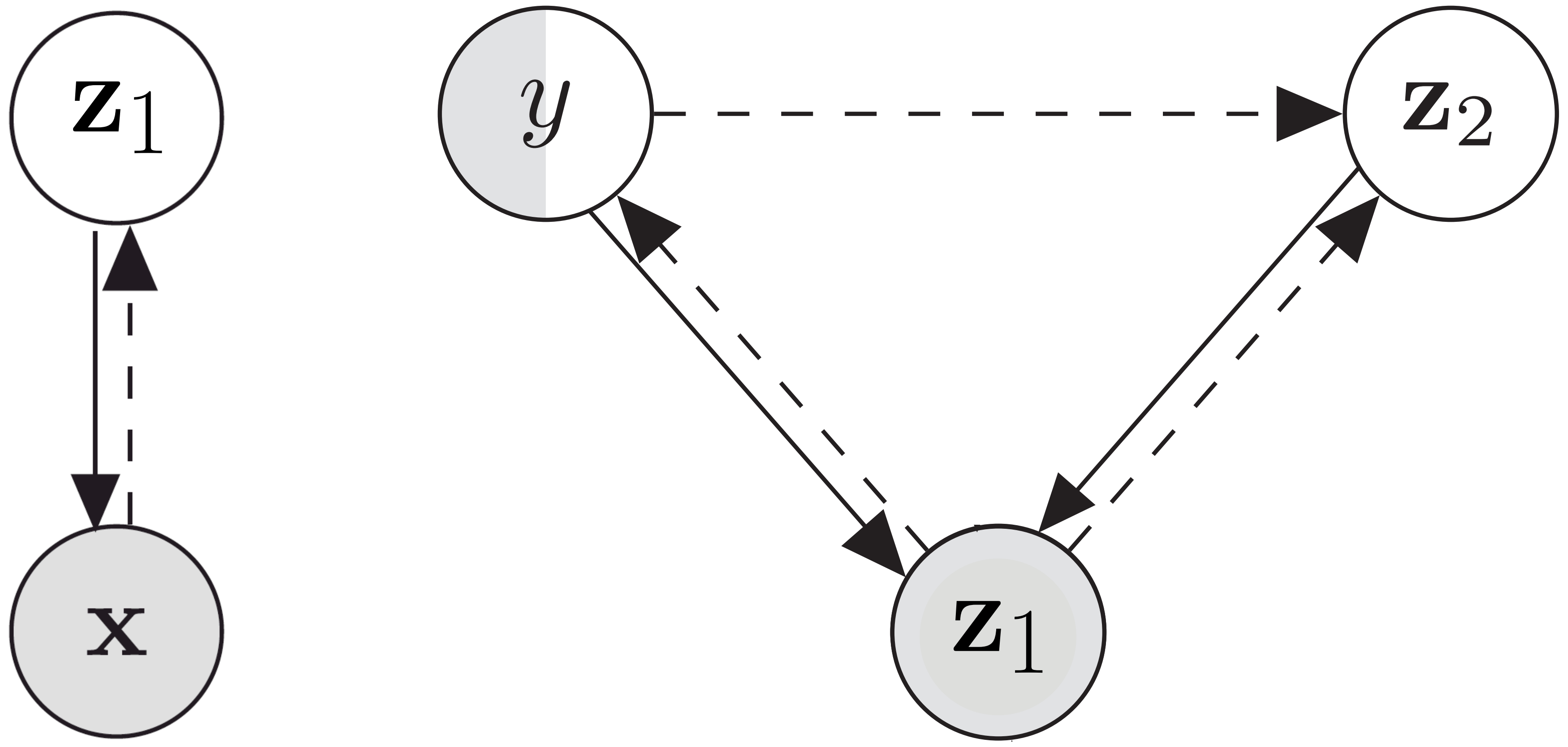
$$\boxed{\mathcal{J} = \sum_{(\mathbf{x},y)\in\mathcal{S}_l} \big(\mathcal{L}(\mathbf{x},y) + \alpha\mathcal{J}_{cls}(\mathbf{x},y)\big) + \sum_{\mathbf{x}\in\mathcal{S}_u} \mathcal{U}(\mathbf{x})}$$

Difference between our SDGM and M1 + M2 [Kingma et al. 2014]

Our SDGM

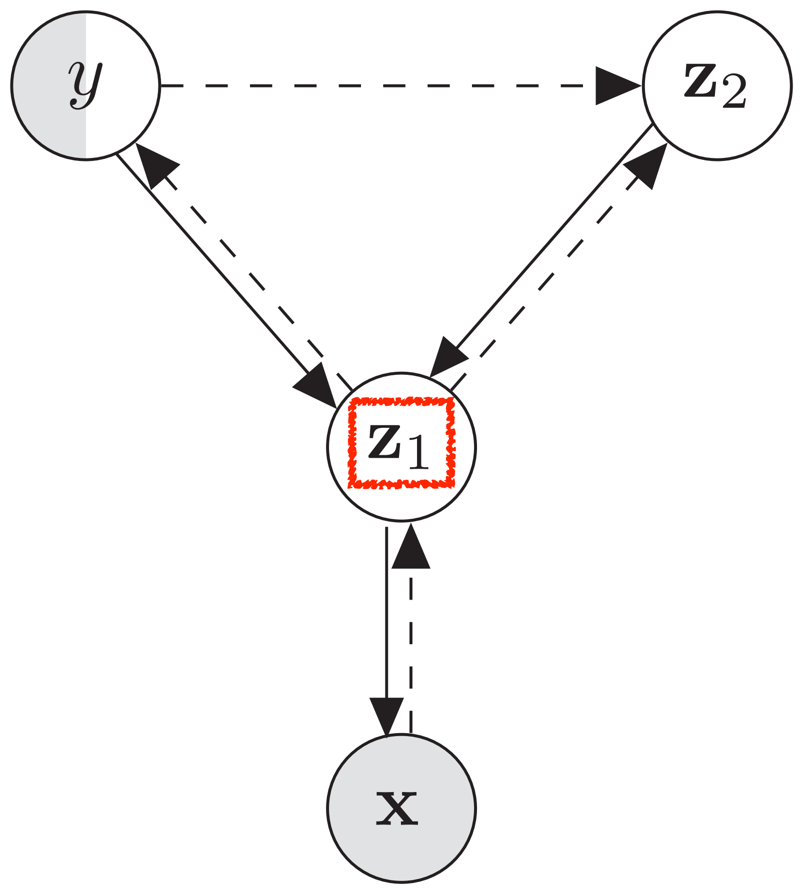


M1+M2



Difference between our SDGM and M1 + M2 [Kingma et al. 2014]

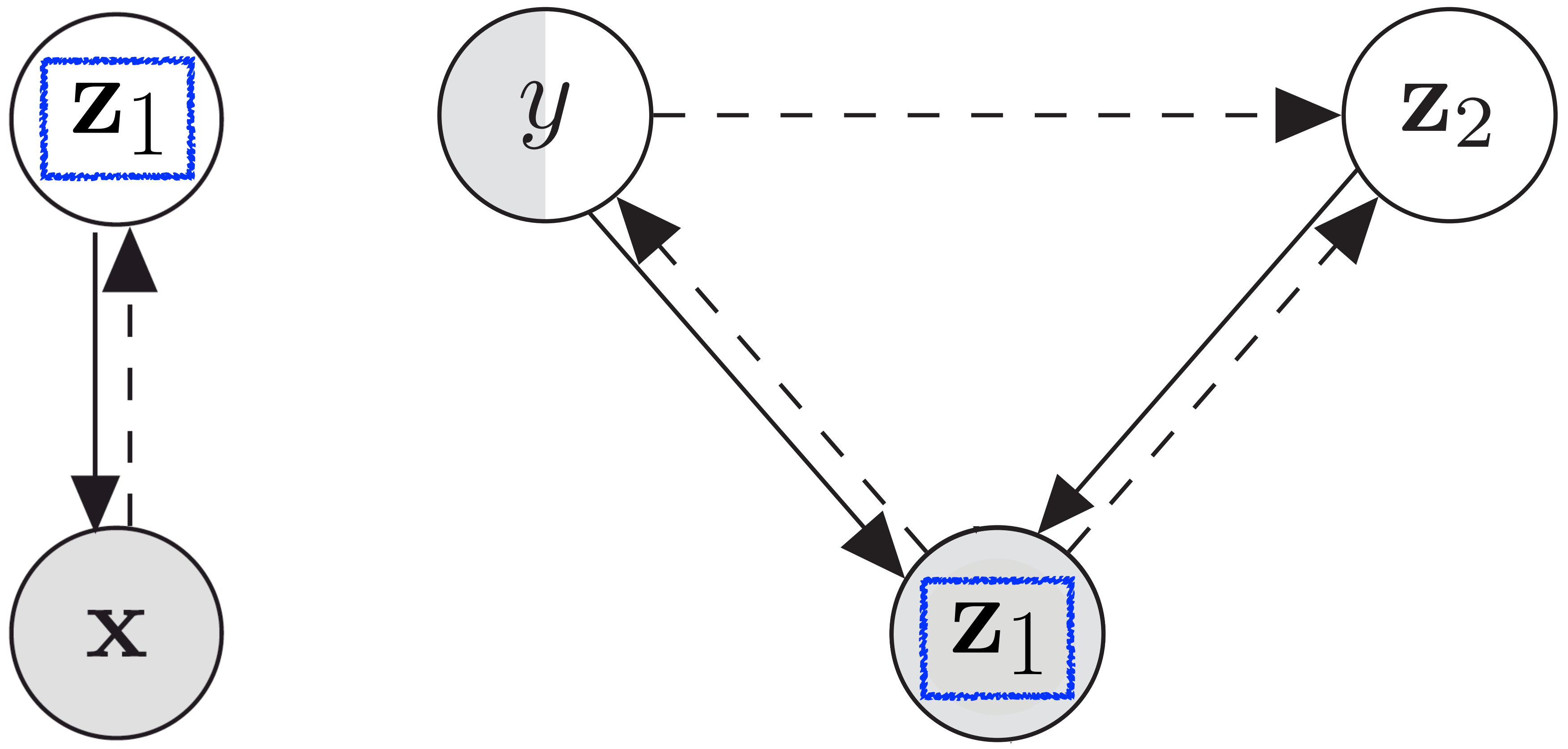
Our SDGM



$$\mathbb{E}_{q_\phi(\boxed{\mathbf{z}_1},\mathbf{z}_2|\mathbf{x},y)}\left[\log\frac{p_\theta(\mathbf{x},y,\boxed{\mathbf{z}_1},\mathbf{z}_2)}{q_\phi(\boxed{\mathbf{z}_1},\mathbf{z}_2|\mathbf{x},y)}\right]\le\log p(\mathbf{x},y)$$

$$\mathbb{E}_{q_\phi(\boxed{\mathbf{z}_1},\mathbf{z}_2,y|\mathbf{x})}\left[\log\frac{p_\theta(\mathbf{x},y,\boxed{\mathbf{z}_1},\mathbf{z}_2)}{q_\phi(\boxed{\mathbf{z}_1},\mathbf{z}_2,y|\mathbf{x})}\right]\le\log p(\mathbf{x})$$

M1+M2

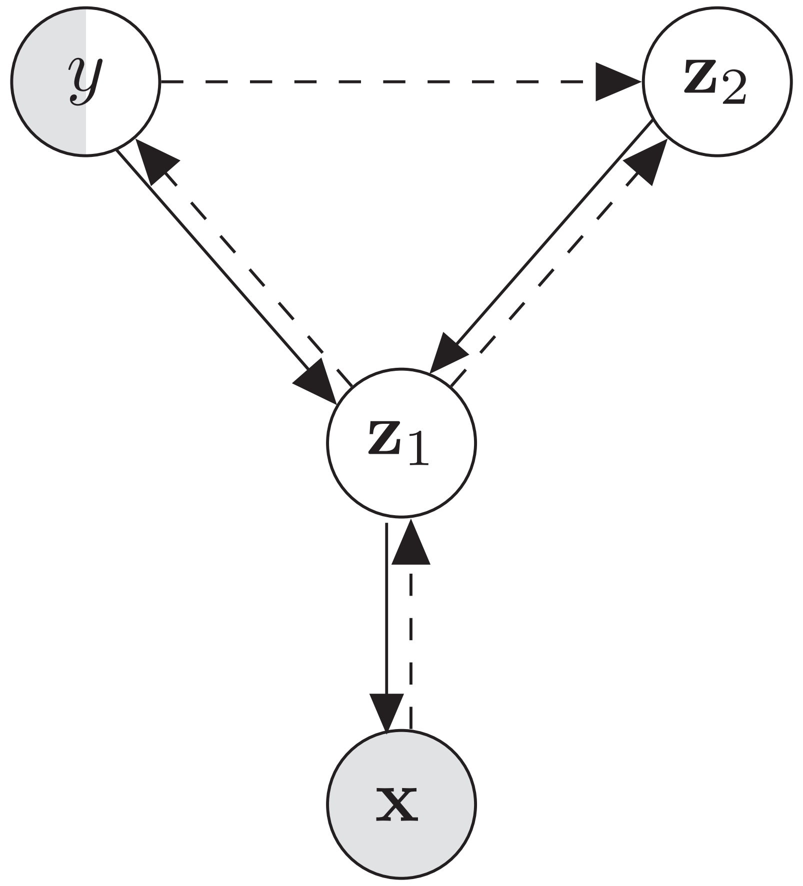


$$\mathbb{E}_{q_\phi(\mathbf{z}_2|\boxed{\mathbf{z}_1},y)}\left[\log\frac{p_\theta(\boxed{\mathbf{z}_1},y,\mathbf{z}_2)}{q_\phi(\mathbf{z}_2|\boxed{\mathbf{z}_1},y)}\right]\le\log p(\boxed{\mathbf{z}_1},y)$$

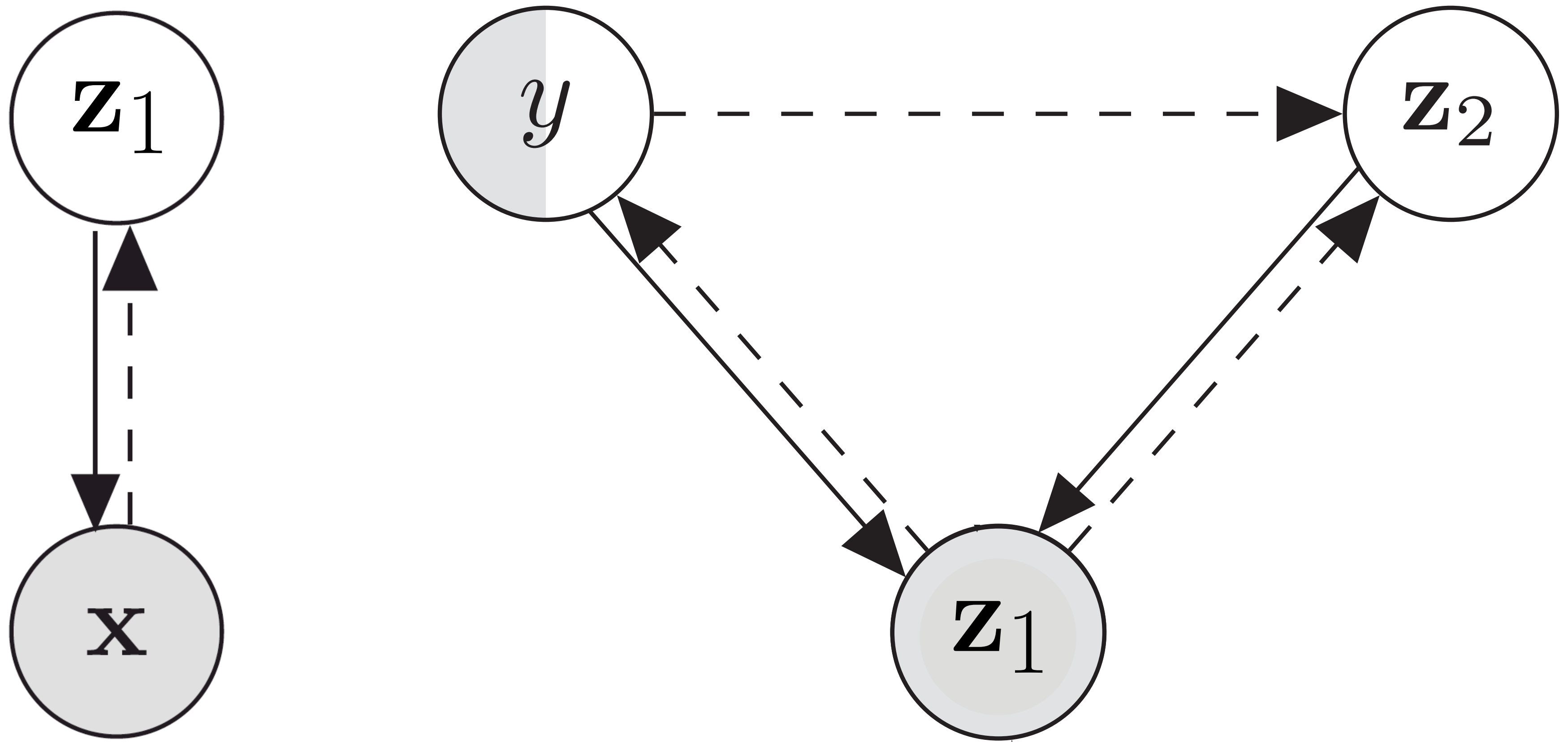
$$\mathbb{E}_{q_\phi(\mathbf{z}_2,y|\boxed{\mathbf{z}_1})}\left[\log\frac{p_\theta(\boxed{\mathbf{z}_1},y,\mathbf{z}_2)}{q_\phi(\mathbf{z}_2,y|\boxed{\mathbf{z}_1})}\right]\le\log p(\boxed{\mathbf{z}_1})$$

Difference between our SDGM and M1 + M2 [Kingma et al. 2014]

Our SDGM

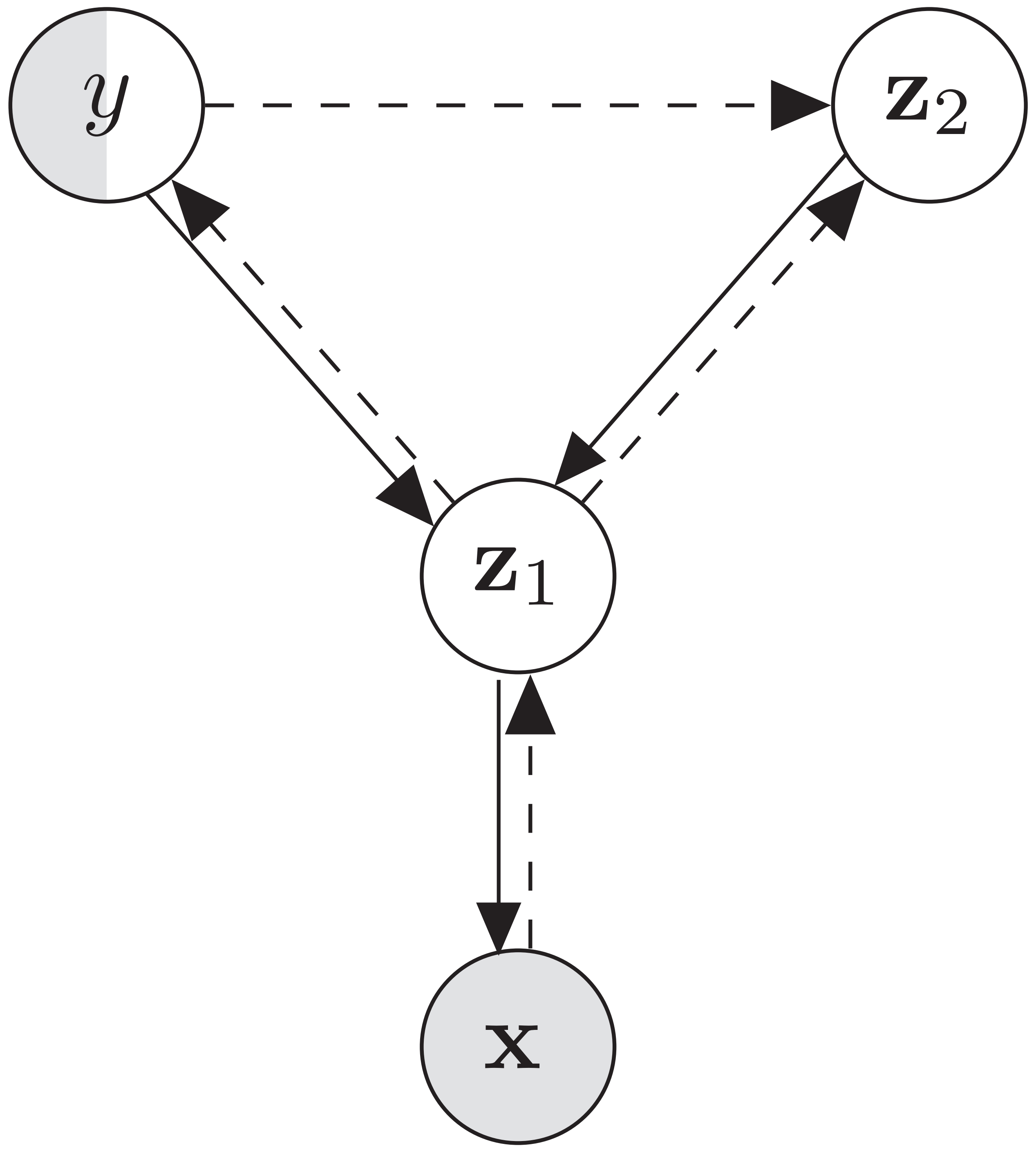


M1+M2

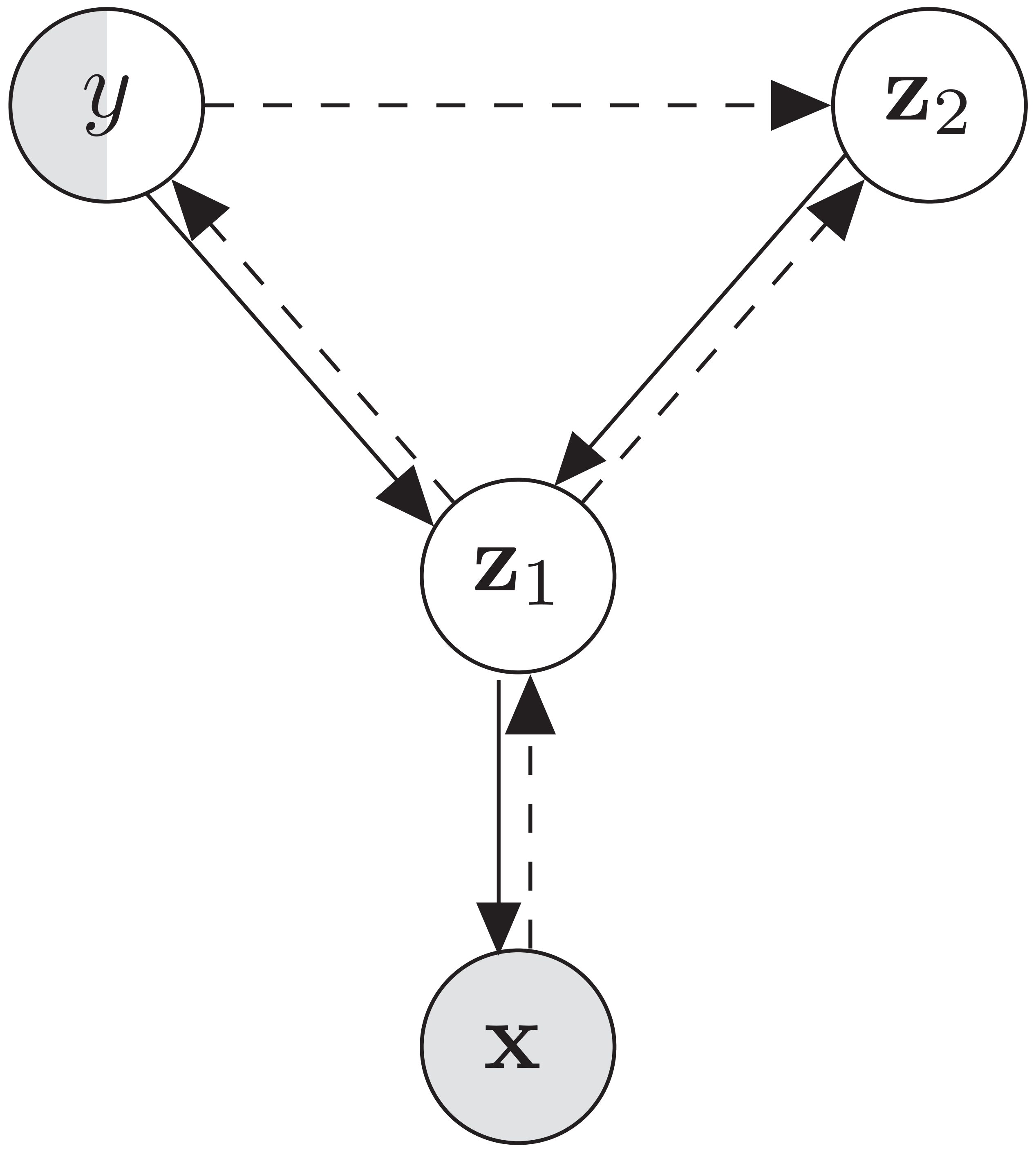


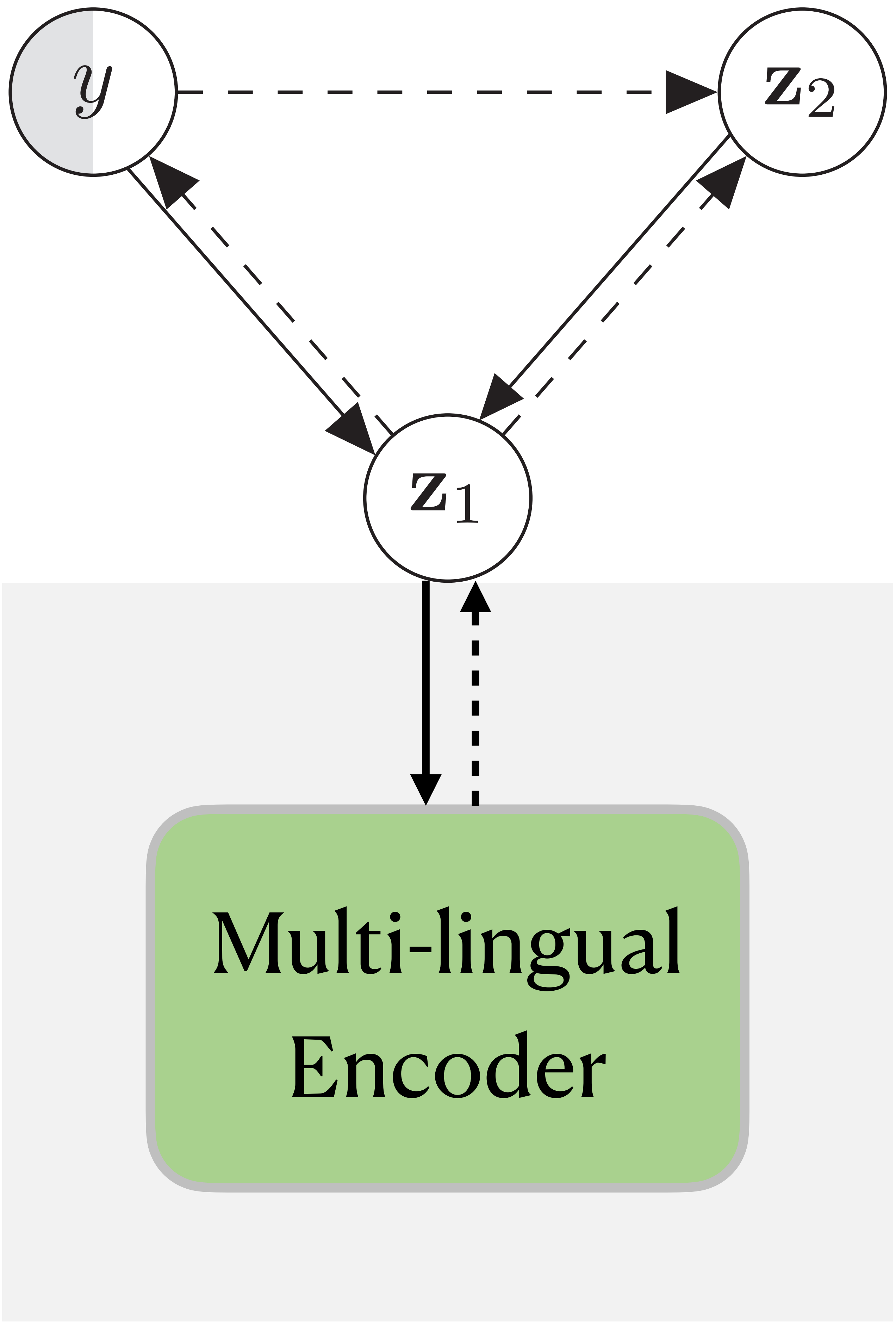
- M1 + M2
  - Layer-wise training
  - The parameters between $\mathbf{x}$ and $\mathbf{z}_1$ are fixed during semi-supervised learning
- Our SDGM
  - End-to-end training
  - Mathematical reformulation of ELBOs into reconstruction and KL terms, more stable optimisation schedule

SDGMs + Multi-lingual pretraining

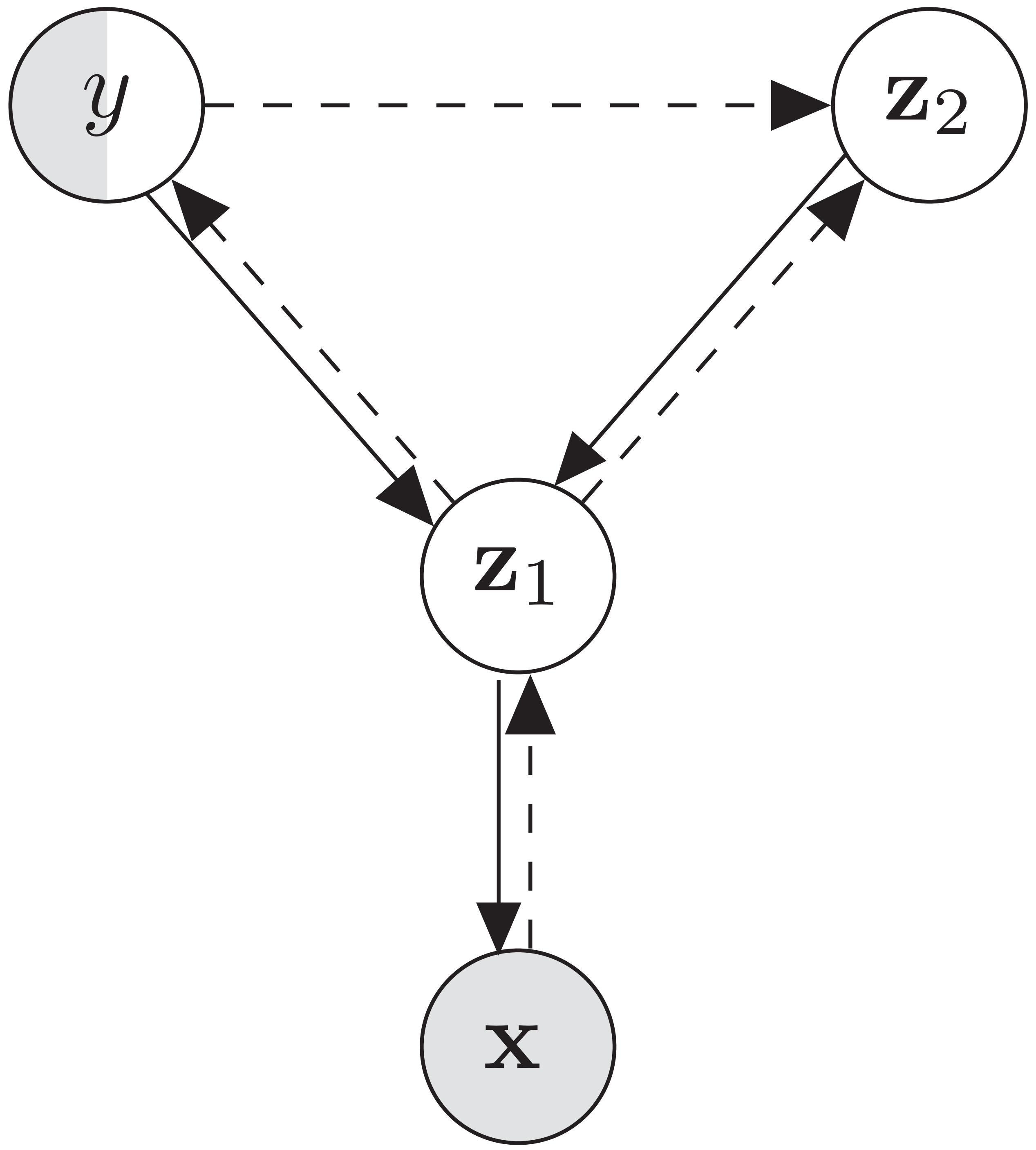


SDGMs + Multi-lingual pretraining

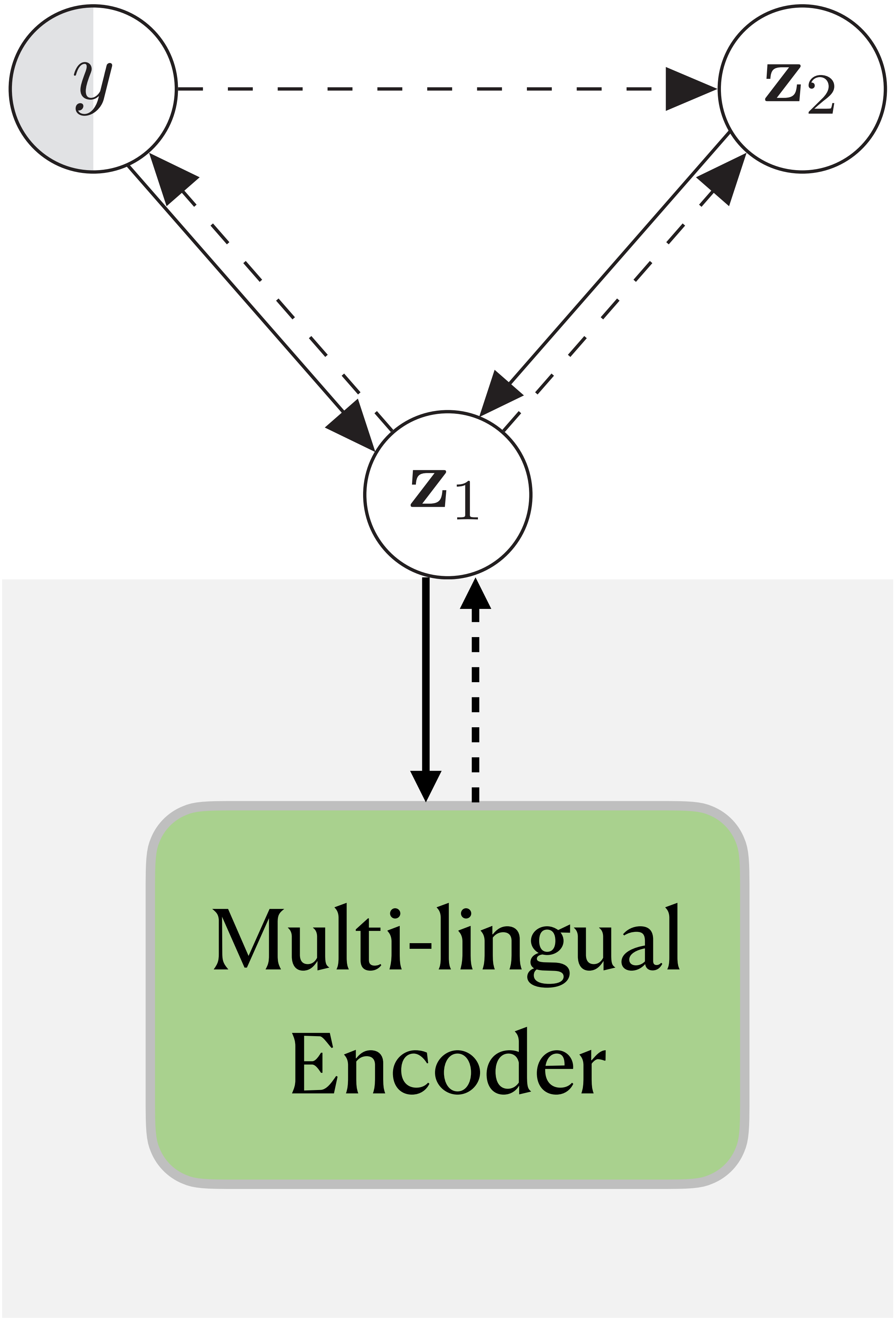




SDGMs + Multi-lingual pretraining



SDGMs can operate on top of the multi-lingual encoder *independently* of encoding architecture



Experiments

Document classification in multiple languages

- Multilingual document classification corpus (MLDoc; Schwenk and Li [2018])

Experiments

Document classification in multiple languages

- Multilingual document classification corpus (MLDoc; Schwenk and Li [2018])
- Five languages in total: English (EN), German (DE), French (FR), Russian (RU), Chinese (ZH)

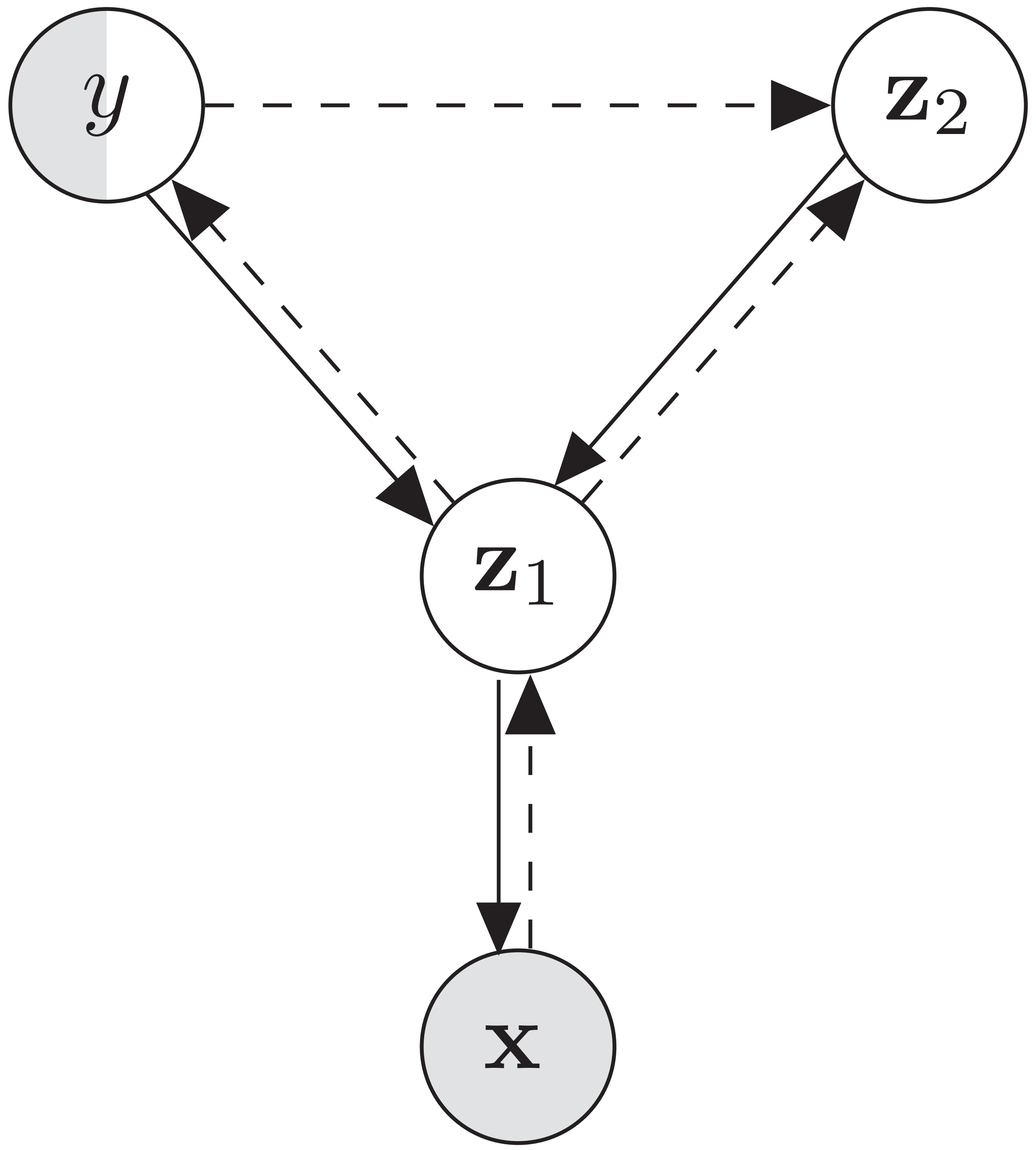
Experiments

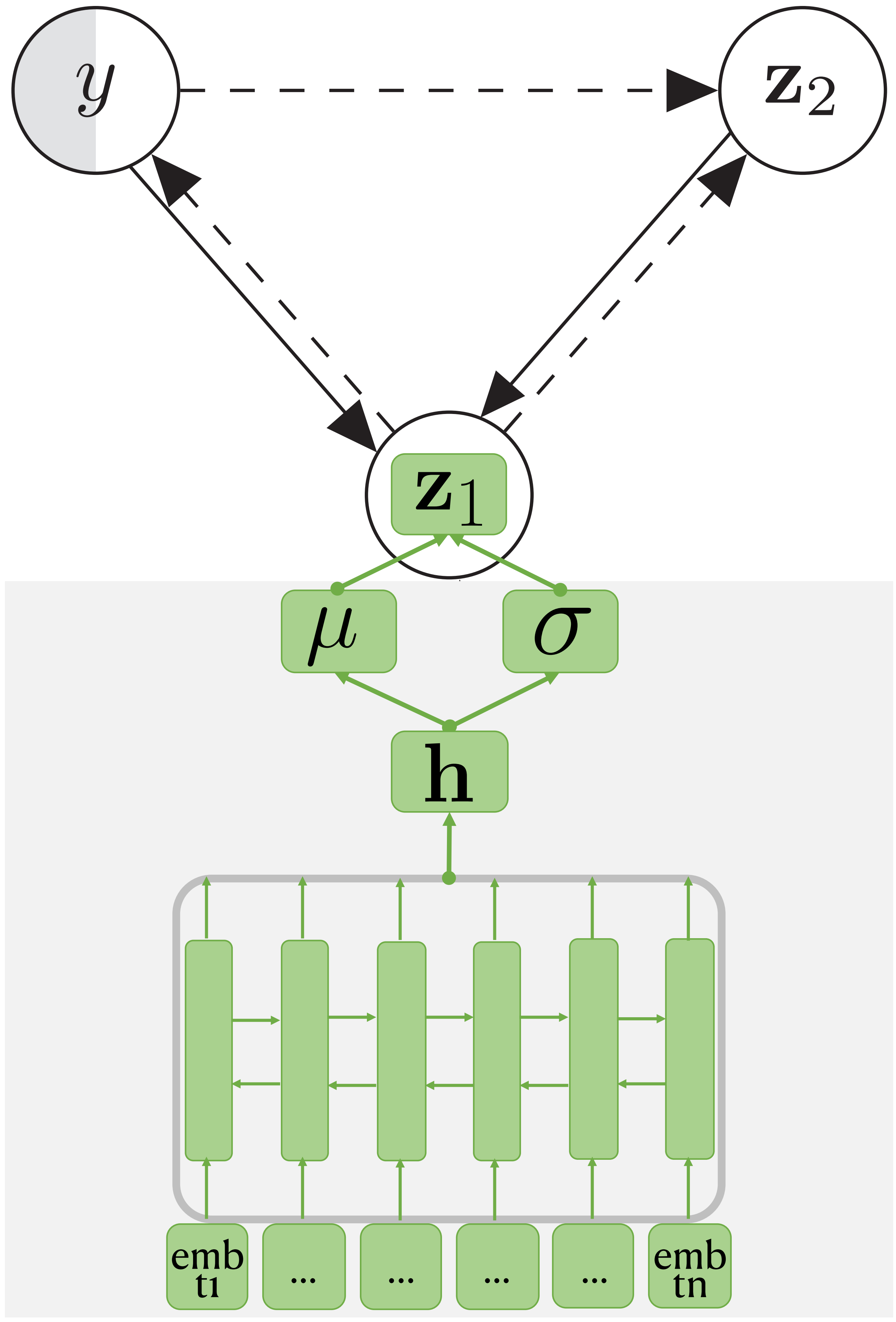
Document classification in multiple languages

- Multilingual document classification corpus (MLDoc; Schwenk and Li [2018])
- Five languages in total: English (EN), German (DE), French (FR), Russian (RU), Chinese (ZH)
- For 1k training corpus, vary labelled data size, use the rest as unlabelled data

SDGMs + Multi-lingual pretraining

Non-parallel cross-lingual VAE (NXVAE)





Qualitative results

EN-DE

Word pair	Lang	kNNs ($k = 3$)
president (EN)	EN	mr, madam, gentlemen
	DE	präsident, herr, kommissar
präsident (DE)	EN	president, mr, madam
	DE	herr, kommissar, herren
great (EN)	EN	deal, with, a
	DE	große, eine, gute
groß (DE)	EN	striking, gets, lucrative
	DE	gering, heikel, hoch
said (EN)	EN	already, as, been
	DE	gesagt, mit, dem
sagte (DE)	EN	he, rightly, said
	DE	vorhin, kollege, kommissar

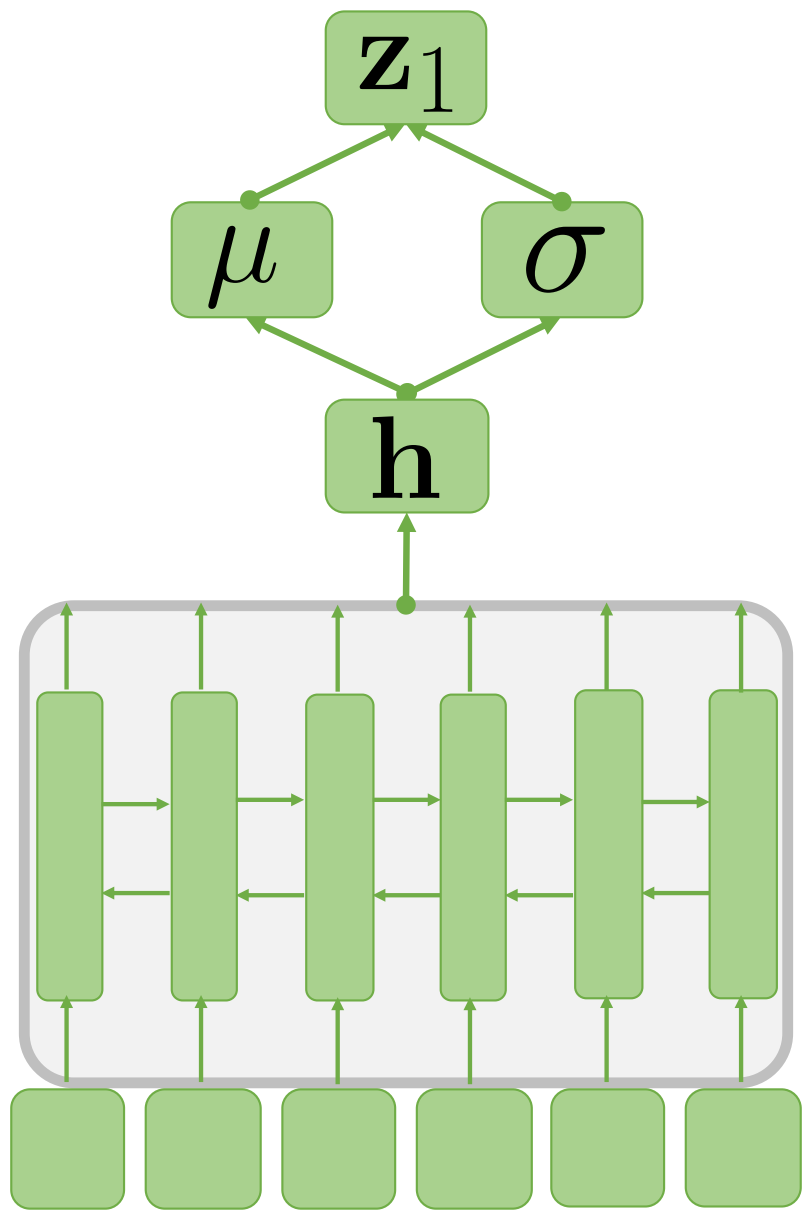
Our SDGM - NXVAE

- Our SDGM with bag-of-word (BOW) decoder (**M1** + **M2** + **BOW**)
- Our SDGM with GRU decoder (**M1** + **M2** + **GRU**)

Baselines - NXVAE

Supervised models

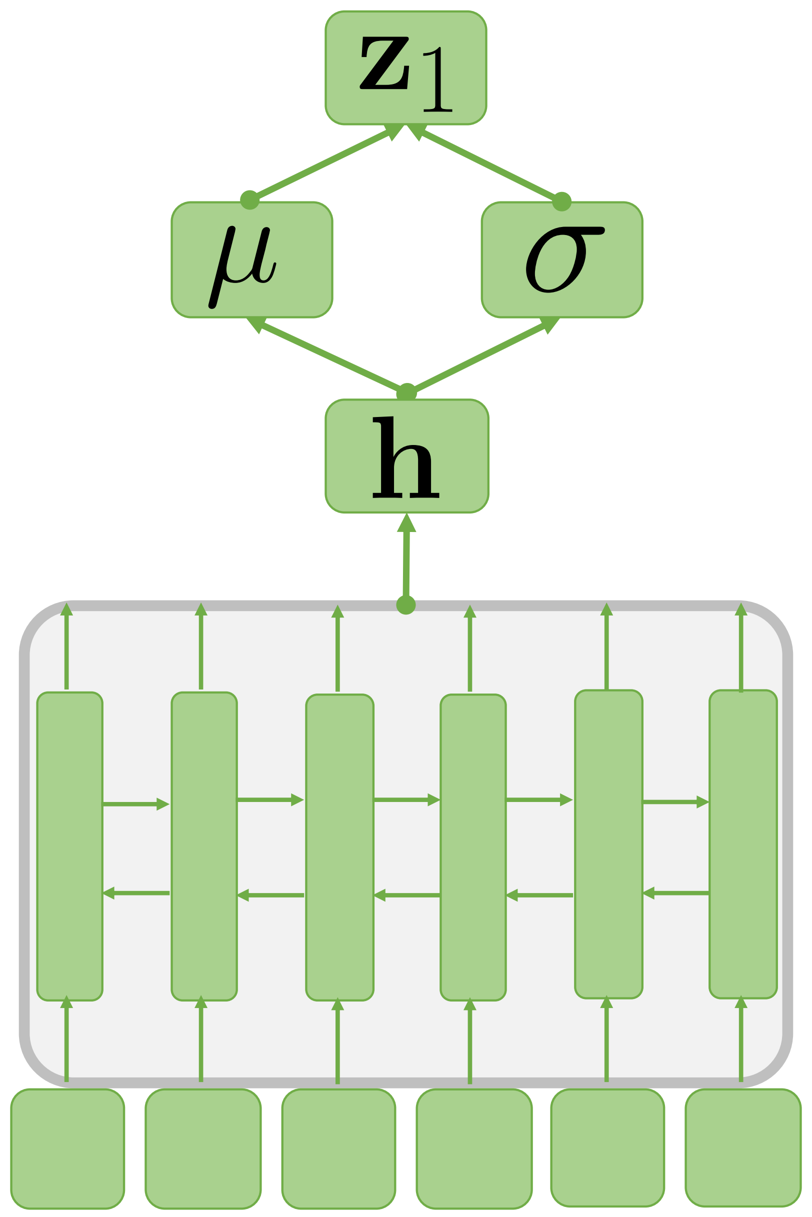
NXVAE-Z1



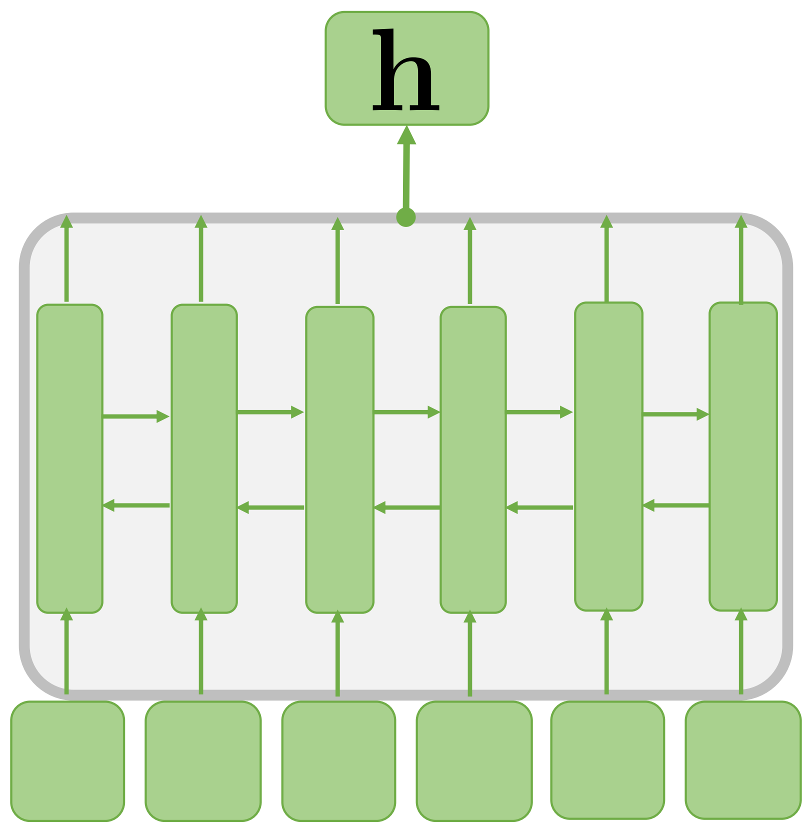
Baselines - NXVAE

Supervised models

NXVAE-Z1



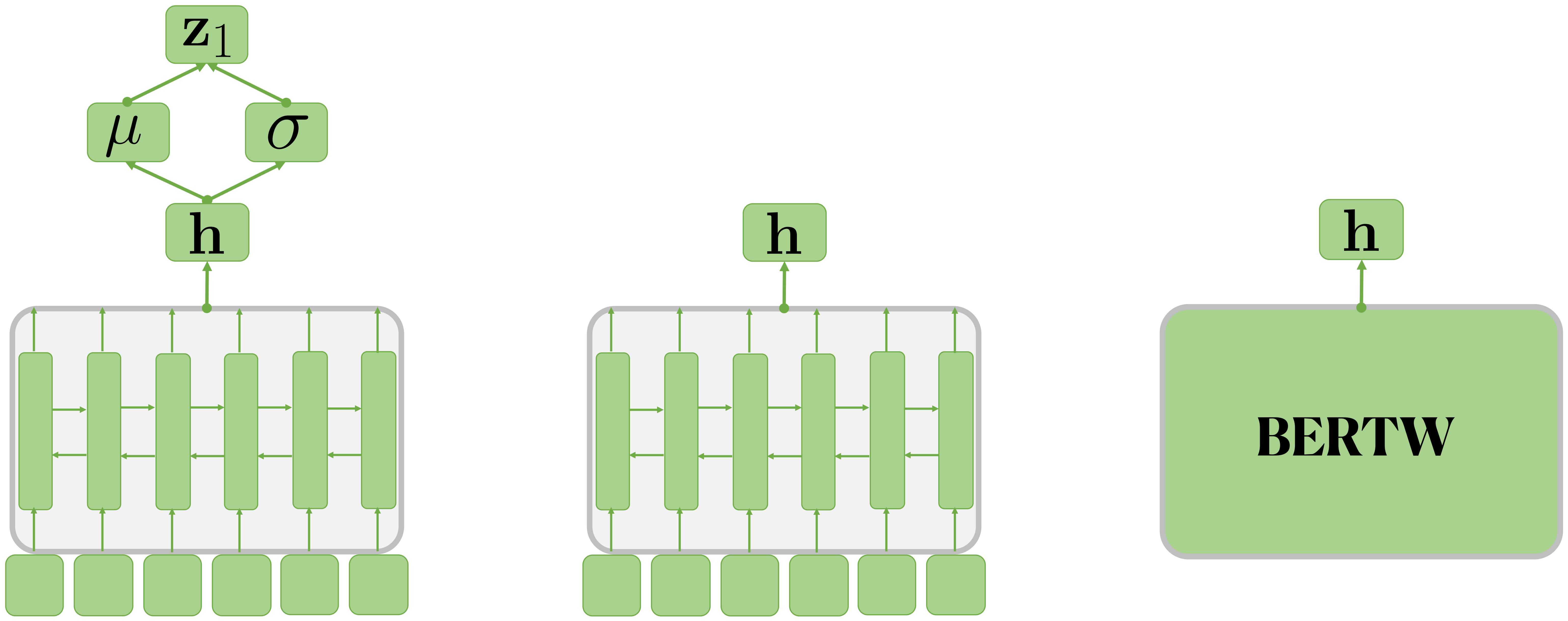
NXVAE-h



Baselines - NXVAE

Supervised models

NXVAE-Z1 NXVAE-h BERTW



Baselines - NXVAE

Semi-supervised models

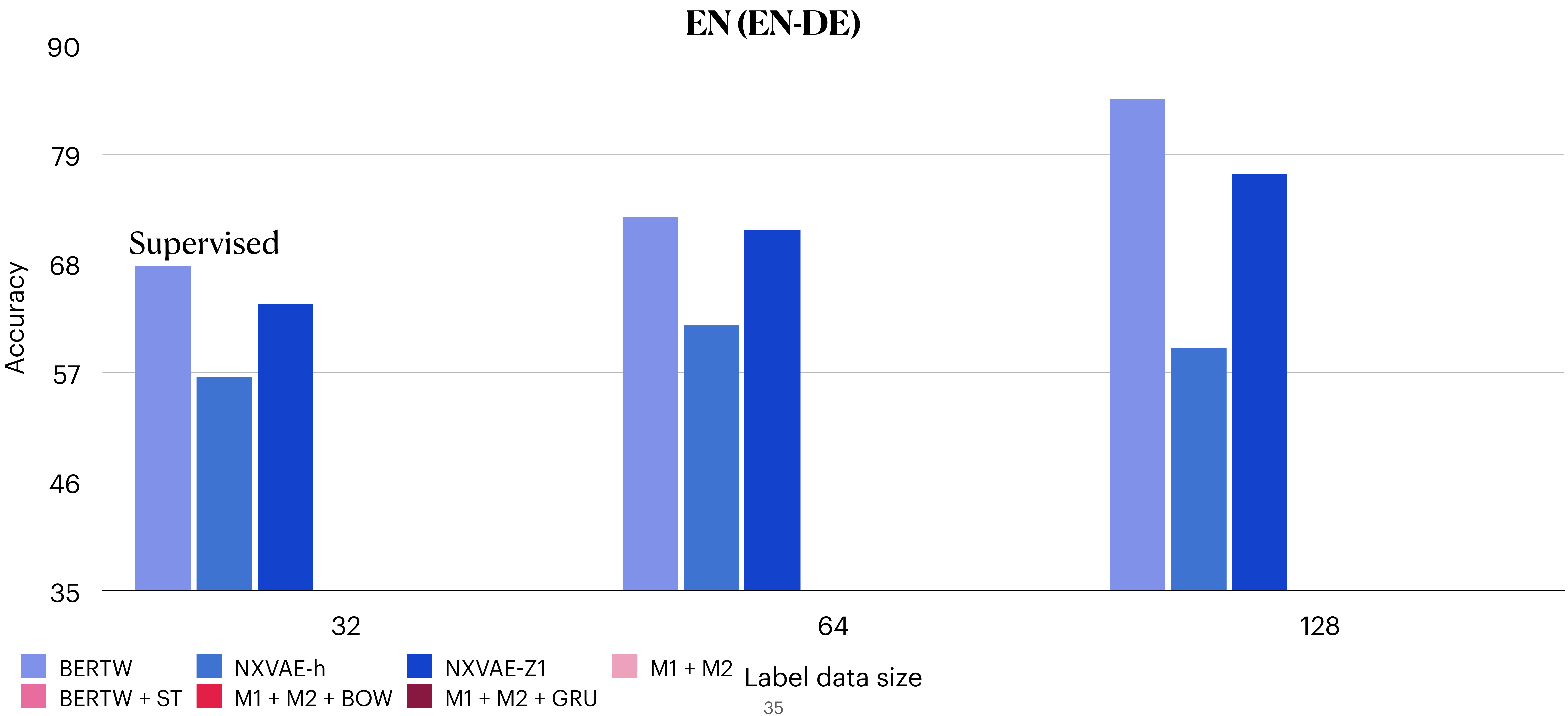
- **M1 + M2**

Baselines - NXVAE

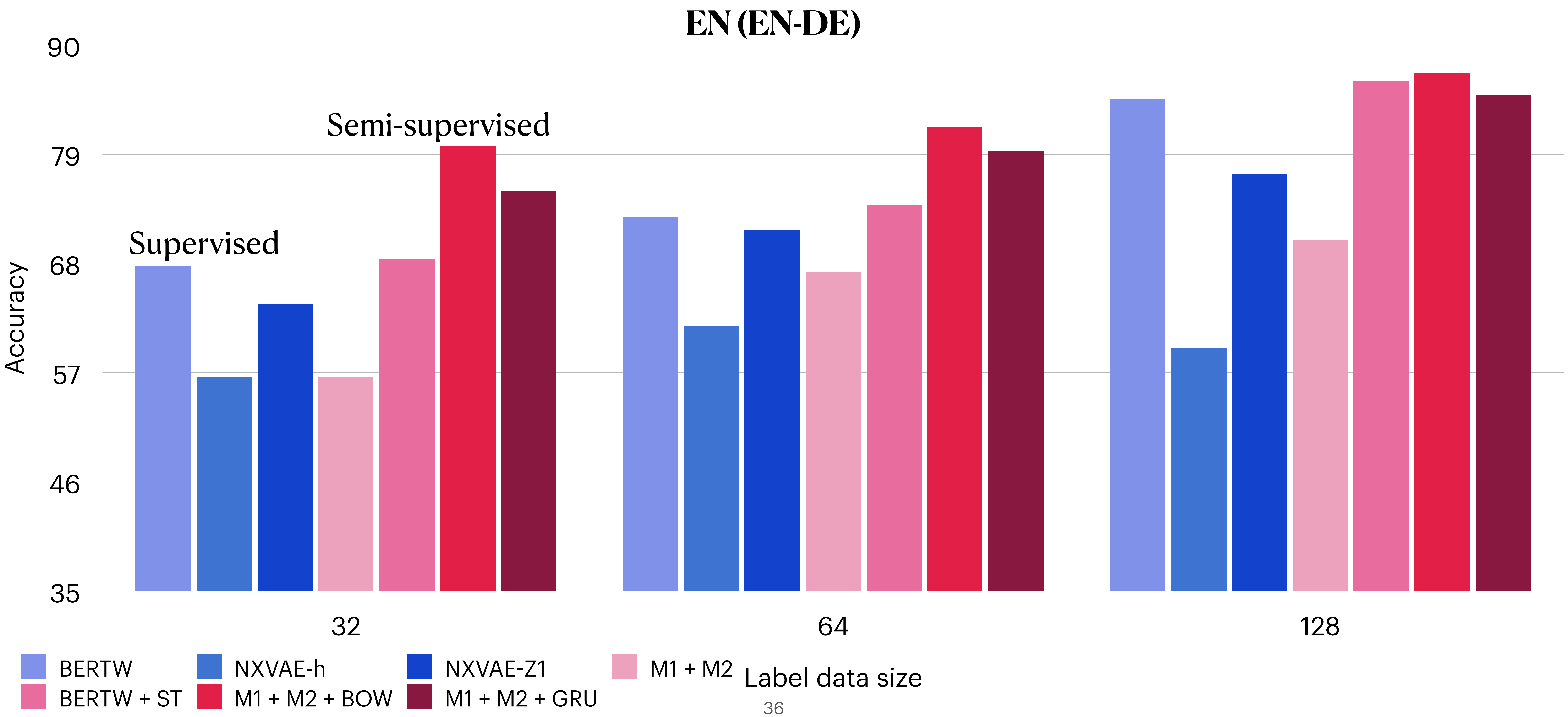
Semi-supervised models

- **M1** + **M2**
- BERTW with self-training (**BERTW** + **ST**)

Results

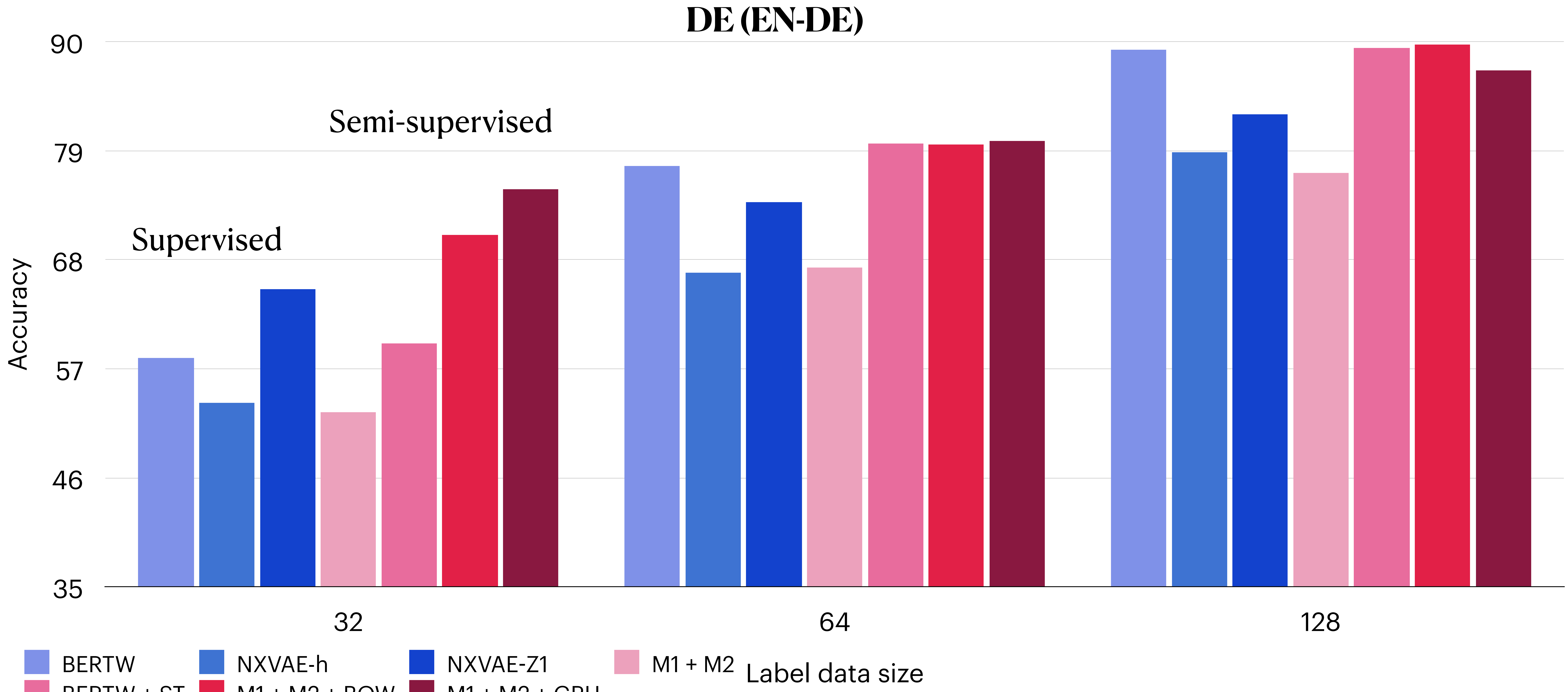

EN (EN-DE)
90
79
68
57
46
35
Accuracy
Supervised
32
64
128
Label data size
BERTW
NXVAE-h
NXVAE-Z1
M1 + M2
BERTW + ST
M1 + M2 + BOW
M1 + M2 + GRU

Results



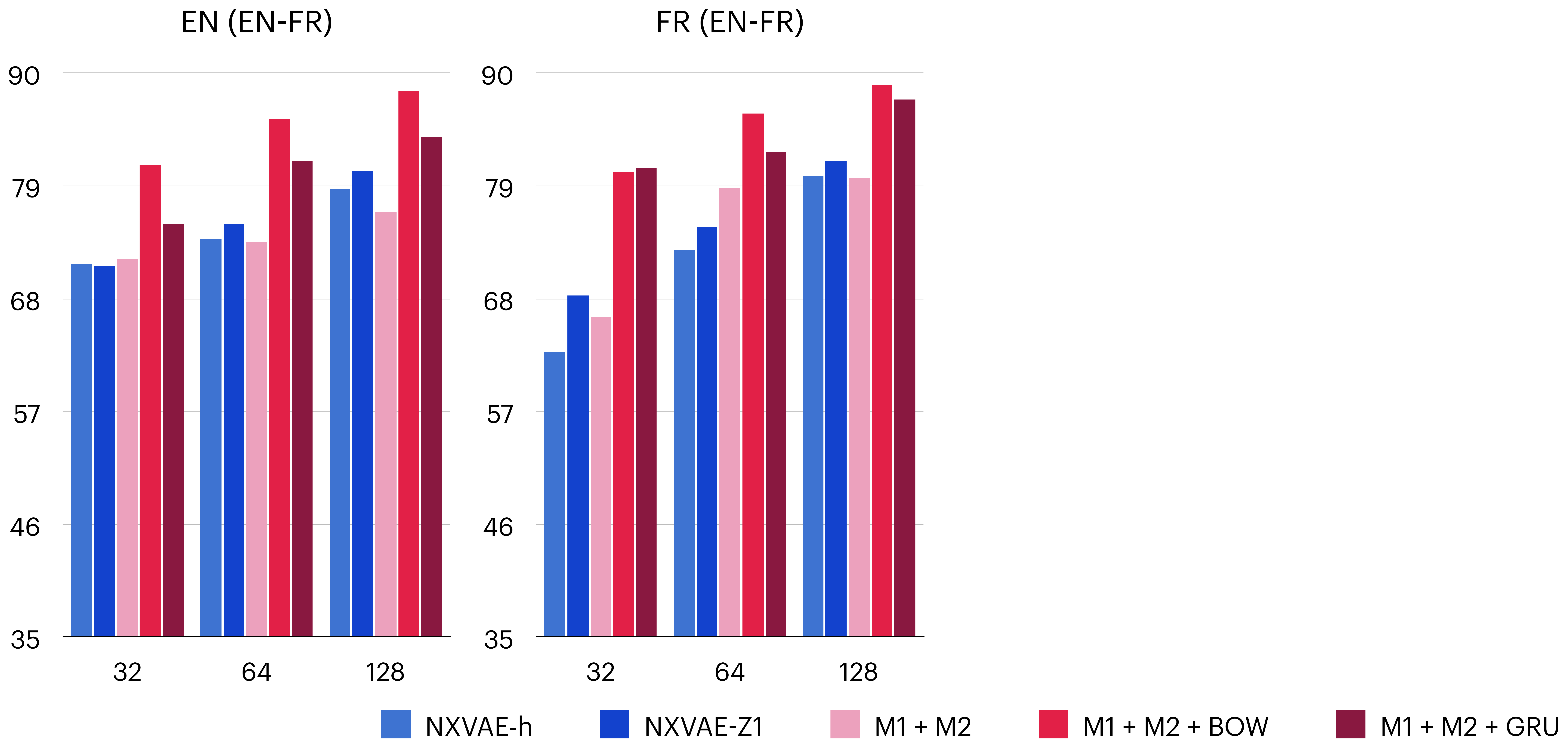
EN (EN-DE)
90
79
68
57
46
35
Accuracy
Semi-supervised
Supervised
32
64
128
BERTW
NXVAE-h
NXVAE-Z1
M1 + M2
BERTW + ST
M1 + M2 + BOW
M1 + M2 + GRU
Label data size

Results


DE (EN-DE)
90
79
68
57
46
35
Accuracy
Semi-supervised
Supervised
32
64
128
BERTW
NXVAE-h
NXVAE-Z1
M1 + M2
BERTW + ST
M1 + M2 + BOW
M1 + M2 + GRU
Label data size

Results

NXVAE for EN-FR and DE-FR

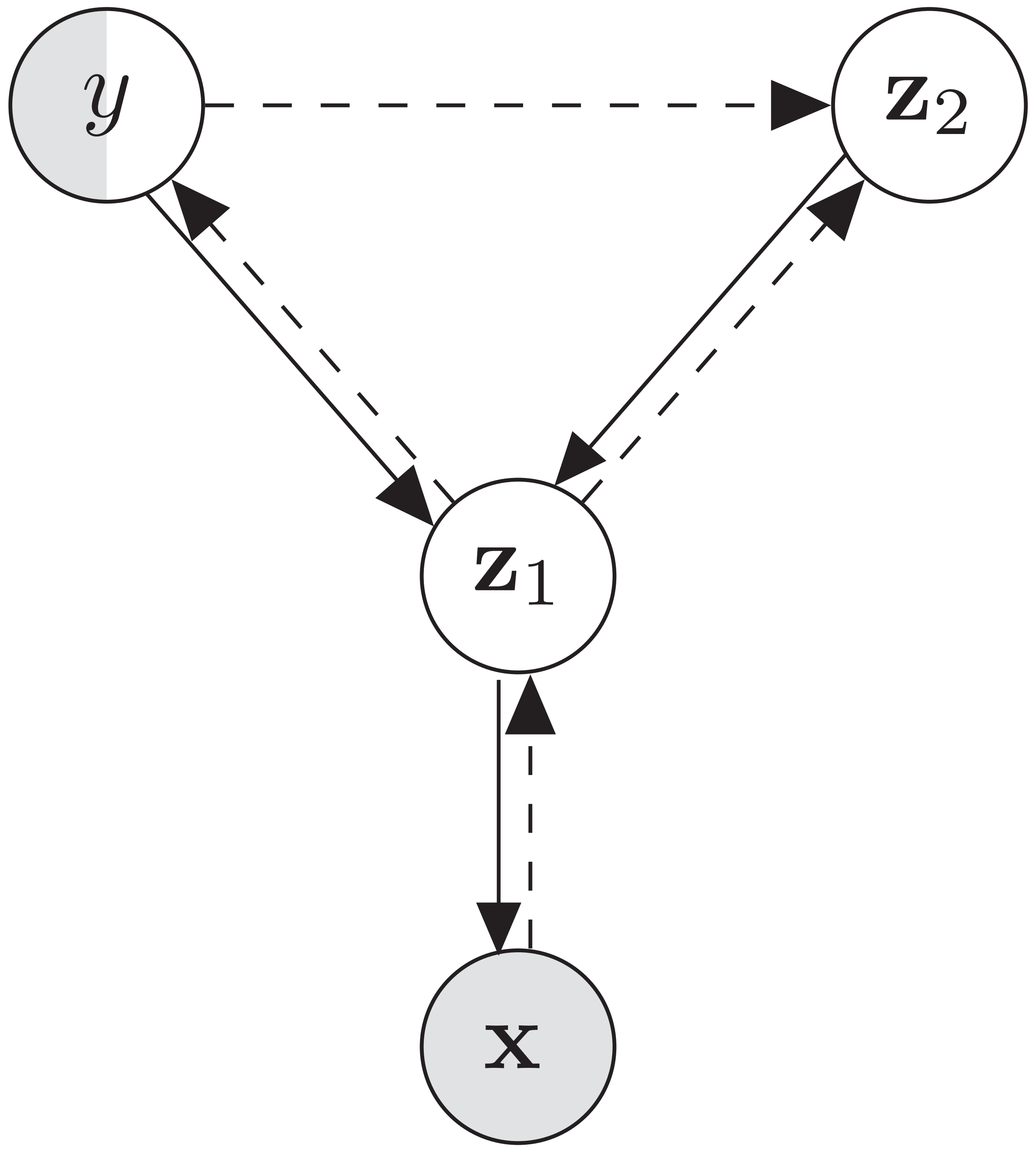


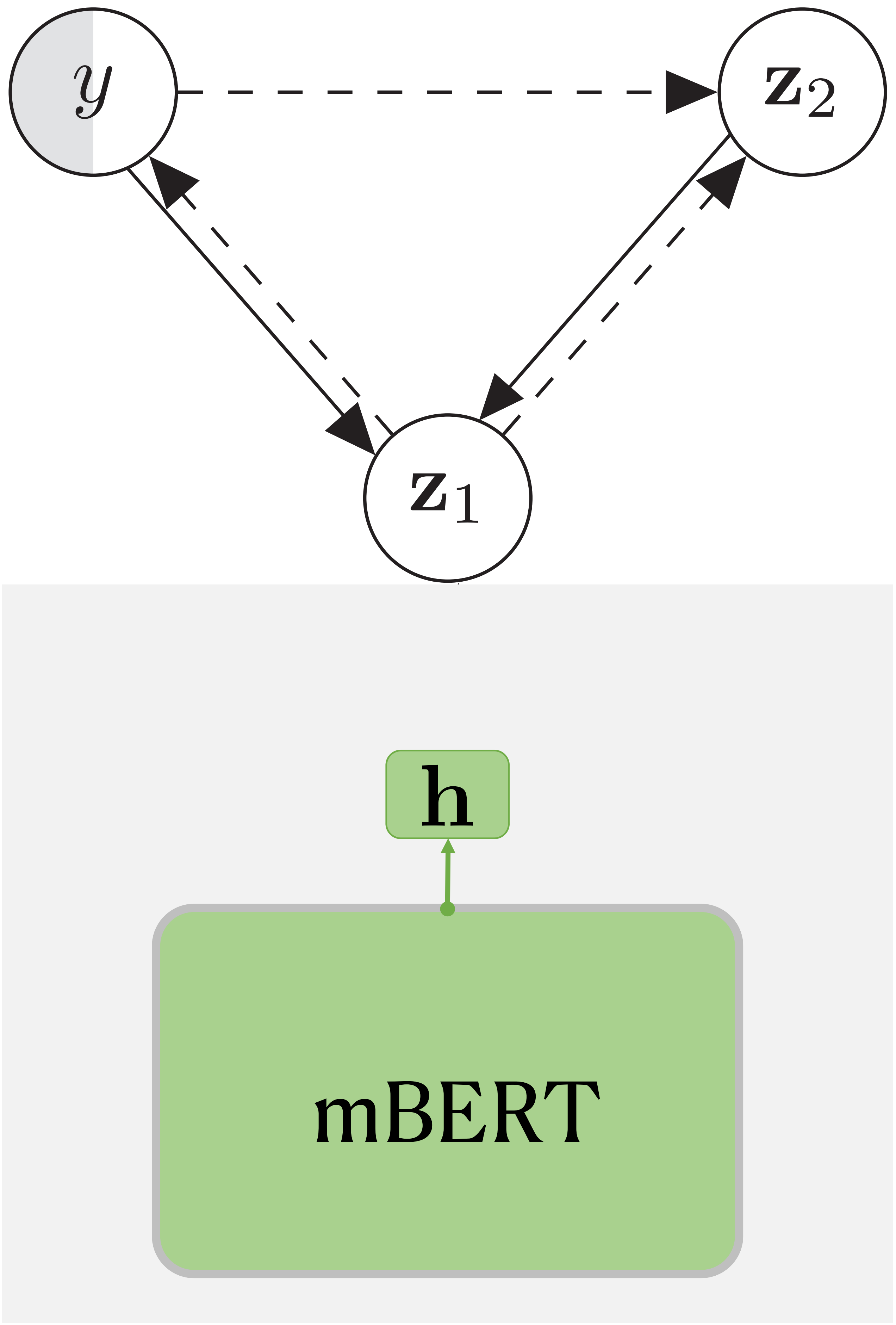
Results

NXVAE for EN-FR and DE-FR

SDGMs + Multi-lingual pretraining

Multi-lingual BERT (mBERT)

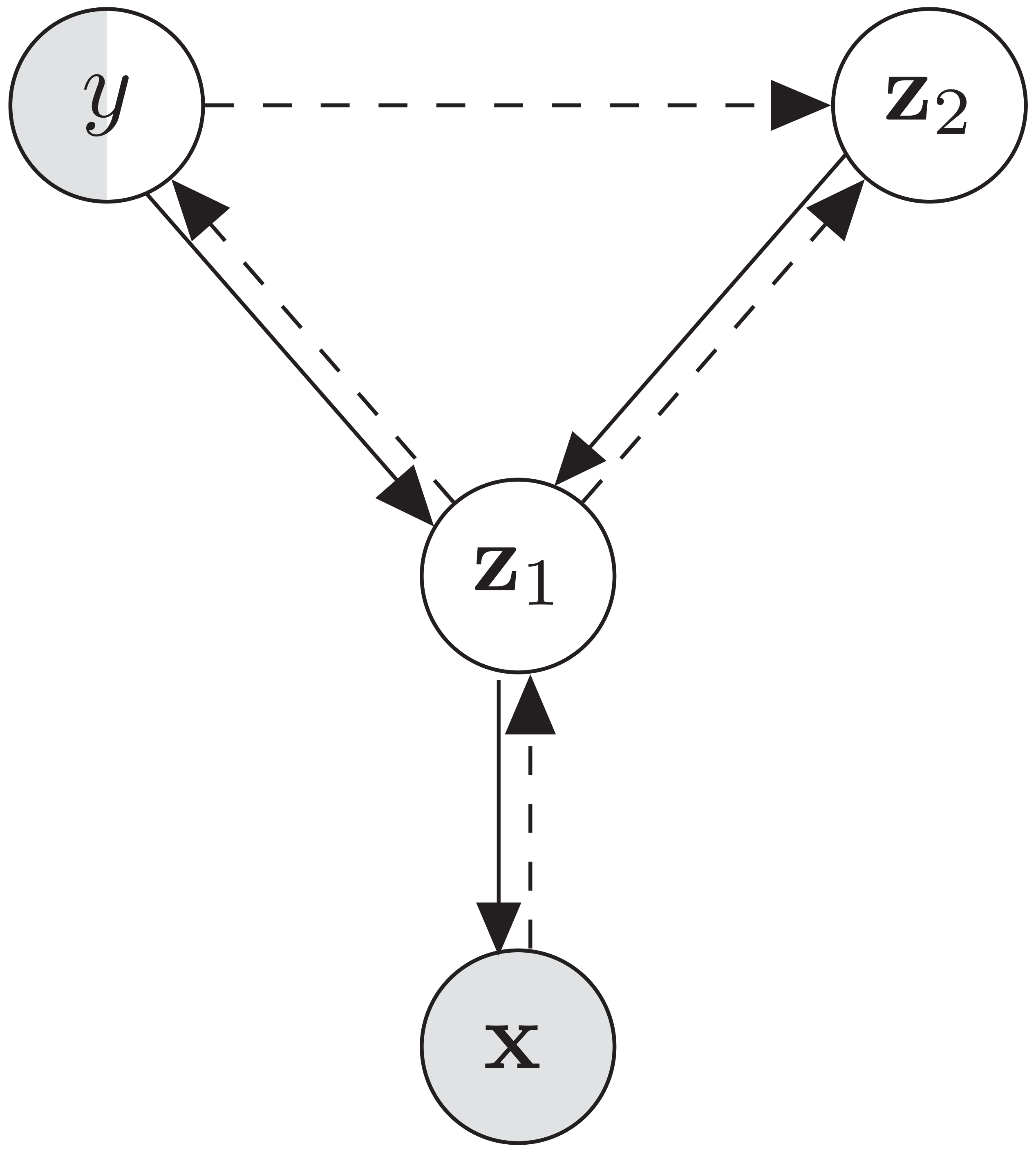


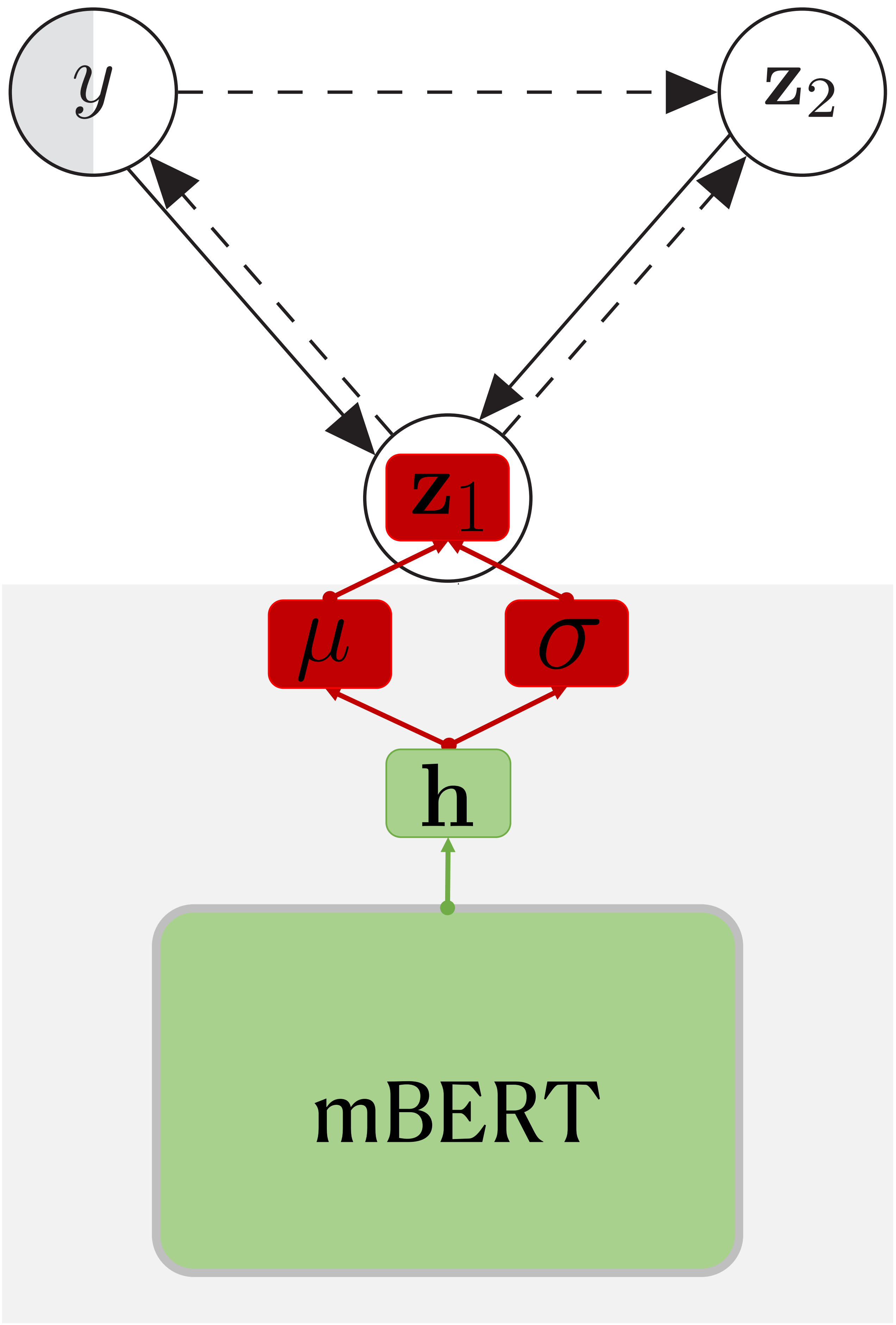


[Devlin et al. 2019]

SDGMs + Multi-lingual pretraining

Multi-lingual BERT (mBERT)



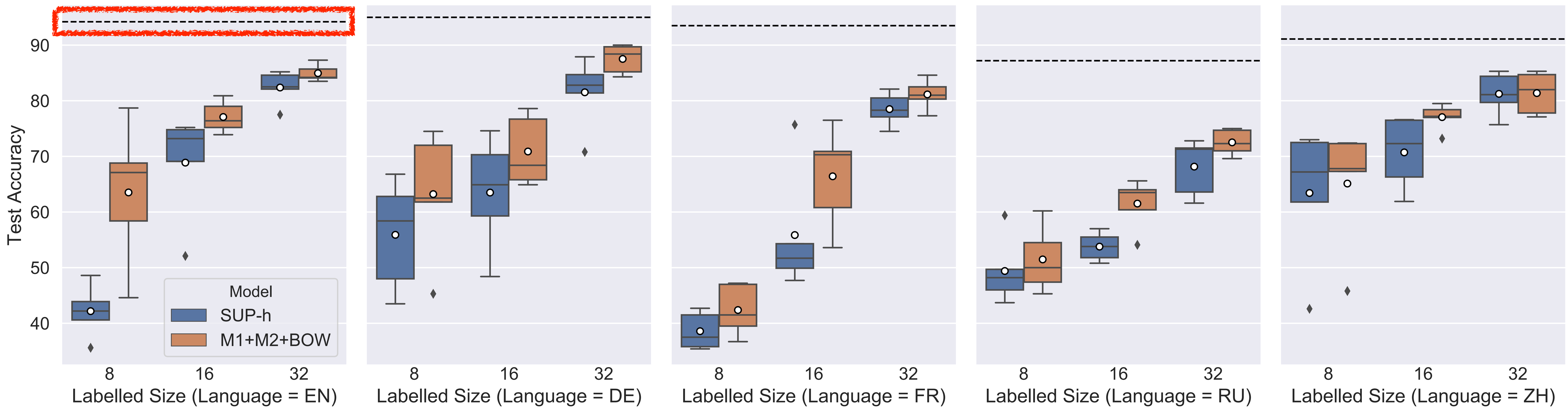


[Devlin et al. 2019]

Results

mBERT

SUP-h with full training data (1k)



Conclusion

- We bridged between multi-lingual pretraining and deep generative models to form a semi-supervised learning framework for multi-lingual document classification

Conclusion

- We bridged between multi-lingual pretraining and deep generative models to form a semi-supervised learning framework for multi-lingual document classification
- Our framework outperformed competitive baseline models, including supervised mBERT
- The benefits of SDGMs are orthogonal to the encoding architecture, opening up a new avenue for SDGMs in low-resource NLP

Thank you! Questions?

Contact: yz568@cam.ac.uk