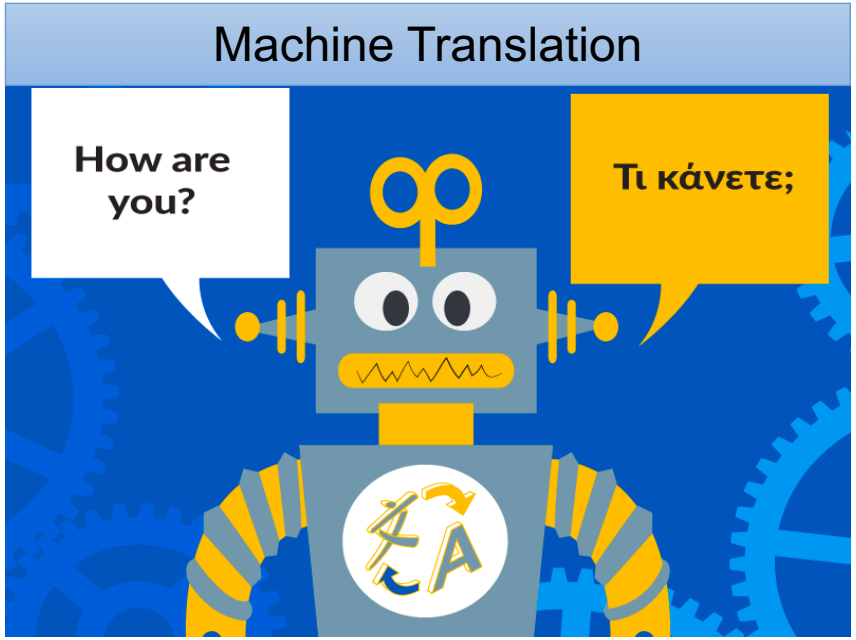
ControlVAE: Controllable Variational Autoencoder

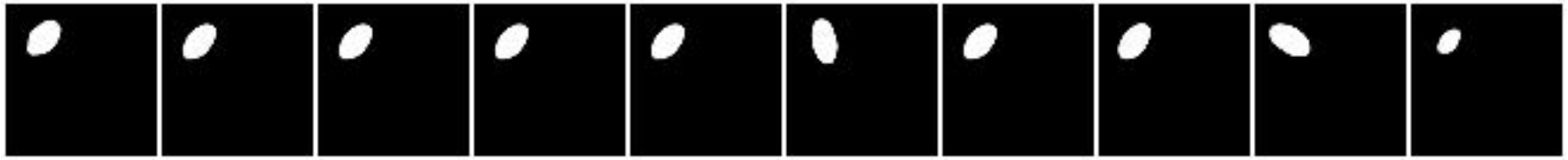
Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang,
Shengzhong Liu, Dongxin LiuJun Wang, Tarek Abdelzaher

University of Illinois at Urbana-Champaign
Amazon Web Services Deep Learning
Alibaba Inc. at Seattle

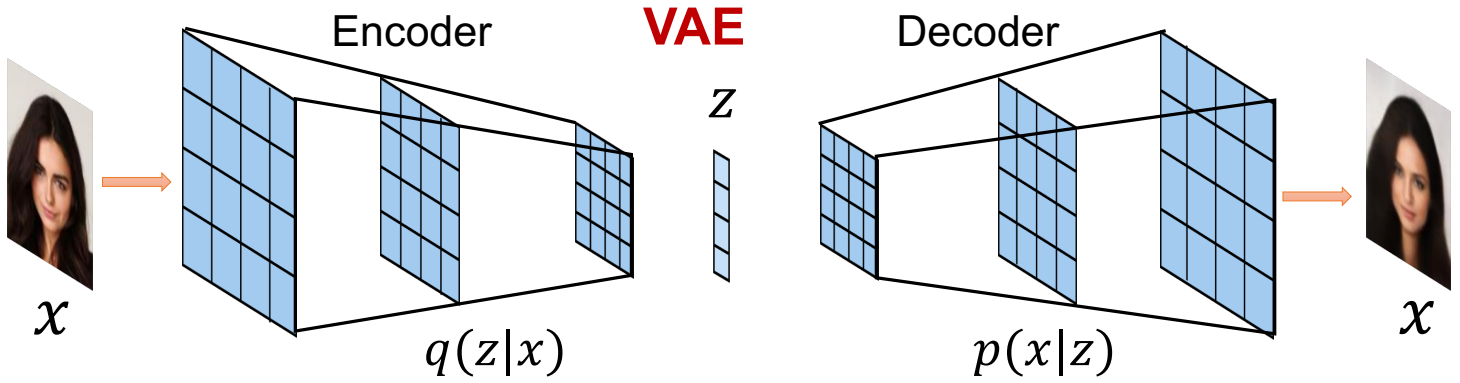
Background--VAE



Disentanglement representation learning



VAE model



Fig, The basic VAE model

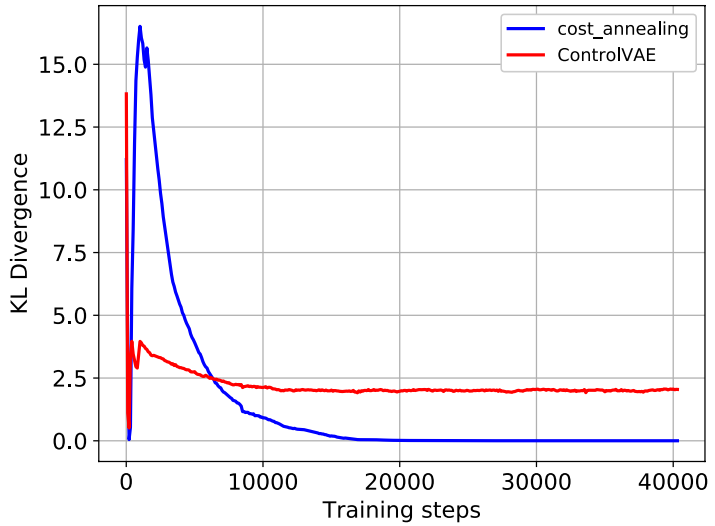
ELBO objective function

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

Recon. term **KL- divergence**

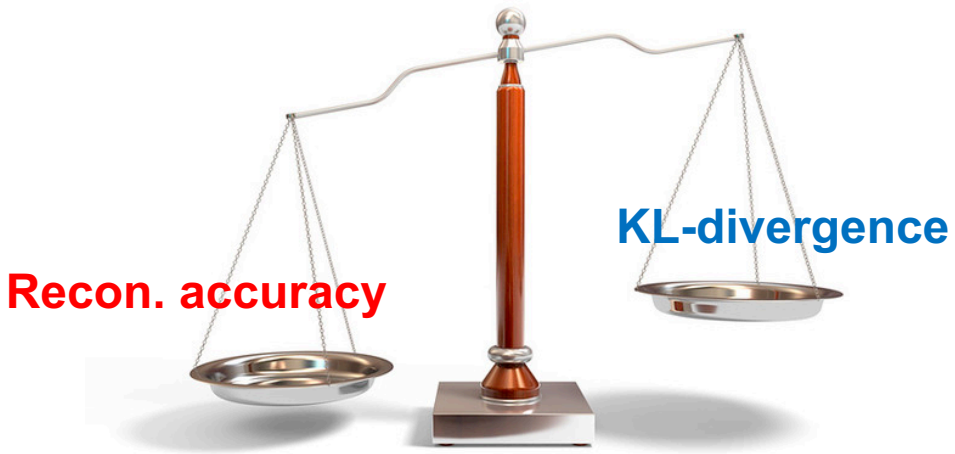
Background

- KL-vanishing (posterior collapse)
  - ➢ KL tends to zero during model training



KL vanishing

- Trade-off between KL-divergence and reconstruction quality



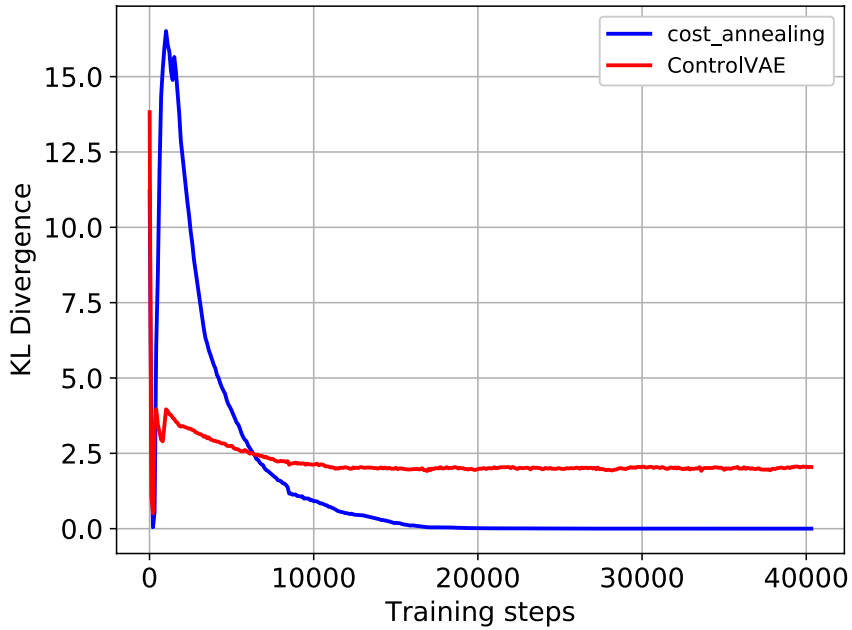
Related work

Study	Description	Cons
Cost annealing (bowman2015)	increase weight β on KL from 0 until to 1 using sigmoid function after *N* steps	still suffer from KL-vanishing
β-VAE (higgins2017)	assign a large and fixed weight to KL term i	high recon. error
TamingVAE (rezende2018)	f c	inima
FactorVAE (kim2018)	Decompose KL into three terms: Index_code, total correlation and wise-KL	fixed weight, has high recon. error
infoVAE (zhao2017)	Add a mutual information maximization term to encourage mutual information between x and z	• fixed weight • cannot explicitly control KL value

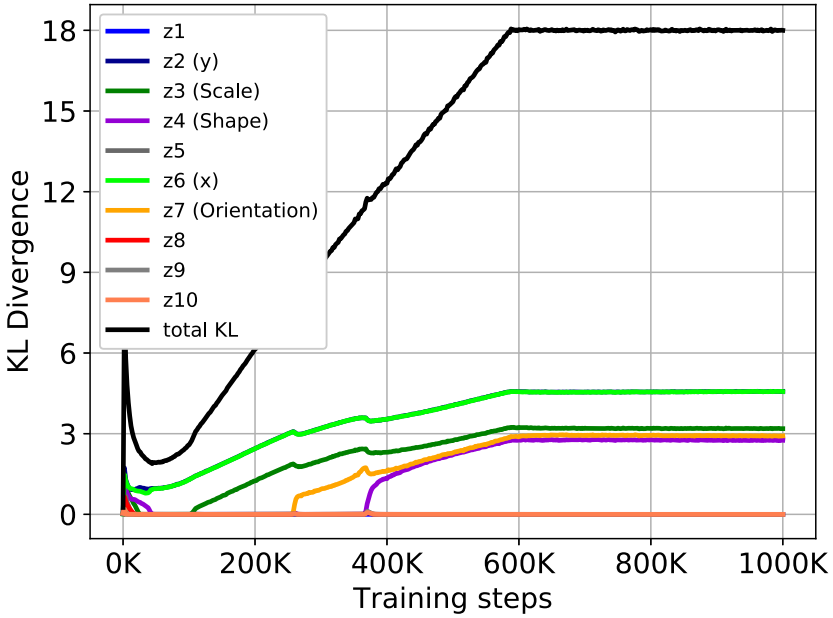
Drawback of existing work:

1. Fixed weight on KL term, leading to high recon. error
2. KL vanishing (posterior collapse)

Motivation



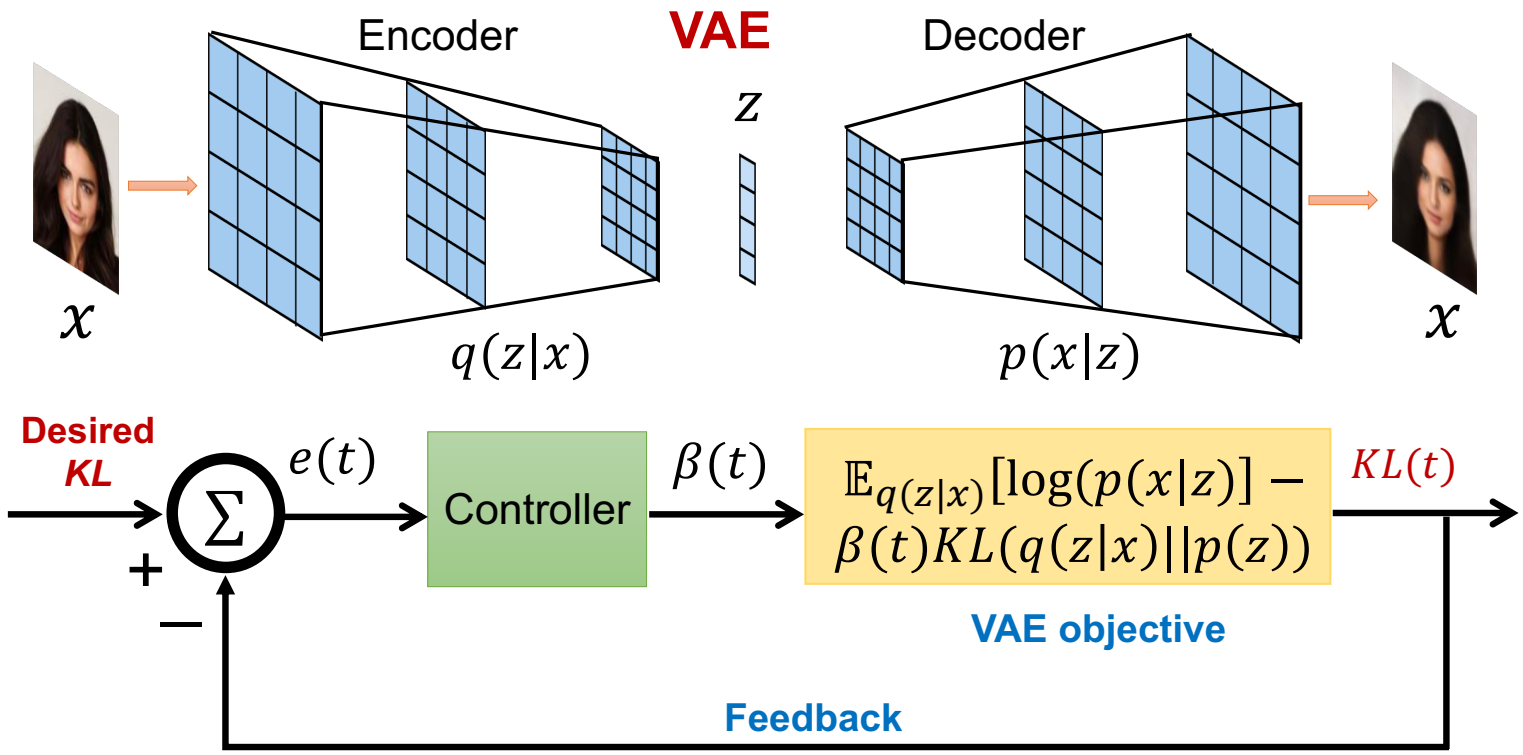
[1] Language modeling: KL vanishing



[2] Disentanglement: information capacity (KL-divergence)

Control KL-divergence !!!

ControlVAE Framework



Fig, Framework of ControlVAE via dynamic learning

ControlVAE Model

Objective function:

$$\mathcal{L} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta(t) D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})),$$

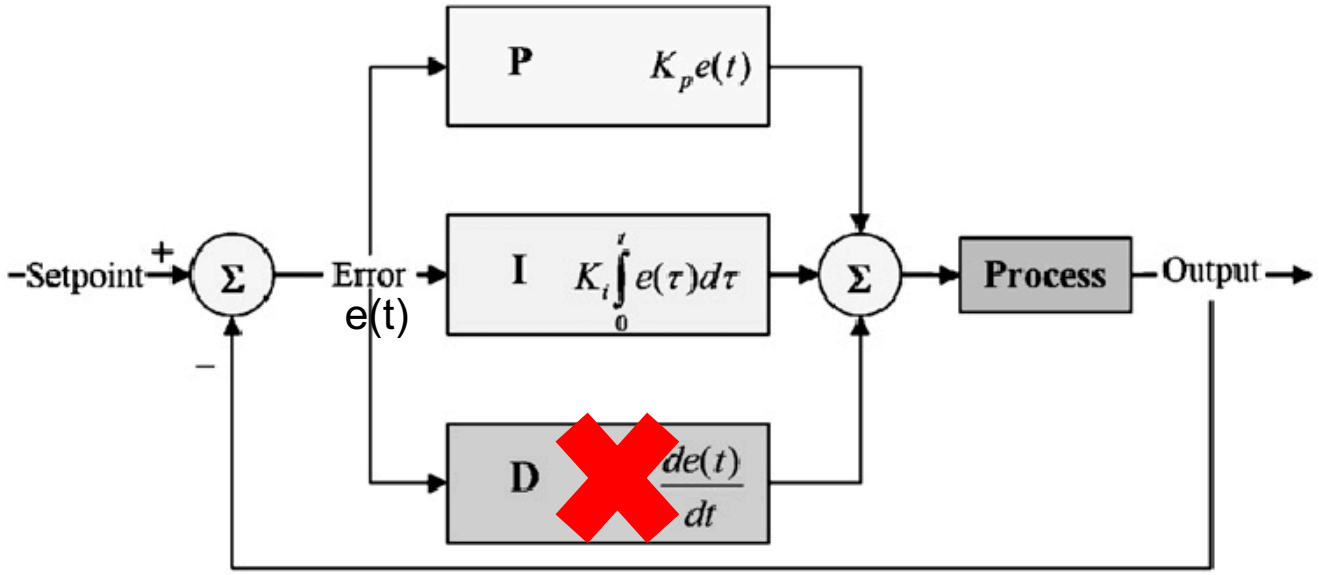
Where β (t) is the output of a controller

PID control algorithm

PID algorithm

P term | I term | D term

$$\beta(t) = K_p e(t) + K_i \int_0^t e(\tau)d\tau + K_d \frac{de(t)}{dt},$$



- *e(t)* is the error between the real KL-divergence and the set point
- K_p is the coefficient for proportional (P) term
- K_i is the coefficient for integer (I) term
- K_d is the coefficient for derivative (D) term

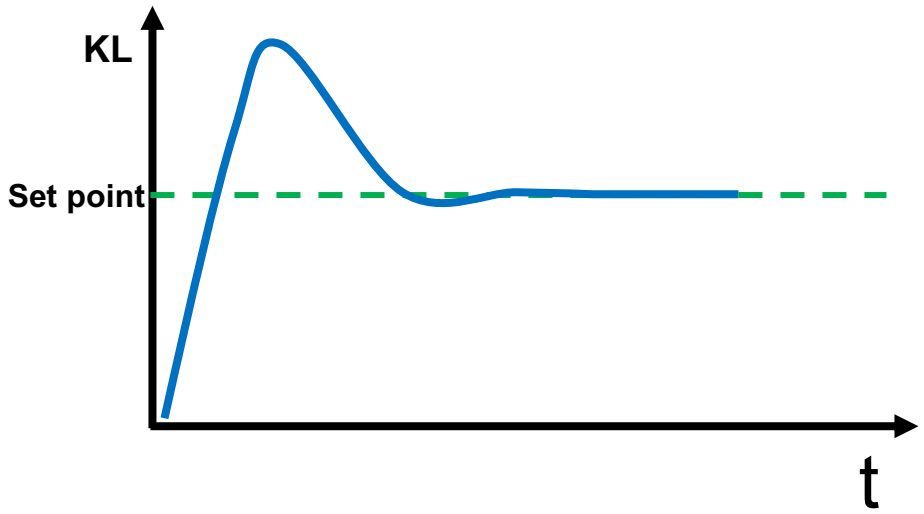
Non-linear PI Controller

application-specific constant

$$\beta(t) = \frac{K_p}{1 + \exp(e(t))} - K_i \sum_{j=0}^{t} e(j) + \beta_{min},$$

Insight of PI controller

- When e(t) > 0: output $\widehat{KL}$(t) is very small, reduce $\beta(t)$, boost KL value;
- When e(t) < 0: output $\widehat{KL}$(t) is larger than set point, increase $\beta(t)$ to optimize KL term;



Non-linear PI Controller

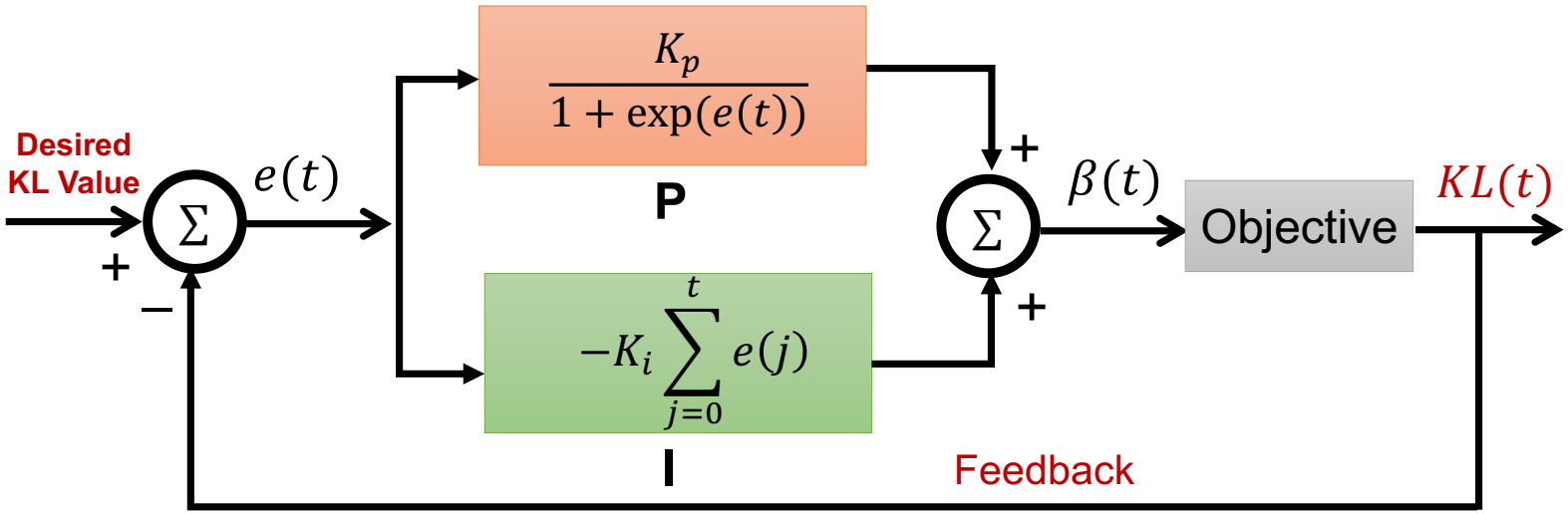


Fig. PI controller

Evaluation

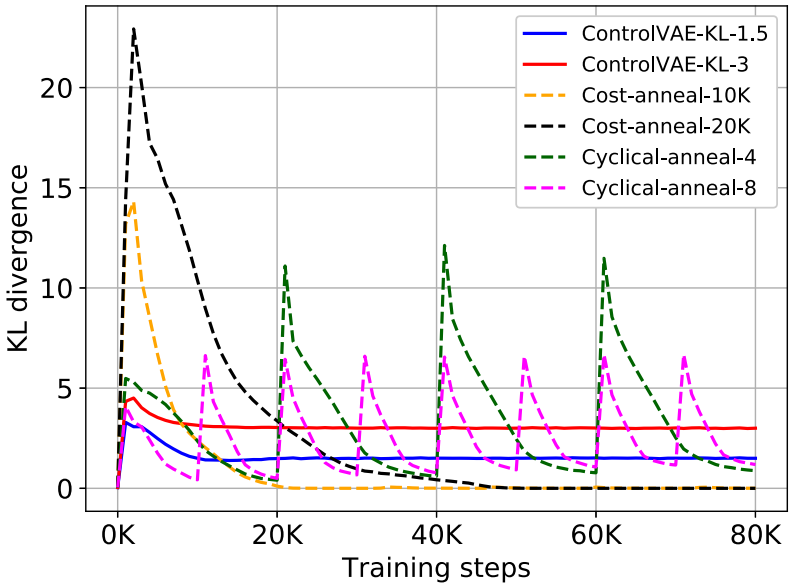
Applications:

- ❖ **Language modeling: text and dialog generation**
- ❖ **Disentanglement representation learning**
- ❖ **Image generation**

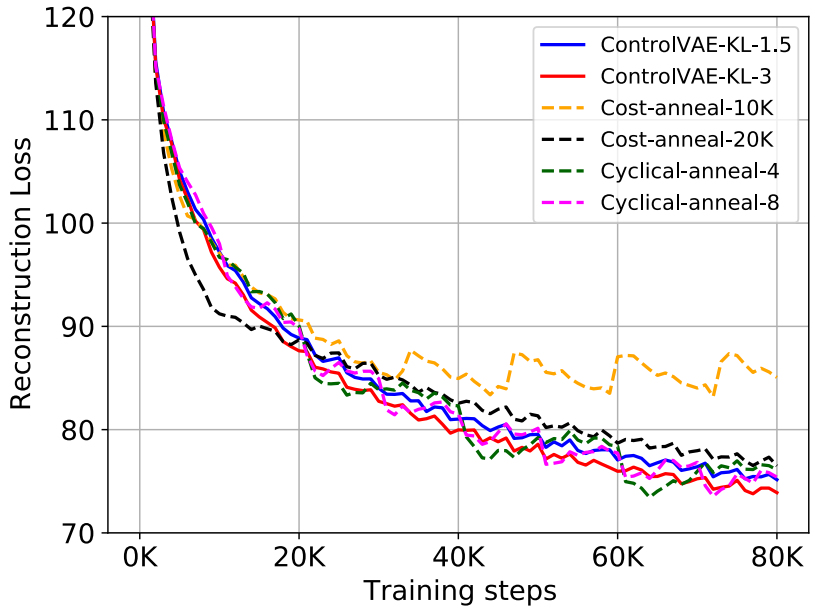
Benchmark datasets:

- **Language modeling:** [1] Penn Tree Bank (PTB) [2] Switchboard(SW) telephone conversation
- **Disentanglement:** DSprites
- **Image generation:** CelebA

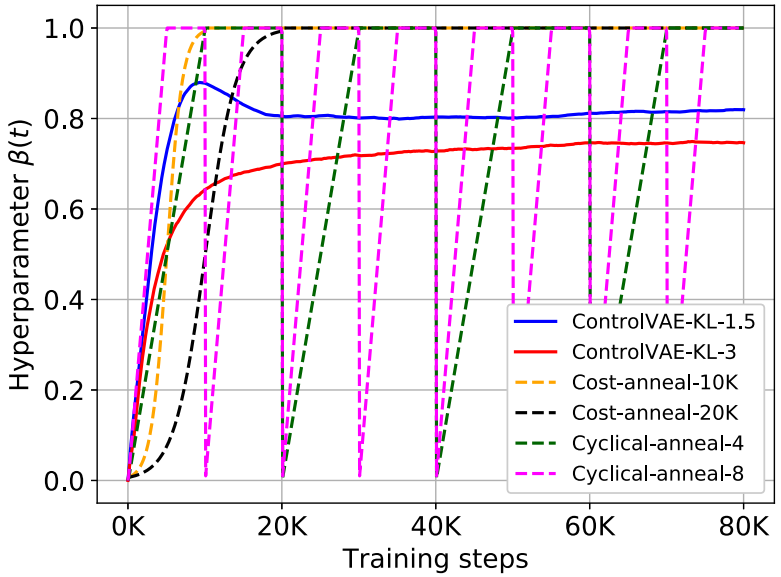
Evaluation: Language modeling (PTB data)



(a) KL divergence



(b) Recon. loss



(c) Weight $\beta(t)$

Baselines:

1) **Cost annealing**: gradually increases the weight on KL-divergence from 0 until to 1 after N steps using Sigmoid function
2) **Cyclical annealing**: splits the training process into M cycles and each increases the weight from 0 until to 1 using a linear function

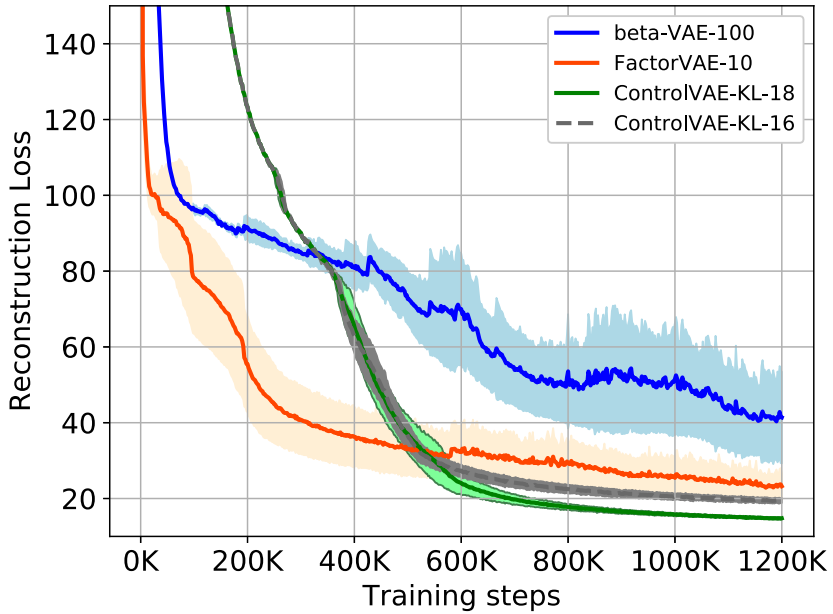
Evaluation: Language modeling

Switchboard (SW) to measure the diversity of generated text

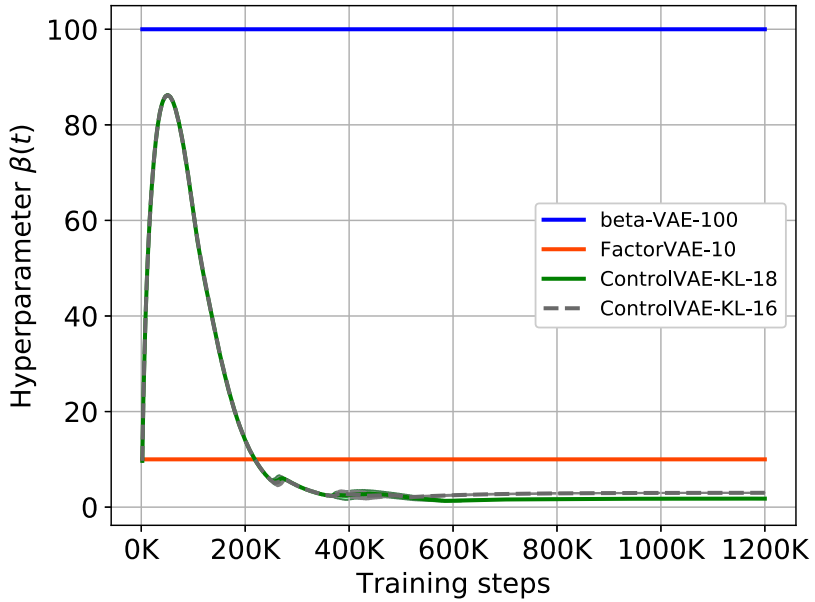
Table 1. Performance comparison for different methods on dialog-generation using SW data over 5 random seeds. Dis-n: higher is better. PPL: lower is better, and self-BLEU lower is better.

Methods/metric	Dis-1	Dis-2	self-BLEU-2	self-BLEU-3	PPL
ControlVAE-KL-35	**6.27K** $\pm$ 41	**95.86K** $\pm$ 1.02K	**0.663** $\pm$ 0.012	**0.447** $\pm$ 0.013	**8.81** $\pm$ 0.05
ControlVAE-KL-25	6.10K $\pm$ 60	83.15K $\pm$ 4.00K	0.698 $\pm$ 0.006	0.495 $\pm$ 0.014	12.47 $\pm$ 0.07
Cost anneal-KL-17	5.71K $\pm$ 87	69.60K $\pm$ 1.53K	0.721 $\pm$ 0.010	0.536 $\pm$ 0.008	16.82 $\pm$ 0.11
Cyclical (KL = 21.5)	5.79K $\pm$ 81	71.63K $\pm$ 2.04K	0.710 $\pm$ 0.007	0.524 $\pm$ 0.008	17.81 $\pm$ 0.33

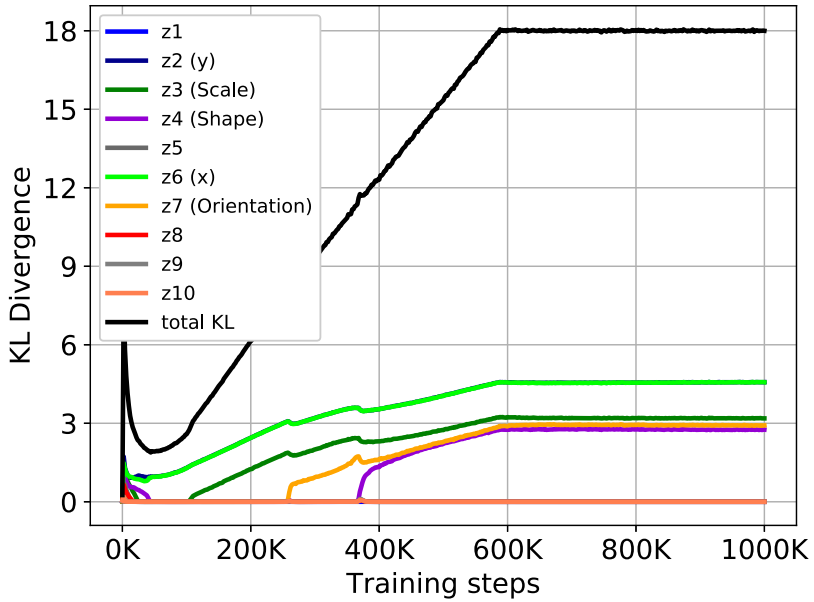
Evaluation: Disentanglement (Dsprites data)



(a) Recon. error



(b) Weight $\beta(t)$



(c) Disentangled factors

Baselines:

1) **\beta-VAE**: Burgess, C. P., Higgins, I., Pal, A., Matthey,et al. (2018). Understanding disentangling in $\beta $-VAE. arXiv preprint arXiv:1804.03599.
2) **FactorVAE**: Kim, Hyunjik, and Andriy Mnih. "Disentangling by Factorising." In International Conference on Machine Learning, pp. 2649-2658. 2018.

Evaluation: Disentanglement

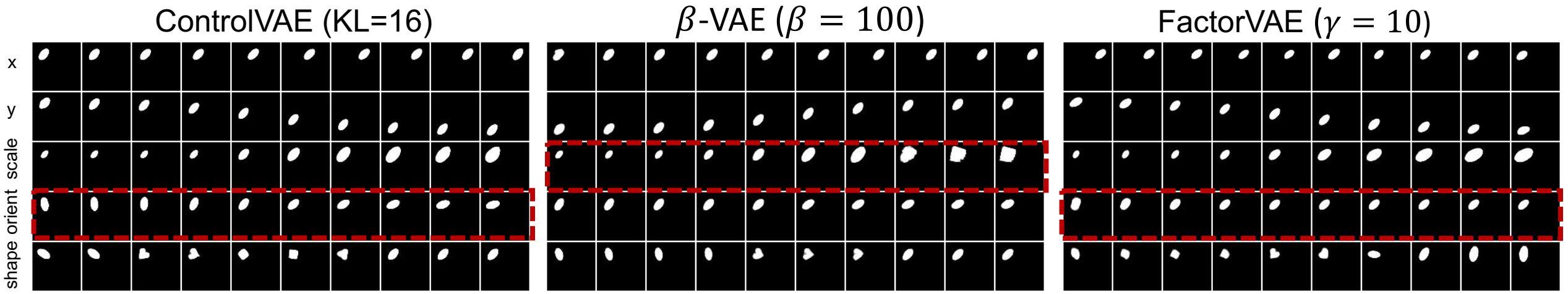
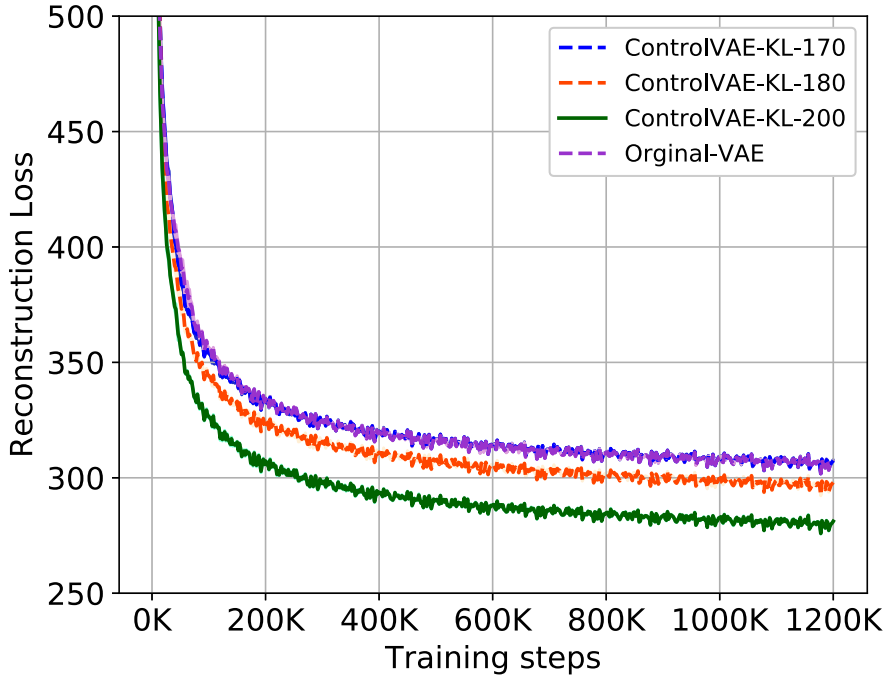


Fig., Example of traverse a single latent dimension in a range of [-3, 3]

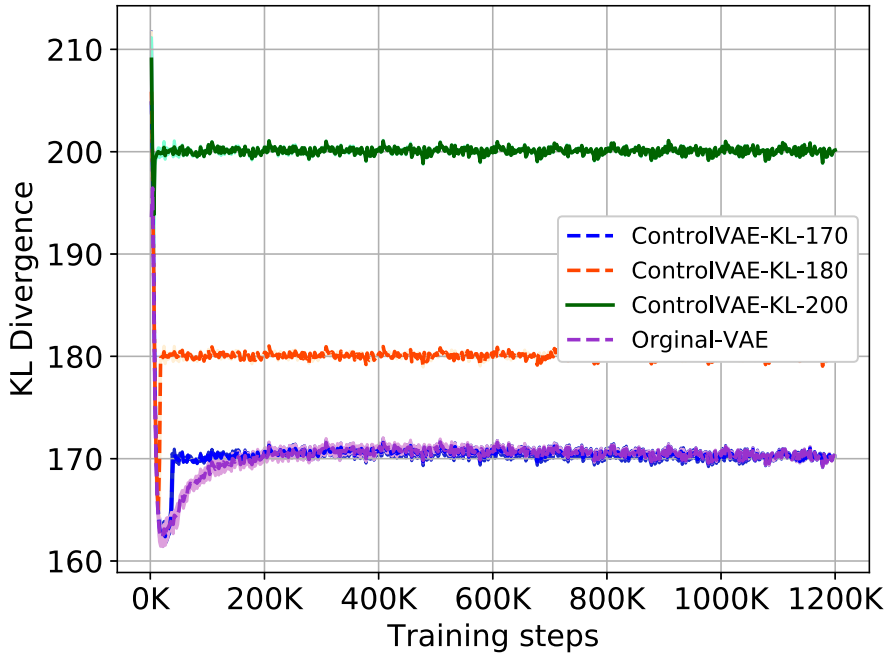
Table 2. Performance comparison of different methods using disentanglement metric, MIG score, averaged over 5 random seeds. The higher is better. ControlVAE (KL=16) has a comparable MIG score but lower variance than the FactorVAE with the default parameters.

Metric	ControlVAE (KL=16)	ControlVAE (KL=18)	β-VAE ($\beta = 100$)	FactorVAE ($\gamma = 10$)
MIG	**0.5628 ± 0.0222**	0.5432 ± 0.0281	0.5138 ± 0.0371	0.5625 ± 0.0443

Evaluation: Image generation



(a) Recon. loss



(b) KL divergence

Table 3. Performance comparison for different methods on CelebA data over 3 random seeds. FID: lower is better. SSIM: higher is better.

Methods/metric	FID	SSIM
ControlVAE-KL-200	**55.16** $\pm$ 0.187	**0.687** $\pm$ 0.0002
ControlVAE-KL-180	57.57 $\pm$ 0.236	0.679 $\pm$ 0.0003
ControlVAE-KL-170	58.75 $\pm$ 0.286	0.675 $\pm$ 0.0001
Original VAE	58.71 $\pm$ 0.207	0.675 $\pm$ 0.0001

Conclusion

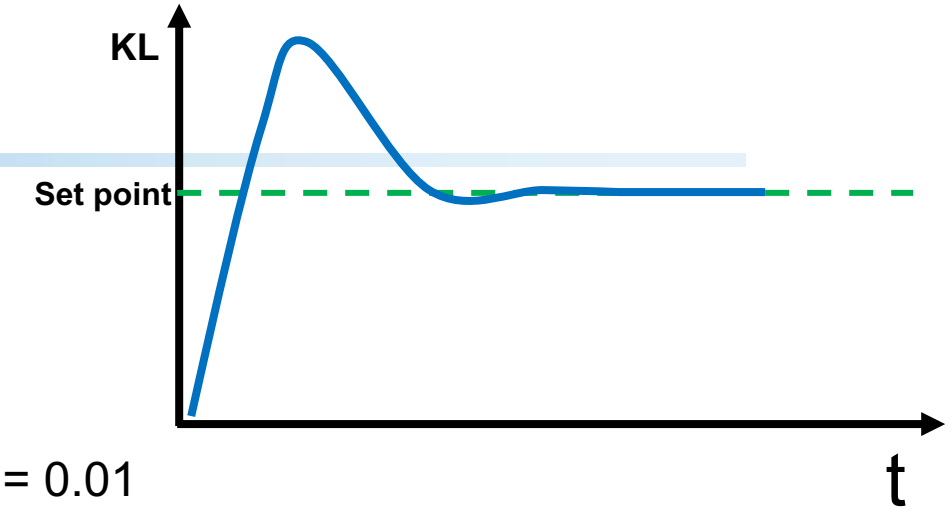
- Propose a new controllable VAE, ControlVAE, that combines a PI controller, with the basic VAE model.
- Design a new non-linear PI controller, to automatically tune the weight in the VAE objective.
- ControlVAE can not only avert the KL-vanishing, but also control the diversity of generated text.
- Achieve better disentangling and reconstruction quality than the existing methods.

Thank you very much!!

Q&A

Backup

PI Parameter Tuning



- Tune K_p , when output $\widehat{KL}$(t) is very small, error >>0,

P term

$$\frac{K_p}{1+\exp(e(t))} \leq \epsilon,$$

e.g., K_p = 0.01

- Tune K_i , when output $\widehat{KL}$(t) is very large, e(t) < 0

I term

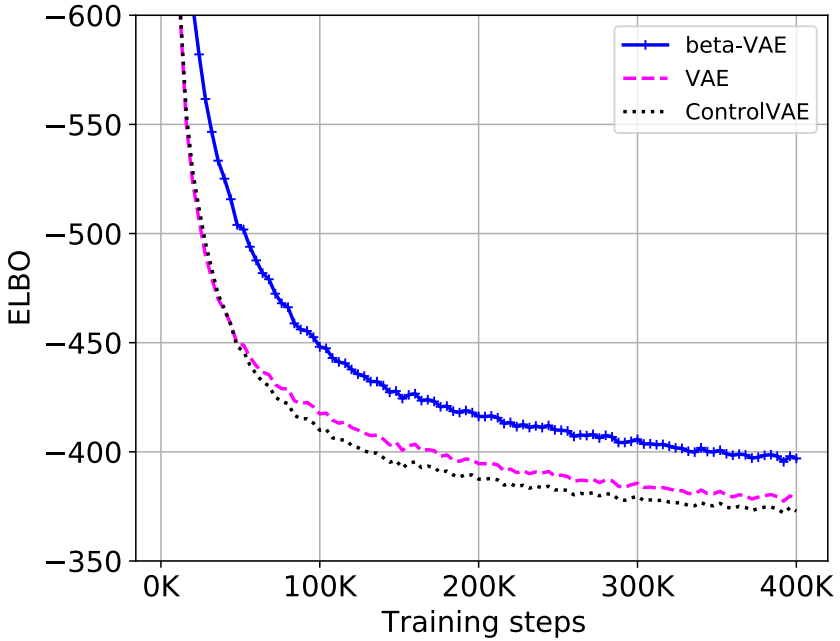
$$-K_i \sum_{j=0}^{t} e(j)$$

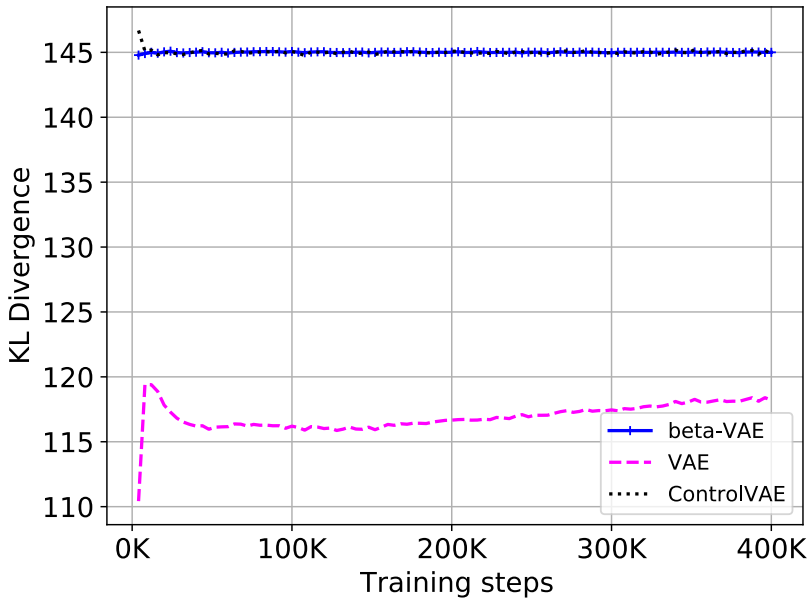
e.g., K_i = 0.001 or 0.0001

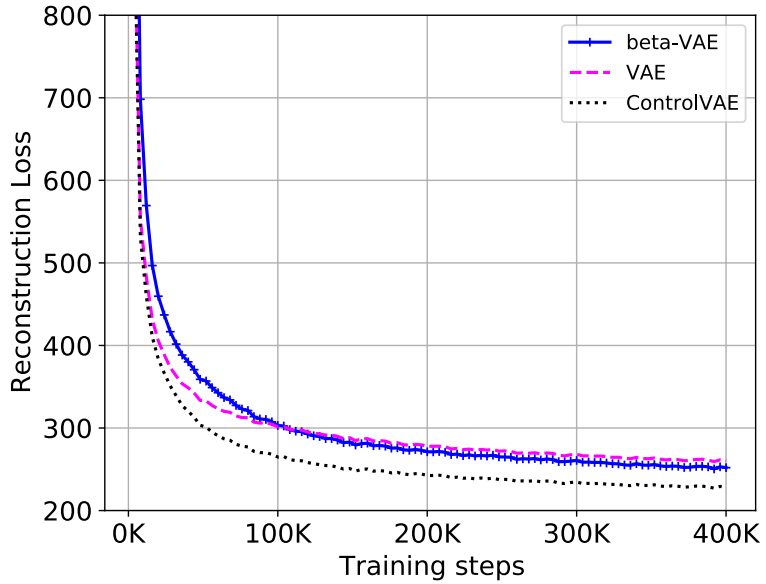
Set Point Guideline

- The set point of KL-divergence is largely application specific.
  - **Text generation**: slightly increase the KL-divergence, denoted by KL_{vae}, produced by the basic VAE or by Cost annealing method.
  - **ELBO improvement**: KL should be increased within the following bound

$$0 \leq d \leq 2 + 2\sqrt{2KL_{vae} + 1}.$$







ELBO improvement