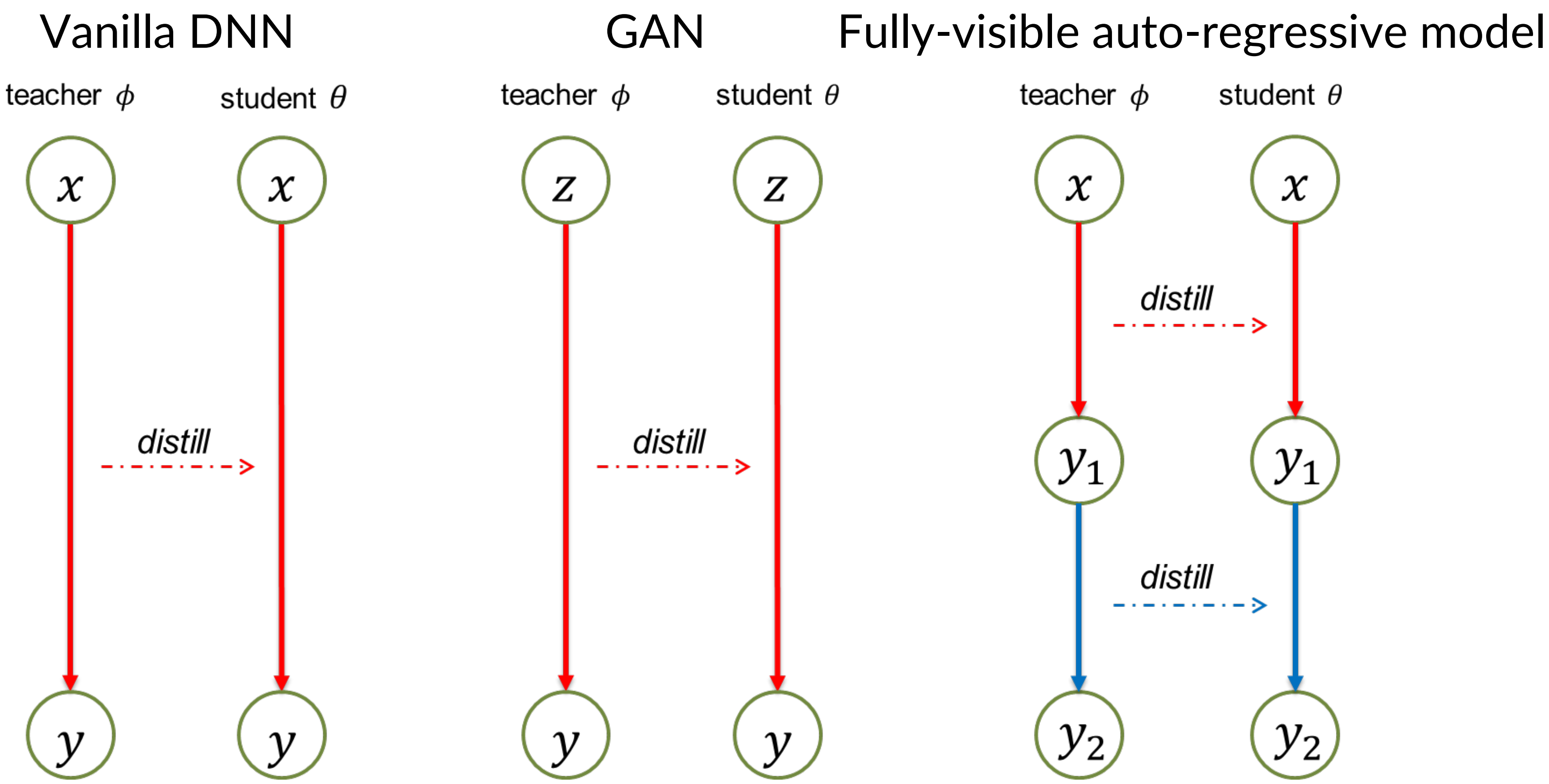
A Unified Knowledge Distillation Framework for Deep Directed Graphical Models

Yizhuo Chen, Kaizhao Liang, Zhe Zeng, Shuochao Yao, Huajie Shao

University of Illinois Urbana-Champaign
College of William & Mary
University of California, Los Angeles
George Mason University

Summary

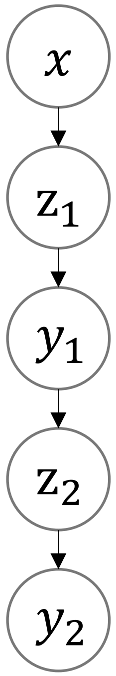
- **Knowledge Distillation (KD)：** Transferring knowledge from a teacher to a student model.
- **Motivation**: all existing KD methods were only applicable to **limited**, **specific** types of Directed Graphical Models (DGM) exclusively.
- **Goal**: propose a unified framework enabling KD for **general** DGMs.



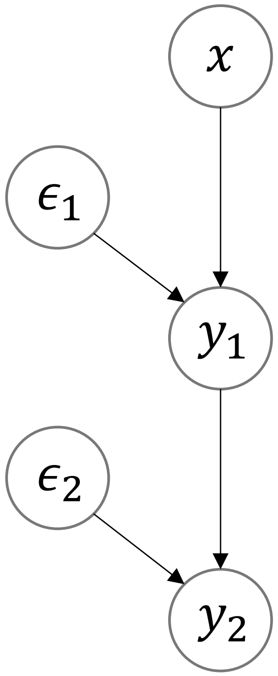
Summary

- **Semi-auxiliary Form:** reparametrize all latent variables
- **Novel KD loss on semi-auxiliary form**
  - Tractable
  - Shallower
  - Upper bound to vanilla KD
  - Proper generalization to multiple KD methods

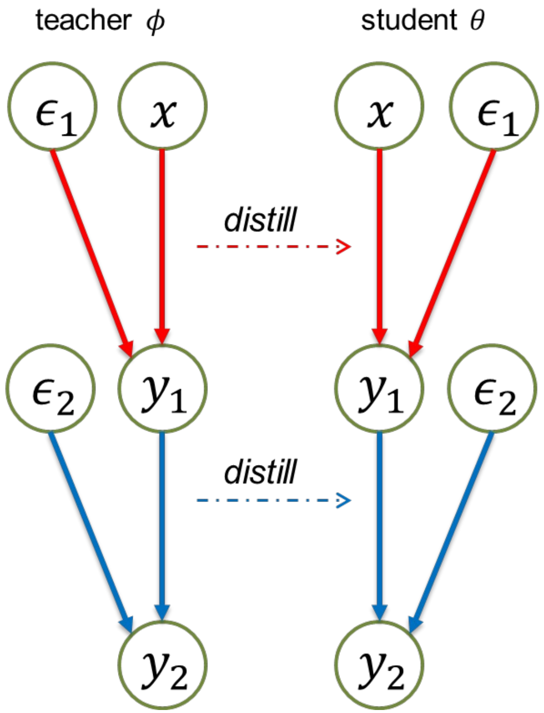
Original form



Semi-auxiliary form



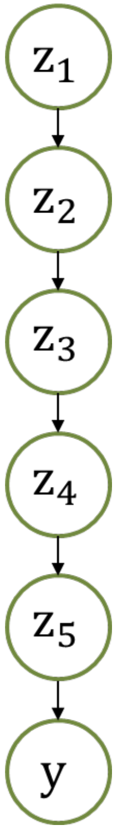
Our KD loss



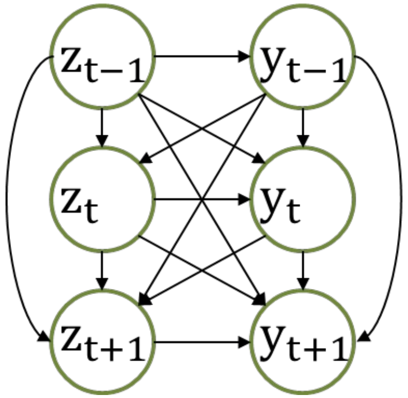
Summary

- **Evaluation Results:** Our method showed better performance on

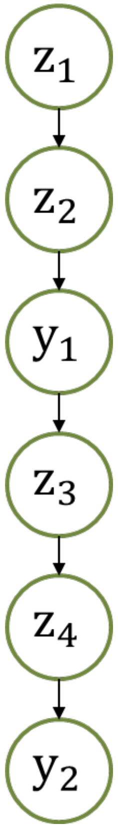
Data-free VAE compression



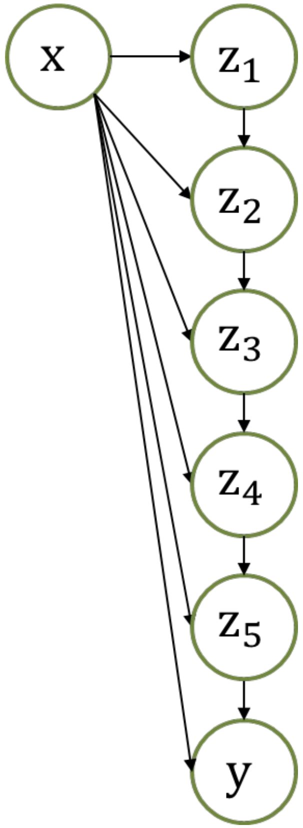
Data-free VRNN compression



Data-free HM compression

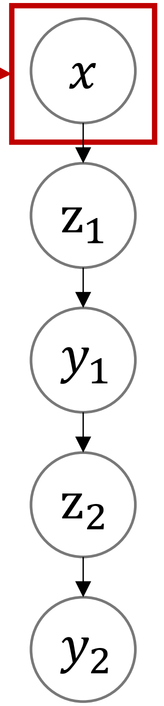


VAE continual learning



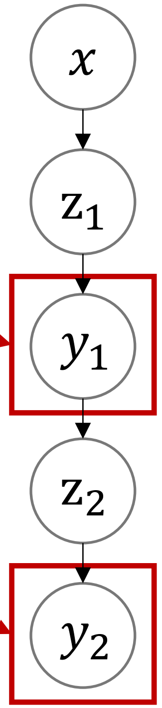
Background

- **Directed Graphical Model**
  - Multiple **input**, target as well as latent variables



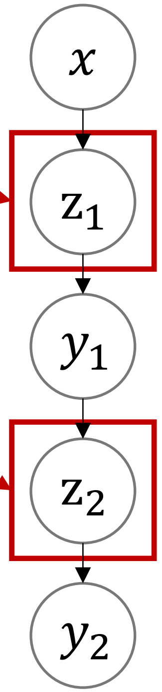
Background

- **Directed Graphical Model**
  - Multiple input, **target** as well as latent variables



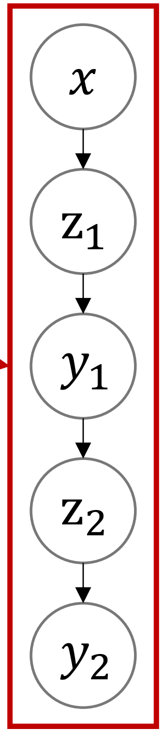
Background

- **Directed Graphical Model**
  - Multiple input, target as well as **latent** variables



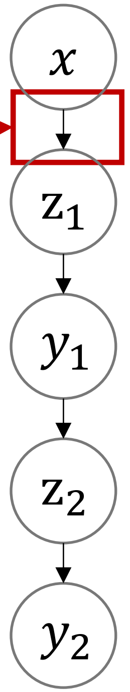
Background

- **Directed Graphical Model**
  - Multiple input, target as well as latent variables
  - Complex dependence structure



Background

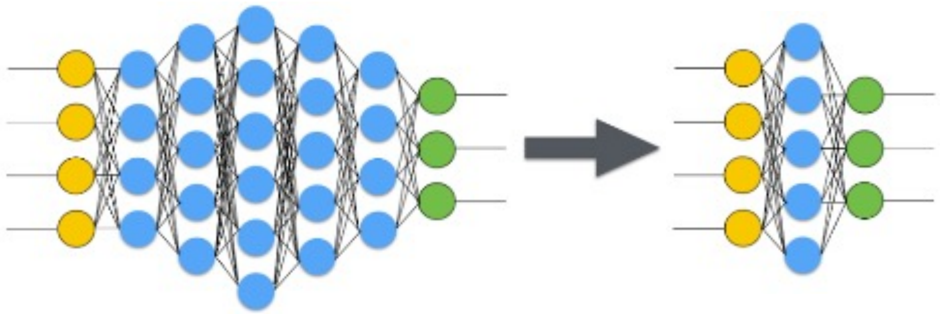
- **Directed Graphical Model**
  - Multiple input, target as well as latent variables
  - Complex dependence structure
  - Parameterized by **Deep Neural Networks**



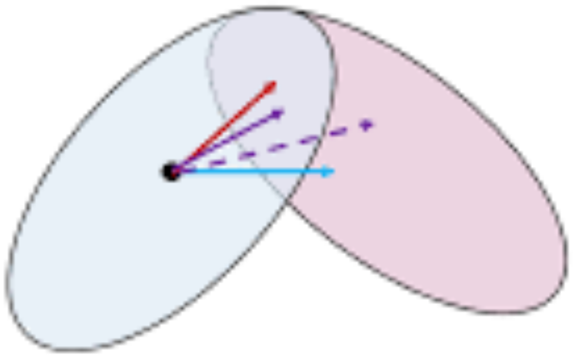
Background

- **Knowledge Distillation (KD)**
  - Transferring knowledge from a teacher to a student model
  - Applications:

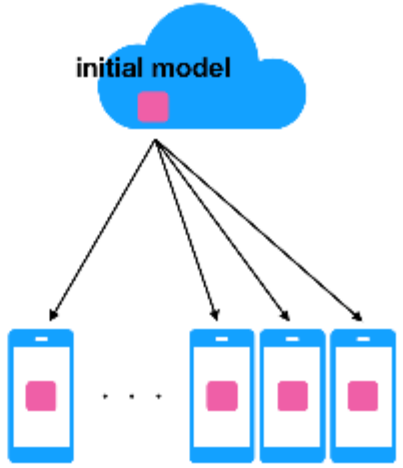
Model Compression



Continual learning



Federated Learning



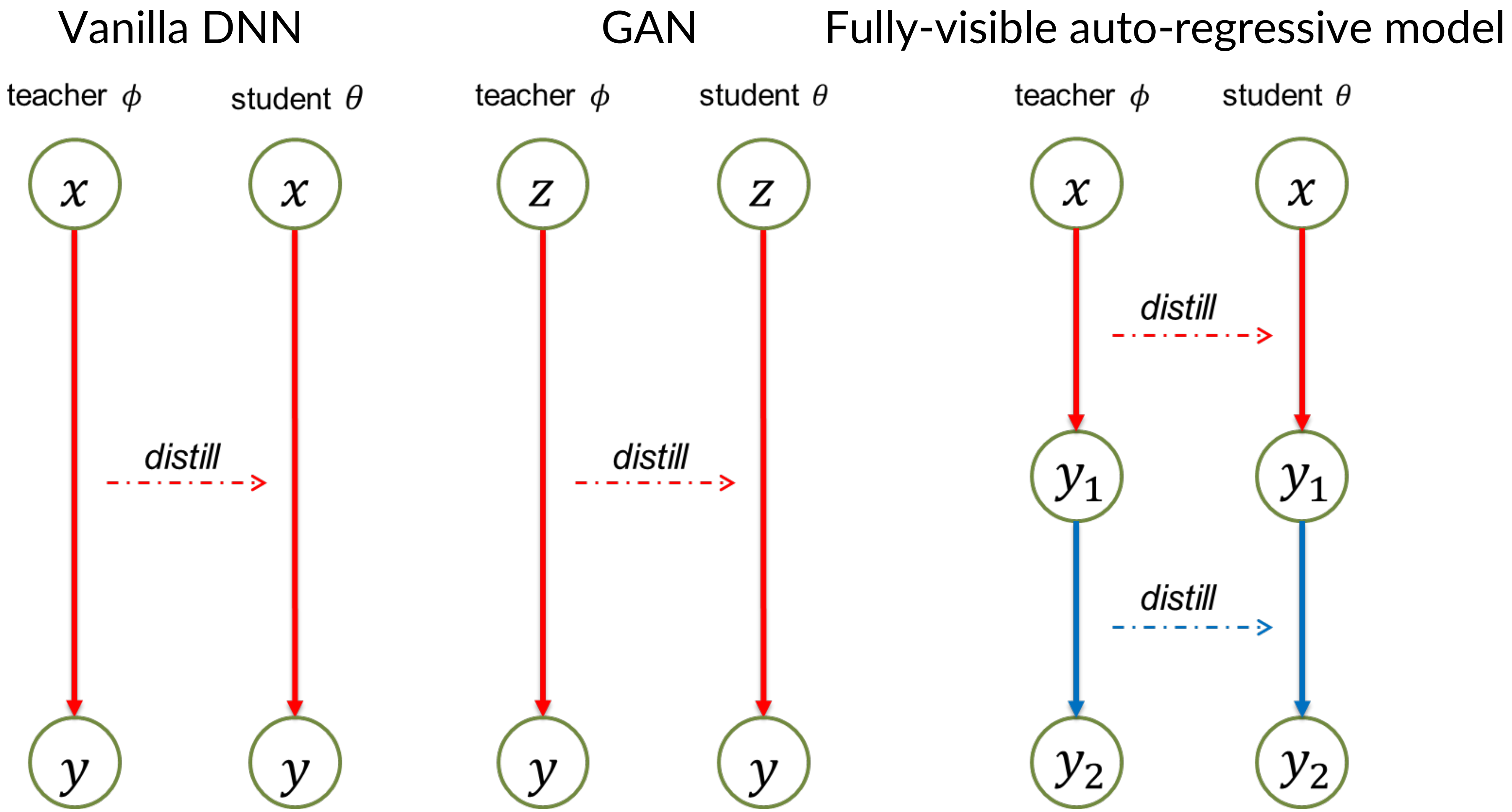
Motivation

- **Motivation**: all existing KD methods were applicable to **limited, specific** types of DGMs exclusively.

		#input variable x	#latent variable z	#target variable y
KD	for vanilla DNN	1	0	1/many
	for GAN	0	1	1
	for fully-visible auto-regressive model	0/1/many	0	many

Motivation

- **Motivation**: all existing KD methods were applicable to **limited, specific** types of DGMs exclusively.



Challenge

- **Goal**: Propose a unified framework enabling KD for general DGMs.
- **Vanilla KD loss:**

$$\mathcal{L}_{kd} = \mathbb{E}_{p_{data}(\boldsymbol{x})}\left[d(p_\phi(\boldsymbol{y}|\boldsymbol{x}), p_\theta(\boldsymbol{y}|\boldsymbol{x}))\right]$$

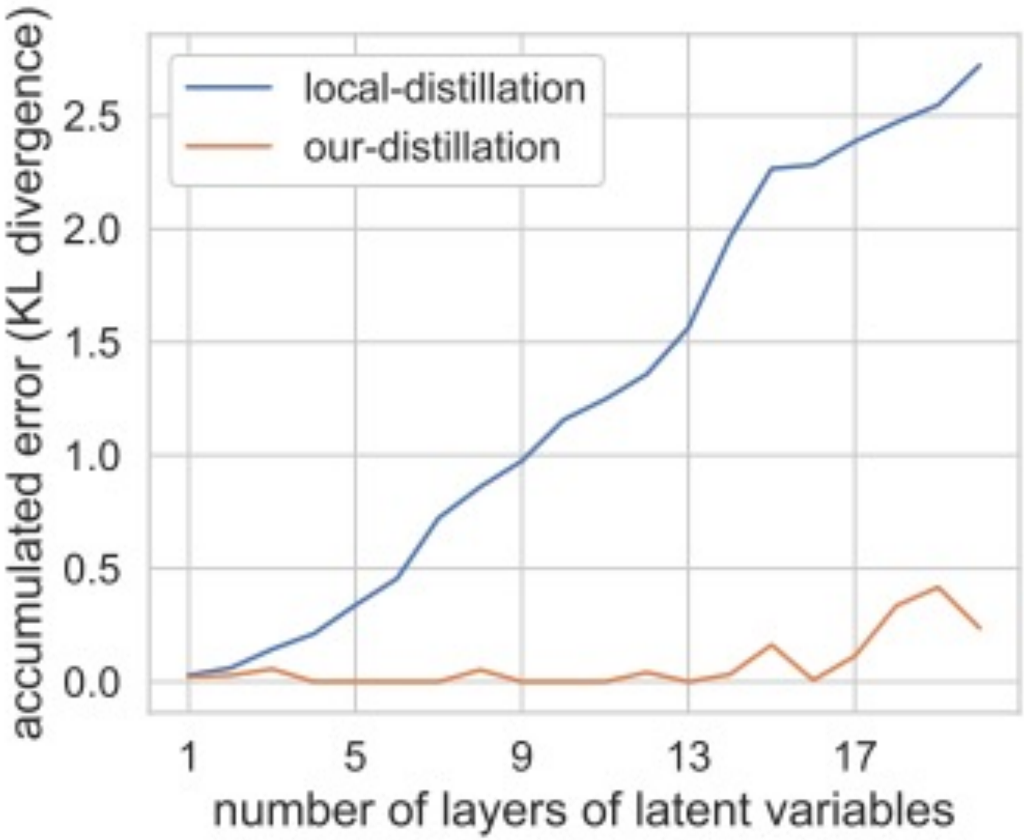
- **Naïve Method 1: Marginalized Distillation:** marginalize all latent variables

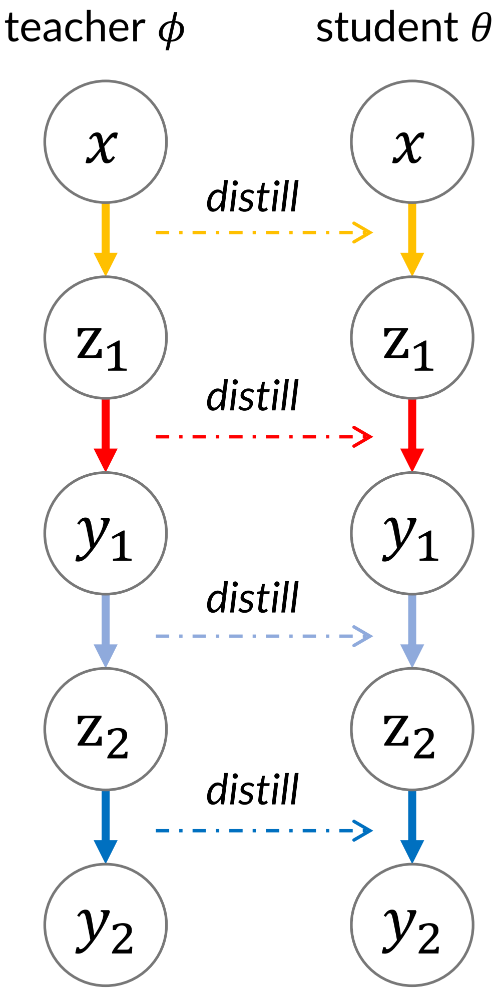
$$p(\boldsymbol{y}|\boldsymbol{x}) = \int p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x})\mathrm{d}\boldsymbol{z}.$$

  - Generally Intractable

Challenge

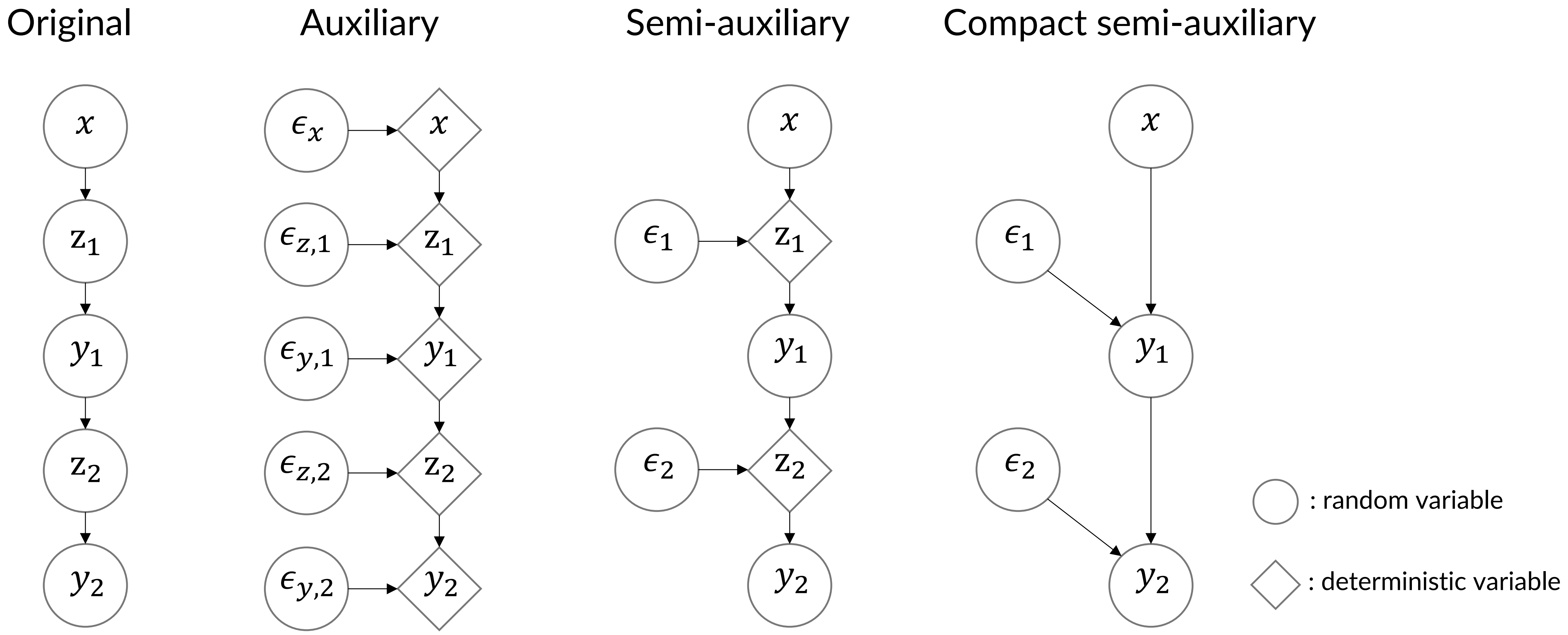
- **Naïve Method 2: Local Distillation**: apply distillation in a layer-wise manner
  - Imitation error accumulation





Method

- **Semi-auxiliary Form:** reparametrize all latent variables

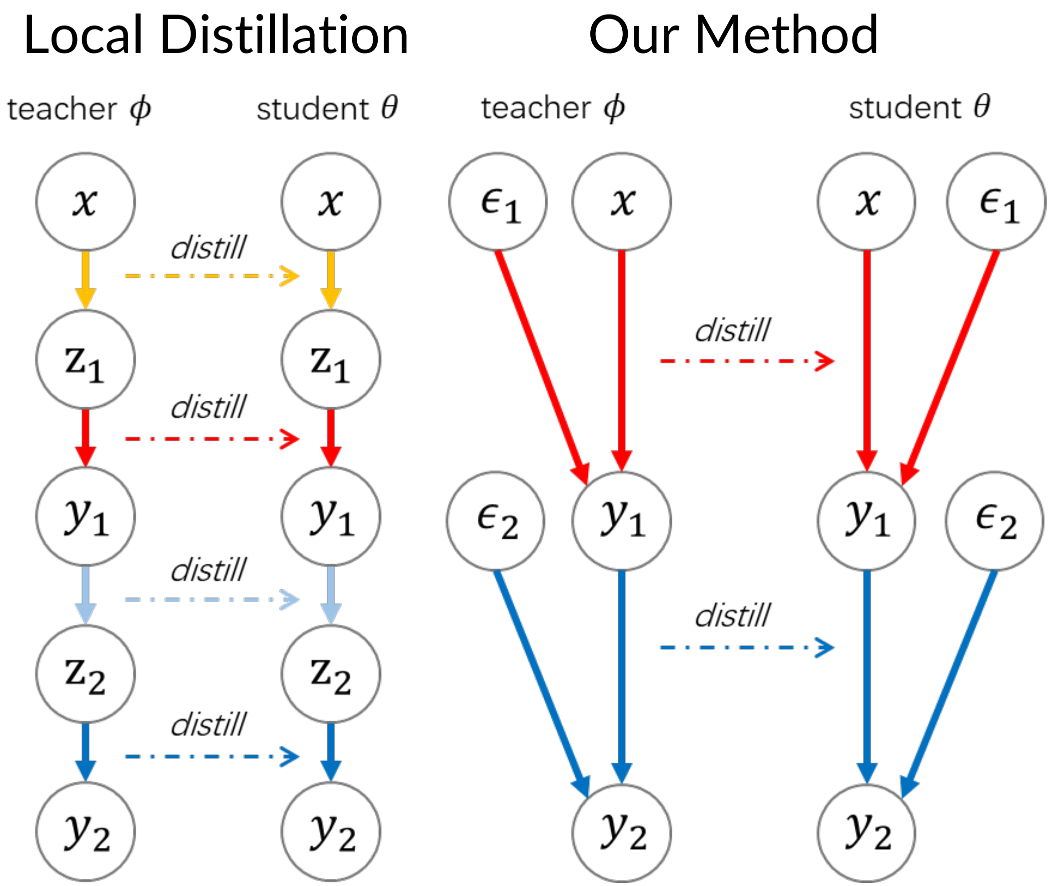


Method

- **Surrogate Distillation Loss**

$$\mathcal{L}_{sd} = \mathbb{E}_{p_\phi(\epsilon)p_{data}(\boldsymbol{x})}\left[d(p_\phi(\boldsymbol{y}|\epsilon,\boldsymbol{x}), p_\theta(\boldsymbol{y}|\epsilon,\boldsymbol{x}))\right],$$

  - **Upper bound** of vanilla KD loss
  - Compared with marginalized distillation: **tractability**
  - Compared with local distillation: **shallowness**

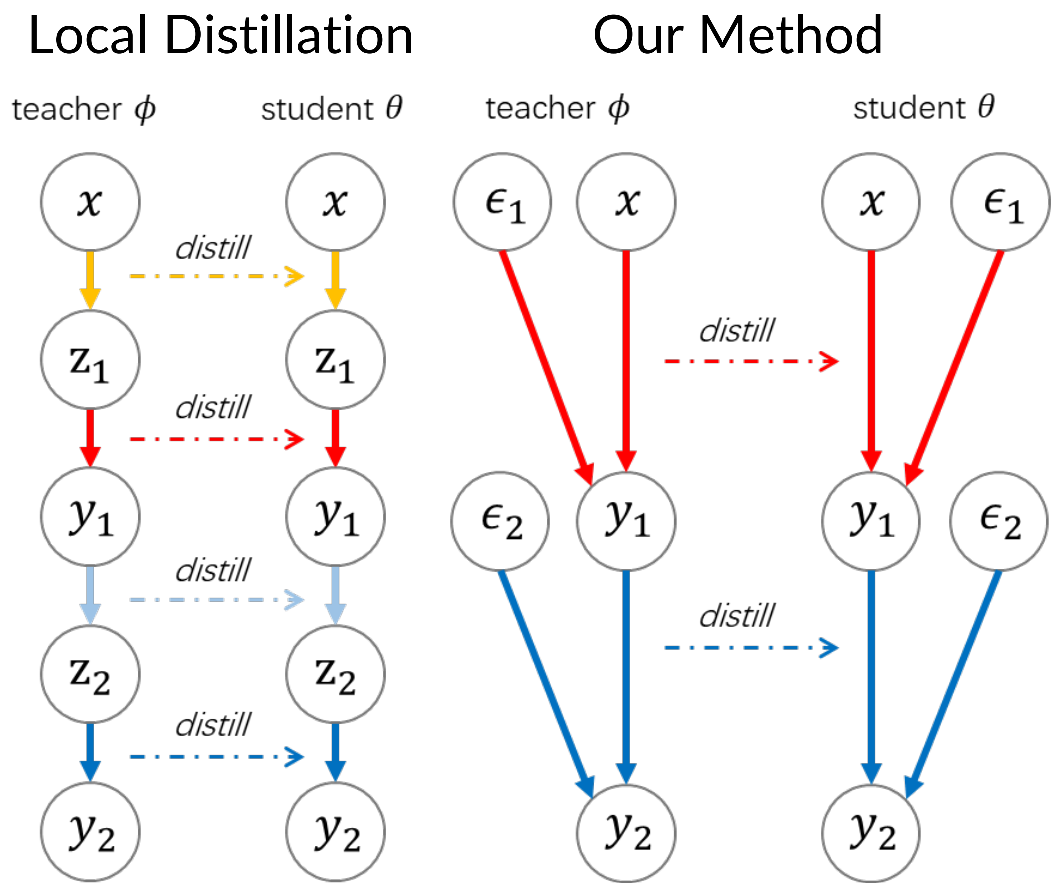


Method

- **Surrogate Distillation Loss**

$$\mathcal{L}_{sd} = \mathbb{E}_{\boxed{p_\phi(\epsilon)} p_{data}(\boldsymbol{x})} \left[d(p_\phi(\boldsymbol{y}|\boxed{\epsilon}, \boldsymbol{x}), p_\theta(\boldsymbol{y}|\boxed{\epsilon}, \boldsymbol{x})) \right],$$

  - Upper bound of vanilla KD loss
  - Compared with marginalized distillation: tractability
  - Compared with local distillation: shallowness



Method

- **Surrogate Distillation Loss**
  - **Is unable to back-propagate** when there is discrete latent variable
  - Might be **hard to optimize** when structure is deep

- **Latent Distillation Loss**: penalize dissimilarity of latent variables

$$\mathcal{L}_{z,i} = \mathbb{E}_{p_\phi(\epsilon)p_{data}(\boldsymbol{x})}\left[d(p_\phi(\boldsymbol{z}_i|\boldsymbol{\epsilon}_{<i}, \boldsymbol{x}), p_\theta(\boldsymbol{z}_i|\boldsymbol{\epsilon}_{<i}, \boldsymbol{x}))\right]$$

- **Our Final Loss:**

$$\mathcal{L}_{our} = \mathcal{L}_{sd} + \lambda \sum_i \mathcal{L}_{z,i},$$

Method

- **Surrogate Distillation Loss**
  - **Is unable to back-propagate** when there is discrete latent variable
  - Might be **hard to optimize** when structure is deep

- **Latent Distillation Loss**: penalize dissimilarity of latent variables

$$\mathcal{L}_{z,i} = \mathbb{E}_{\boxed{p_\phi(\epsilon)} p_{data}(\boldsymbol{x})} \left[d(p_\phi(\boldsymbol{z}_i | \boxed{\boldsymbol{\epsilon}_{<i}}, \boldsymbol{x}), p_\theta(\boldsymbol{z}_i | \boxed{\boldsymbol{\epsilon}_{<i}}, \boldsymbol{x})) \right]$$

- **Our Final Loss:**

$$\mathcal{L}_{our} = \mathcal{L}_{sd} + \lambda \sum_i \mathcal{L}_{z,i},$$

Method

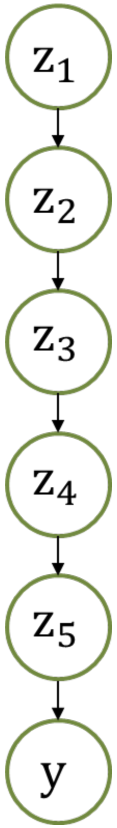
- **Fast Implementation**: Requires only an **ordinary forward pass** of teacher and student model to calculate the loss

- Our method is a **proper generalization** of:
  - **Vanilla KD and Sequence-Level KD:**
    when no latent variable
  - **Feature based KD:**
    when choose Wasserstein Distance in Latent Distillation Loss
  - **GAN distillation:**
    when no input variable & choose Wasserstein Distance in Latent Distillation Loss

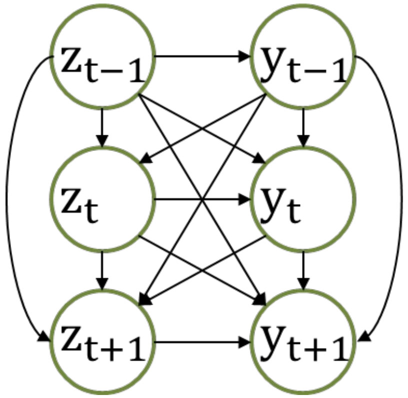
Evaluation

- **Evaluation Results:** Our method showed better performance on

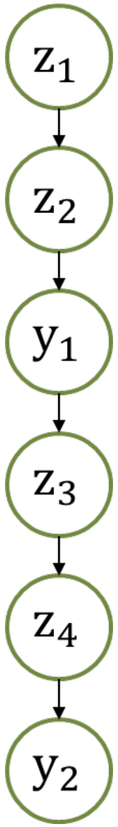
Data-free VAE compression



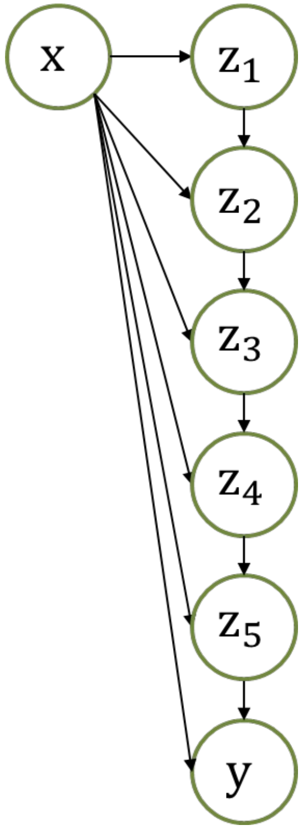
Data-free VRNN compression



Data-free HM compression



VAE continual learning



Evaluation

- **Results of Data-free VAE Compression**

dataset	method	#param	FID (↓)	EMD (↓)	MMD (↓)	1NN (↓)	FID-T (↓)	EMD-T (↓)	MMD-T (↓)	1NN-T (↓)
CelebA	teacher	6.60M	4.95	8.54	0.24	0.89	-	-	-	-
	our	0.44M	**5.38 ± 0.10**	**8.77 ± 0.06**	**0.27 ± 0.01**	**0.92 ± 0.01**	**0.019 ± 0.002**	**6.48 ± 0.04**	**0.12 ± 0.00**	**0.17 ± 0.01**
	local	0.44M	6.23 ± 0.17	9.25 ± 0.11	0.33 ± 0.01	0.95 ± 0.00	0.052 ± 0.006	8.32 ± 0.14	0.26 ± 0.02	0.82 ± 0.02
	scratch	0.44M	6.10 ± 0.31	9.08 ± 0.16	0.33 ± 0.02	0.95 ± 0.01	0.052 ± 0.016	8.34 ± 0.36	0.26 ± 0.05	0.82 ± 0.05
	our	0.12M	**5.96 ± 0.12**	**9.06 ± 0.09**	**0.31 ± 0.01**	**0.95 ± 0.00**	**0.036 ± 0.005**	**8.04 ± 0.11**	**0.23 ± 0.01**	**0.79 ± 0.02**
	local	0.12M	8.95 ± 0.19	11.24 ± 0.16	0.50 ± 0.01	0.99 ± 0.00	0.157 ± 0.018	10.82 ± 0.17	0.47 ± 0.01	0.99 ± 0.00
	scratch	0.12M	8.18 ± 0.15	10.50 ± 0.14	0.45 ± 0.01	0.99 ± 0.00	0.095 ± 0.007	9.98 ± 0.03	0.43 ± 0.00	0.97 ± 0.00
	our	0.04M	**8.20 ± 0.12**	**10.66 ± 0.12**	**0.45 ± 0.01**	**0.99 ± 0.00**	**0.069 ± 0.004**	**9.91 ± 0.06**	**0.40 ± 0.00**	**0.98 ± 0.00**
	local	0.04M	11.08 ± 0.27	12.79 ± 0.22	0.62 ± 0.01	1.00 ± 0.00	0.139 ± 0.015	12.80 ± 0.28	0.64 ± 0.01	1.00 ± 0.00
	scratch	0.04M	9.57 ± 0.14	11.46 ± 0.11	0.55 ± 0.01	1.00 ± 0.00	0.093 ± 0.004	11.19 ± 0.13	0.56 ± 0.02	1.00 ± 0.00
SVHN	teacher	5.39M	4.19	7.98	0.17	0.80	-	-	-	-
	our	0.10M	**4.38 ± 0.05**	**7.94 ± 0.04**	**0.19 ± 0.00**	**0.81 ± 0.00**	**0.028 ± 0.006**	**6.90 ± 0.05**	**0.14 ± 0.01**	**0.47 ± 0.02**
	local	0.10M	5.93 ± 0.50	8.66 ± 0.32	0.30 ± 0.04	0.95 ± 0.02	0.108 ± 0.019	9.29 ± 0.41	0.40 ± 0.04	0.98 ± 0.01
	scratch	0.10M	4.69 ± 0.16	8.04 ± 0.12	0.21 ± 0.01	0.85 ± 0.01	0.037 ± 0.006	7.86 ± 0.10	0.22 ± 0.01	0.80 ± 0.02
	our	0.03M	**4.81 ± 0.06**	**8.10 ± 0.03**	**0.22 ± 0.01**	**0.87 ± 0.01**	**0.031 ± 0.012**	**7.82 ± 0.08**	**0.23 ± 0.01**	**0.82 ± 0.03**
	local	0.03M	6.95 ± 0.35	9.40 ± 0.28	0.37 ± 0.02	0.98 ± 0.01	0.153 ± 0.017	10.39 ± 0.21	0.49 ± 0.02	1.00 ± 0.00
	scratch	0.03M	5.84 ± 0.32	8.62 ± 0.18	0.31 ± 0.02	0.92 ± 0.01	0.080 ± 0.009	9.10 ± 0.22	0.39 ± 0.03	0.95 ± 0.01
	our	0.01M	**6.71 ± 0.38**	**9.22 ± 0.31**	**0.36 ± 0.03**	**0.96 ± 0.01**	**0.055 ± 0.011**	**9.13 ± 0.11**	**0.37 ± 0.02**	**0.98 ± 0.00**
	local	0.01M	8.26 ± 0.37	10.40 ± 0.35	0.45 ± 0.02	1.00 ± 0.00	0.170 ± 0.015	11.25 ± 0.25	0.55 ± 0.01	1.00 ± 0.00
	scratch	0.01M	7.73 ± 0.28	9.95 ± 0.25	0.43 ± 0.01	0.99 ± 0.00	0.063 ± 0.015	10.22 ± 0.18	0.49 ± 0.01	0.99 ± 0.00
Cifar10	teacher	5.39M	4.63	7.57	0.25	0.89	-	-	-	-
	our	0.10M	**5.47 ± 0.23**	**8.16 ± 0.20**	**0.31 ± 0.02**	**0.92 ± 0.01**	**0.024 ± 0.006**	**6.29 ± 0.08**	**0.19 ± 0.01**	**0.48 ± 0.01**
	local	0.10M	6.22 ± 0.05	8.61 ± 0.06	0.37 ± 0.00	0.94 ± 0.00	0.036 ± 0.003	7.58 ± 0.08	0.31 ± 0.01	0.91 ± 0.01
	scratch	0.10M	6.19 ± 0.25	8.54 ± 0.15	0.38 ± 0.02	0.95 ± 0.01	0.034 ± 0.005	7.27 ± 0.12	0.28 ± 0.03	0.83 ± 0.03
	our	0.03M	**6.11 ± 0.16**	**8.59 ± 0.11**	**0.36 ± 0.01**	**0.95 ± 0.01**	**0.036 ± 0.013**	**7.29 ± 0.11**	**0.28 ± 0.01**	**0.82 ± 0.02**
	local	0.03M	7.55 ± 0.09	9.66 ± 0.05	0.45 ± 0.00	0.97 ± 0.00	0.052 ± 0.003	9.08 ± 0.07	0.44 ± 0.01	0.99 ± 0.00
	scratch	0.03M	6.74 ± 0.36	8.95 ± 0.27	0.42 ± 0.03	0.96 ± 0.01	0.037 ± 0.007	7.84 ± 0.29	0.35 ± 0.03	0.93 ± 0.02
	our	0.01M	**7.61 ± 0.91**	**9.65 ± 0.79**	**0.47 ± 0.05**	**0.98 ± 0.01**	**0.045 ± 0.016**	**8.48 ± 0.78**	**0.42 ± 0.05**	**0.96 ± 0.02**
	local	0.01M	10.53 ± 0.74	12.10 ± 0.71	0.64 ± 0.03	1.00 ± 0.00	0.085 ± 0.015	11.38 ± 0.69	0.62 ± 0.02	1.00 ± 0.00
	scratch	0.01M	10.17 ± 1.13	11.67 ± 1.01	0.64 ± 0.06	1.00 ± 0.00	0.067 ± 0.023	10.56 ± 0.99	0.61 ± 0.05	1.00 ± 0.00

Conclusion

- We proposed a unified KD framework for general DGMs, which
  - converted DGMs to the proposed semi-auxiliary form
  - combined 2 novel KD losses
  - generalized to multiple existing methods
  - applied to various types of DGMs and tasks

Thank you