Benchmarking Multimodal Variational Autoencoders: CdSprites+ Dataset and Toolkit Supplementary Material

Gabriela Sejnova Michal Vavrecka Karla Stepanova Czech Institute of Informatics, Robotics and Cybernetics Czech Technical University in Prague, Prague, Czech Republic gabriela.sejnova@cvut.cz

1 Overview

In Section 2, we provide additional details for the CdSprites+ dataset. In Section 3, we describe the technical information and detailed results for the experiments presented in the paper and Section 4 reports the reproduced results from original papers using our toolkit.

2 CdSprites+ dataset statistics

The presented version of the benchmark dataset comprises 5 different levels of difficulty, where each level varies in the number of included features (see Table 1 for their overview). The default size of the dataset is depicted in Table 1 for each level, although the user can easily generate a larger set of the data. The scripts for calculation of the cross- and joint- coherency accuracies use separate batches of testing data provided in the toolkit.

In all levels, the source of noise in the images is their random position and rotation - in levels 1-4, the shapes are located around the whole image with a variance of 25 pixels along both x and y axes. In Level 5, the shapes are shifted in random quadrants where their position also varies with a variance of 8 pixels. The positions are configured so that the whole shapes are always fitting the image. In Levels 2-5 where we vary the size, the proportion of the small objects to the big objects is 1:5.

You can also see a PCA visualization of the CdSprites+ Level 5 dataset in Fig. 4.

Table 1: Statistics of the CdSprites+ benchmark dataset. We show the number of train/validation samples and the number of various shapes, colours, object poses (meaning quadrants which are distinguished in captions) and backgrounds used in each difficulty level. The text captions only describe features that vary (e.g. in level 1, the text descriptions only include the shape name). The colours and backgrounds are all textured when they vary.

Level	Train Samples	Validation Samples	Shapes	Sizes	Colours	Positions	Backgrounds
1	67 500	7 500	3	1	1	1	1
2	108 000	12 000	3	2	1	1	1
3	270 000	30 000	3	2	5	1	1
4	540 000	60 000	3	2	5	4	1
5	864 000	96 000	3	2	5	4	2

Table 2: Training and inference times for each model trained on our CdSprites+ dataset. The models were trained for 150 epochs on Levels 1-2 and for 250 epochs on Levels 3-5, we thus show these times separately. We show the mean values over all seeds and different latent dimensionalities, the standard deviation is shown as \pm .

Model	Per epoch (s)	Per training (min) (Levels 1-2)	Per training (min) (Levels 3-5)	Inference time (s)
MMVAE	203 ± 2	525 ± 8	860 ± 10	152 ± 6
MVAE	150 ± 20	397 ± 25	645 ± 32	135 ± 19
MoPoE	127 ± 12	324 ± 9	542 ± 10	126 ± 11
DMVAE	200 ± 3	500 ± 6	841 ± 7	148 ± 8

2.1 Using character-wise embeddings

We choose to use character-wise embeddings for CdSprites+ rather than word embeddings. While this choice was made to increase the difficulty of the text modality in our dataset, character embeddings have been recently used also in several other works as this approach brings specific advantages.

Firstly, character-wise embedding does not require a pre-defined vocabulary of possible input words. This can be useful e.g. in incremental learning scenarios where the whole vocabulary is not known prior to training beginning. Secondly, the model can be tested for robustness after training by inputting sentences with misspelt words (e.g., "square" instead of "square") to see if the model can generate correct images. With word-level embedding, this is not possible as replacing entire words will change the feature or create a nonsensical query (e.g., "left square" instead of "blue square").

Please note that we expect the users to use the same encoder and decoder networks (i.e. character transformer networks) for the CdSprites+ benchmark to provide a restricted and fair comparison to other models. Should the users want to use CdSprites+ outside our toolkit for their custom evaluation, they can as well use word-level embeddings as we provide raw strings for the CdSprites+ text modality.

3 Benchmark study results

Here we provide the specific training configuration and hyperparameters used for the experiments on the CdSprites+ dataset as listed in the paper. We also report the detailed results for hyperparameter grid search in terms of the cross- and joint-generation accuracies.

3.1 Training configuration

All our experiments were trained with the GeForce GTX 1080 and NVIDIA Tesla V100 GPU cards, the mean computation times for training and inference are shown in Table 2. We used the Adam optimizer, the learning rate of $1e^{-4}$ and all experiments were repeated for 5 seeds (we report standard deviations for the results in the tables). We trained for 150 epochs for Levels 1 and 2 and for 250 epochs in the case of Levels 3-5. In the hyperparameter grid search, we varied the latent dimensionality (16, 24, 32) for all 5 dataset levels and the MVAE, MMVAE and MoPoE models. In the case of DMVAE, the latent dimensionality was different as there are private (modality-dependent) and shared latents. We thus chose different values for the comparison. We used a fixed value of 10 for both private latents and varied the shared latents with values 10, 16 and 24. In Tables 1-5, we show this as the total number of latent dimensions, i.e. 30 (10 shared and 2×10 private), 36 (16 shared and 2×10 private) and 46 (24 shared and 2×10 private).

We used the default training dataset size and validation split as reported in the statistics Table 1. In Tables 3, 4, 5, 6 and 7, we show results for the MVAE, MMVAE, DMVAE and MoPoE models and the compared latent dimensionalities. Standard deviations over 5 seeds are shown in brackets.

Table 3: Level 1 comparison of accuracies for the four evaluated models trained on our CdSprites+ dataset. *Strict* refers to percentage of completely correct samples (sample pairs in joint generation), *Feats* shows the average percentage of correct features (Level 1 has only 1 feature and *Feats* and *Strict* are thus the same) and *Letters* shows the mean percentage of correctly reconstructed letters.(*Dim*) is the latent space dimensionality.

Model (Dim)	Txt→Img	Txt→Img	Img→Txt	Img→Txt	Img→Txt	Joint	Joint
	Strict %	Feats %	Strict %	Feats %	Letters %	Strict %	Feats %
MMVAE (16-D)	47 (14)	N/A	64 (3)	N/A	88 (2)	17 (10)	N/A
MVAE (16-D)	52 (3)	N/A	63 (8)	N/A	86 (2)	5 (9)	N/A
DMVAE (30-D)	33 (4)	N/A	4 (5)	N/A	25 (2)	4 (6)	N/A
MoPoE (16-D)	33 (3)	N/A	10 (17)	N/A	26 (7)	16 (27)	N/A
MMVAE (24-D)	55 (15)	N/A	42 (3)	N/A	31 (12)	0 (0)	N/A
MVAE (24-D)	55 (4)	N/A	61 (3)	N/A	82 (1)	3 (2)	N/A
DMVAE (36-D)	36 (1)	N/A	3 (3)	N/A	21 (2)	9 (13)	N/A
MoPoE (24-D)	35 (3)	N/A	4 (2)	N/A	24 (6)	1 (1)	N/A
MMVAE (32-D)	48 (3)	N/A	36 (2)	N/A	26 (2)	0 (0)	N/A
MVAE (32-D)	53 (5)	N/A	60 (2)	N/A	82 (2)	1 (1)	N/A
DMVAE (46-D)	34 (2)	N/A	3 (2)	N/A	20 (9)	0 (0)	N/A
MoPoE (32-D)	36 (5)	N/A	2 (1)	N/A	23 (7)	0 (0)	N/A

3.2 Used architecture

For all evaluated models, we used the standard ELBO loss function with the β parameter fixed to 1. For the MVAE Wu & Goodman (2018) model, we used the sub-sampling approach where the model is trained on all subsets of modalities (i.e. images only, text only and images+text). For the image encoder and decoder, we used 4 fully connected layers with ReLU activations. In the case of the text, we used a Transformer network with 8 layers, 2 attention heads, 1024 hidden features and a dropout of 0.1.

3.3 Evaluation metrics

After training, we used the script for automated evaluation (provided in our toolkit) to compute the cross- and joint-coherency of the models. For cross-coherency, we generated a 10000-sample test dataset using the dataset generator and used first the images, and then captions as input to the model to reconstruct the missing modality. For joint coherency, we generated 1000 traversal samples over each dimension of the latent space (i.e. 32000 samples for a 32-D latent space) and fed these latent vectors into the models to reconstruct both captions and images.

For both the cross- and joint-coherencies, we report the following metrics: *Strict, Feat*, and *Letters* to provide more information on what the models are capable to do. In the first (*Strict*) metrics, we considered the text sample as accurate only if all letters in the description were 100 % accurate, i.e. we did not tolerate any noise. For the image outputs, we considered the images as correct only if all the attributes for the given difficulty level could be detected using our pre-trained classifiers (i.e. correct classification for the shape, colour, size, position or background). For joint coherency, we considered the generated pair as correct only when both the image and captions fulfilled these criteria and were semantically matching.

For the feature-level metrics, we calculated the percentage of correctly reconstructed/generated features (e.g. whole words or image attributes such as shape) and reported the mean percentage of correct features per sample. For the image-caption cross-generation accuracy, we also calculated the average percentage of correct letters per output sample.

In the following section, we report the mean accuracies for both cross- and joint-coherency - these numbers describe the proportion of the correct outputs to all outputs.

Table 4: Level 2 comparison of accuracies for the 4 models trained on our CdSprites+ dataset. *Strict* refers to the percentage of completely correct samples (sample pairs in joint generation), *Feats* shows the average percentage of correct features (Level 2 has 2 features) and *Letters* shows the mean percentage of correctly reconstructed letters.(*Dim*) is the latent space dimensionality.

Model (Dim)	Txt→Img	Txt→Img	Img→Txt	Img→Txt	Img→Txt	Joint	Joint
	Strict %	Feats %	Strict %	Feats %	Letters %	Strict %	Feats %
MMVAE (16-D)	18 (4)	0.8 (0.1)/2	41 (20)	1.4 (0.2)/2	85 (4)	3 (3)	0.6 (0.1)/2
MVAE (16-D)	16 (1)	0.8 (0.0)/2	55 (27)	1.5 (0.3)/2	91 (6)	1 (1)	0.3 (0.3)/2
DMVAE (30-D)	15 (2)	0.8 (0.0)/2	4 (1)	0.4 (0.0)/2	30 (2)	0 (0)	0.2 (0.1)/2
MoPoE (16-D)	10 (3)	0.8 (0.0)/2	8 (7)	0.7 (0.1)/2	40 (4)	1 (1)	0.2 (0.1)/2
MMVAE (24-D)	17 (5)	0.4 (0.0)/2	16 (0)	0.4 (0.0)/2	40 (2)	1 (1)	0.2 (0.0)/2
MVAE (24-D)	16 (3)	0.4 (0.0)/2	52 (9)	0.8 (0.0)/2	86 (1)	5 (6)	0.3 (0.0)/2
DMVAE (36-D)	18 (2)	0.9 (0.0)/2	5 (1)	0.4 (0.0)/2	24 (1)	0 (0)	0.2 (0.2)/2
MoPoE (24-D)	8 (3)	0.8 (0.0)/2	13 (3)	0.8 (0.1)/2	35 (3)	1 (1)	0.5 (0.1)/2
MMVAE (32-D)	17 (1)	0.4 (0.0)/2	16 (0)	0.5 (0.0)/2	43 (2)	0 (0)	0.1 (0.0)/2
MVAE (32-D)	16 (4)	0.8 (0.1)/2	40 (13)	1.8 (0.1)/2	87 (1)	11 (9)	0.8 (0.0)/2
DMVAE (46-D)	17 (1)	0.8 (0.0)/2	3 (1)	0.4 (0.1)/2	24 (2)	0 (0)	0.1 (0.1)/2
MoPoE (32-D)	7 (2)	0.8 (0.0)/2	10 (8)	0.8 (0.1)/2	33 (1)	0 (0)	0.3 (0.2)/2

Table 5: Level 3 comparison of accuracies for the 4 models trained on our CdSprites+ dataset. *Strict* refers to the percentage of completely correct samples (sample pairs in joint generation), *Feats* shows the average percentage of correct features (Level 3 has 3 features) and *Letters* shows the mean percentage of correctly reconstructed letters.(*Dim*) is the latent space dimensionality.

Model (Dim)	Txt→Img	Txt→Img	Img→Txt	Img→Txt	Img→Txt	Joint	Joint
	Strict %	Feats %	Strict %	Feats %	Letters %	Strict %	Feats %
MMVAE (16-D)	6 (2)	1.2 (0.2)/3	2 (3)	0.6 (0.2)/3	31 (5)	0 (0)	0.4 (0.1)/3
MVAE (16-D)	6 (0)	1.3 (0.0)/3	22 (12)	2.1 (0.1)/3	85 (3)	0 (0)	0.5 (0.1)/3
DMVAE (30-D)	4 (0)	1.2 (0.0)/3	0 (0)	0.4 (0.1)/3	22 (2)	1 (1)	0.5 (0.1)/3
MoPoE (16-D)	6 (1)	1.6 (0.0)/3	0 (0)	0.1 (0.1)/3	21 (5)	0 (0)	0.0 (0.0)/3
MMVAE (24-D)	4 (3)	1.2 (0.3)/3	1 (1)	0.8 (0.2)/3	31 (6)	0 (0)	0.3 (0.0)/3
MVAE (24-D)	7 (1)	1.3 (0.0)/3	45 (3)	2.4 (0.0)/3	91 (1)	0 (0)	0.6 (0.0)/3
DMVAE (36-D)	3 (2)	1.1 (0.1)/3	0 (0)	0.2 (0.0)/3	18 (1)	0 (0)	0.1 (0.0)/3
MoPoE (24-D)	7 (4)	1.3 (0.1)/3	0 (0)	0.7 (0.1)/3	32 (0)	0 (0)	1.1 (0.1)/3
MMVAE (32-D)	5 (3)	1.1 (0.1)/3	1 (1)	0.6 (0.1)/3	28 (2)	0 (0)	0.0 (0.0)/3
MVAE (32-D)	8 (2)	1.3 (0.0)/3	59 (4)	2.5 (0.1)/3	93 (1)	0 (0)	0.5 (0.1)/3
DMVAE (46-D)	5 (1)	1.1 (0.0)/3	0 (0)	0.1 (0.1)/3	15 (1)	0 (0)	0.1 (0.0)/3
MoPoE (32-D)	8 (2)	1.5 (0.1)/3	0 (1)	0.6 (0.1)/3	28 (1)	0 (0)	0.5 (0.2)/3

Table 6: Level 4 comparison of accuracies for the 4 models trained on our CdSprites+ dataset. *Strict* refers to the percentage of completely correct samples (sample pairs in joint generation), *Feats* shows the average percentage of correct features (Level 4 has only 4 features) and *Letters* shows the mean percentage of correctly reconstructed letters.(*Dim*) is the latent space dimensionality.

Model (Dim)	Txt→Img	Txt→Img	Img→Txt	Img→Txt	Img→Txt	Joint	Joint
	Strict %	Feats %	Strict %	Feats %	Letters %	Strict %	Feats %
MMVAE (16-D)	2 (0)	1.6 (0.2)/4	0 (0)	0.4 (0.4)/4	15 (3)	0 (0)	0.1 (0.1)/4
MVAE (16-D)	0 (0)	1.3 (0.0)/4	0 (0)	0.2 (0.3)/4	16 (5)	0 (0)	0.3 (0.6)/4
DMVAE (30-D)	1 (1)	1.4 (0.0)/4	0 (0)	0.5 (0.1)/4	18 (1)	0 (0)	0.5 (0.1)/4
MoPoE (16-D)	3 (1)	1.6 (0.2)/4	0 (0)	0.5 (0.1)/4	16 (3)	0 (0)	0.1 (0.1)/4
MMVAE (24-D)	3 (3)	1.7 (0.4)/4	1 (2)	0.7 (0.4)/4	27 (9)	0 (0)	0.5 (0.2)/4
MVAE (24-D)	4 (1)	1.2 (0.1)/4	0 (1)	2.4 (0.0)/4	14 (1)	0 (0)	0.2 (0.1)/4
DMVAE (36-D)	0 (1)	1.3 (0.0)/4	0 (0)	0.2 (0.0)/4	14 (1)	0 (0)	0.2 (0.0)/4
MoPoE (24-D)	2 (1)	1.4 (0.0)/4	0 (0)	0.7 (0.1)/4	21 (3)	0 (0)	0.1 (0.2)/4
MMVAE (32-D)	1 (1)	1.6 (0.0)/4	0 (0)	0.9 (0.0)/4	21 (0)	0 (0)	0.2 (0.0)/4
MVAE (32-D)	2 (1)	1.1 (0.1)/4	0 (1)	1.1 (0.0)/4	12 (3)	0 (0)	0.4 (0.2)/4
DMVAE (46-D)	1 (1)	1.2 (0.0)/4	0 (0)	0.1 (0.0)/4	14 (1)	0 (0)	0.1 (0.0)/4
MoPoE (32-D)	4 (0)	1.7 (0.1)/4	0 (0)	0.5 (0.3)/4	20 (3)	0 (0)	0.2 (0.2)/4

Table 7: Level 5 comparison of accuracies for the 4 models trained on our CdSprites+ dataset. *Strict* refers to the percentage of completely correct samples (sample pairs in joint generation), *Feats* shows the average percentage of correct features (Level 5 has 5 features) and *Letters* shows the mean percentage of correctly reconstructed letters. (*Dim*) is the latent space dimensionality.

Model (Dim)	Txt→Img	Txt→Img	Img→Txt	Img→Txt	Img→Txt	Joint	Joint
	Strict %	Feats %	Strict %	Feats %	Letters %	Strict %	Feats %
MMVAE (16-D)	0 (0)	1.8 (0.0)/5	0 (0)	0.4 (0.2)/5	16 (0)	0 (0)	0.7 (0.4)/5
MVAE (16-D)	0 (0)	1.8 (0.0)/5	0 (0)	0.6 (0.0)/5	27 (1)	0 (0)	0.2 (0.2)/5
DMVAE (30-D)	0 (0)	1.8 (0.0)/5	0 (0)	0.6 (0.1)/5	18 (2)	0 (0)	0.7 (0.1)/5
MoPoE (16-D)	0 (0)	1.8 (0.0)/5	0 (0)	0.3 (0.2)/5	15 (1)	0 (0)	0.5 (0.7)/5
MMVAE (24-D) MVAE (24-D) DMVAE (36-D) MoPoE (24-D)	0 (0) 0 (0) 1 (0) 0 (0)	1.8 (0.0)/5 1.8 (0.0)/5 1.8 (0.0)/5 1.8 (0.0)/5 1.8 (0.0)/5	0 (0) 0 (0) 0 (0) 0 (0) 0 (0)	0.6 (0.1)/5 0.6 (0.0)/5 0.6 (0.1)/5 0.7 (0.0)/5	17 (2) 25 (3) 14 (0) 17 (1)	0 (0) 0 (0) 0 (0) 0 (0) 0 (0)	0.5 (0.1)/5 0.3 (0.0)/5 0.5 (0.1)/5 1.0 (0.0)/5
MMVAE (32-D)	0 (0)	1.8 (0.0)/5	0 (0)	0.4 (0.1)/5	15 (0)	0 (0)	0.5 (0.4)/5
MVAE (46-D)	0 (0)	1.8 (0.0)/5	0 (0)	0.6 (0.1)/5	24 (2)	0 (0)	0.6 (0.1)/5
DMVAE (32-D)	0 (0)	1.8 (0.0)/5	0 (0)	0.4 (0.1)/5	14 (1)	0 (0)	0.4 (0.1)/5
MoPoE (32-D)	0 (0)	1.8 (0.0)/5	0 (0)	0.7 (0.3)/5	17 (2)	0 (0)	1.1 (0.1)/5



Figure 1: Results for the MVAE and MMVAE models trained on the MNIST-SVHN dataset using our toolkit. For MMVAE, we used the DREG objective as proposed by the authors, MVAE was trained with ELBO. We used the encoder and decoder networks from the original implementations. The top figures are traversals for each modality, below we show cross-generated samples. The bottom figures are T-SNE visualizations of the latent space - please note that for MVAE we show samples from the single joint posterior, while for MMVAE we show samples for both modality-specific distributions.



Figure 2: T-SNE visualizations for the MVAE model's (16-D) joint latent space trained on CdSprites+ Level 4. We show the latent space for each of the 4 features (size, shape, position and colour) individually.



Figure 3: T-SNE visualizations for the MMVAE model's (24-D) unimodal latent spaces trained on CdSprites+ level 4. We show the latent space for each of the 4 features (size, shape, position and colour) individually.



Figure 4: PCA calculated on the images in our CdSprites+ dataset, Level 5. We show a separate figure for each of the 5 features (size, shape, position and colour).



Figure 5: Image traversals for the MMVAE and MoPoE models for the CdSprites+ Levels 2 and 4. Each row is one out of 32 dimensions of the latent space, each column is the single sampled vector from the traversal range (-6,6).

					MVA	E									м	VVAE				
suc	umal pirk girpatbibbboorrihtsg	unal pirk qijipset.bbbcorrihit	smail pitk o qippertibilioooniihiti	small pink s oprpret biobocoliefhq	small pink I ggurre t bibboooliefftig	bmail bitk c qqurre a bitocellef#p	bial ben gaare at biceoleffig	bigbluenspare at boop lefttspapyyeee	big bluesopare « at bospplet#tabgopy	big beuesqqaro y ut tooppilftatbibqoqyye	sia lemolaurellipsse ce typpiliphrttilli	smalleyrceureolipsse a topprlightetillil	small precuseelipsea a 1500 ricerentiiti	small ynonweilias at beton righttillig	unall yroonweigipas at bottom riightshqill	smail yroenwelojpas a at bostor wrightthol	small yroenwosgarse h at bostoon ihttings.	small prin response at bostown	small yillo esquire at bostoonile	small yeller I ececarit t bootcemieft
š	small yellow applyce at bettern	smoll yrllinw rganse at bettern	small prokrowgpghream ethti	wygyre athleteellefs	small bink applieses biolocolilities	brief biologypureot	bigli son gesore attabaserithity	big blueroogineed. blopnightigblogoginee	kig ble e oppreset teoprightte	big binkergouine d. at two rightigbilitegrapy	biselliselwggelli oobo log93300	bigelilloonengunfeet b lightstillil	big siliconwsquarce at bettern rightstilling	big polewnweggarse at betten rightstilleo	small yreemweggipse at bottom rith/tithglo	small ynoenwelgipse at boetomwrightthry	small pelowellpseedd	ellose at bottom	small golwar elipse at beitum	small pplin elepsean betters nits
me	bialbiblerequirpost babb	brattilineqgipse bbsbecenthineqqipse	brcallbinecylinse i bábboo nihittagagapelo	brisal bineopoarrea E bistos righttoppogyere	Brind Lárk ggurreat Lásso Iefitigi	avat pirk quire atténocoletta	uraal yek weggine addooolieftii	unal yelow qqurre atabasolertza	unal yolee r. quaro at loso Physicscondo	small police square of blog	eellise atkortaorvi	lighter reliver atbodoorer	big ytoskeoffræ anticettori righttibe	en al bistratiposo at battors rightettai	enal yraes elojos at koton dobtet enals	unal yelow opins at bostammrig	unal yelen elipse allikationerigi	snatl jedav snatl jedav t elijse stakston righ	anal yelov elipse etabotton right	oraal yellow elipsostaa tars right
Ģ	small yellow heritaat3betoorwright	gental biologenminht	smail yrliowgorr at bostom rightgogob	smail prok wągarnea biblitoo riihłe	small pink gguireet bbbceoli/Rgc	bmail bitk g gyrreas bitsolif#3gg	bigli sonkoguareest bboollefttpbgggggroee	big blueneggaare afbbooplief#fbbgopge	big blue x sqquree ofbtopile/tatq	big blae ti separacut biopile/#358	small yookrahlvent abbottommightnesh	small prohruhopent i at bottern riighhittehin	small precrumoport at bestso mighhitshih	small yrecnesogurro at bettom rightetthge	arnall yreenweggupes at bottom riighttingh	anal yroenvolipse at bootomerrighttho	small yroenwollpsae c. at bottom righttuqilig	amail pricewellipse at bottom leghttoulili	small yr leonollipse at beidors lefhttsullill	small yfeiglipse at bittom icetttailli
Ţ.	big red segare at bottes right364bigt6ceseso	big helesopureal biotionrikhtätäköigik	big blue I square albhollars rigt	ligtues rigguine albiotoceurite	besol beskppp.recalb	(small bink approved biolocaliefTbg	send yeak weppere alleppietZag	sead yelowgggane Gogaleffagargegag	small yellew p. optyw atkogałieńszyc	small polew q. ophyseca loopole@appo	big yeyeloewsoopper of boetom might (364	sigl polewespape at bostore might bala	small poliniespipse at bestare entgetting	small yrlenwisspapse at beolonirwrightithg	small yreenweligese c. at boolomerright(Dig	small ynemwelgpae at kotiom rightithgha	small promovelipse a bil lediem rithridithejes	small preenwellipse at bettern lightstrepte	small years ellipse al bollars leveldEliphoad	small plukes ellipse at bettarn leftstitplanat
	qqirre atbaceorrights	qqqrrs atboscerright	unal brukn 1 gggrie atläbtosorright	unal br.k oppureraabbecoolieft	small bink gquire a bibbosolieffto	brnall pink y ggwre a bibboolleftsg	biall pinkopgareeatbobs	i bigli pirkoquarreatätid	t biglipinkagurre t bát.toolieftigbogggee	bigipininsegument bibot pp ee	sigil prixinweesopput at bottom left tt bi h	small protrienequipse at bottors light this h	snall greewoequpso wbloctow rightzbis	cidyazedynejd exert Ausocheidriboos	a unal yreeweigipte at bootomerrighteth	unal yreeswelopse at bottom rightsteph	unall yforweigipte Ir at bottan rithttttgllig	small poliniveologia at bottors rightereening	unal yeltsvecipu at batoonwight#####	anal yeinwelpse atboeton rightettille
	bmail rengeare at boppril hittoppopp	opineatbiogernihikt	small renicppporecality	small renkçeştirmaatbi	brool bink ggune a bibbooolieffig	brsall bitk garre t bbbocelieftsp	brial bisk ggirpeot bbbocolieffig	bial bink opport atbioetoermile/Rd	bigirbink guare bi abbotsomileftgibbgg	tigbrben goarre # #bboftomilefbbbgggs	smolweelseen gits opotto mithihituseeng	small soleven s ofts abstommentights	small pollow hquineat bettorimming	small yellow I square of bottom rightsthqualiq	small yritme squase at bottom nightthrapilig	small ycerwelipse at bostommrlightthep	sig pograkn e eligse attabottom rigi	tig volrvereolipseo abbottom rightsbook	big yr er wreeligaee atbietara lightheee	bipgproenwkeelipsee o leghtsbececee
	tog pellew eclipse at topile/Equips	equer at topertable	big yellow equipse anticipitetting	lightics op specasion	guine a bibeoff733g	swal birk papere a lisboculetta	smad reskipppine Materiaansni Bittippppe	snatres eggssartectoorm	snatiren II egypse atabattani ng	small ren # eqqpse attaction ingt b	Experit eclipsecast Educry right difficer	slatjana kollisse al boltani raphittai oso	small renkeelipse at bellarwerlightstat ave	swad prå ellpss at bottom räjtettaat et	ellpse at battom rithett winks	small yellow squpper at lowtonicity	small yellow p squarer at bottom right23-qelliq	snall pelewweeppse at bellars rightlingth	small yellowe agrine at battom righttamenting	small entrane envie at taltars right arweit
	big blue equipse at boopenightts	big bruenegojpso at btoprrighttopppore	big branneqquireeattà	t biglir verkeggir er uttibe	biał błok gęumo a bitzoliełtząc scencesc	umal birk p quiree bibocelieffts	gipper bbboolefets	ursall birk hgirt at boltoolie#sqo	unall bilk c heart at bottom leftspopopopoblobg	small bink heart at bottom lefttsbggggggbbbgg	unal ogl cecetoraa bottos ighttittopliee	smail leal eqgereeaatbeotoom r	small bluke i gguno at bottom riightttogilli	eqpineat/beotoon rit	arnal yreenweggipse at bottom rihittthdia	ursall yroeswelojps. 6 at boetarwrightthqo	sigil prescessilges at bottom righttbugo	big yyyeribreelipse at bottors righttbullip	big spier appelippepet bottom i	big poper popelippepsa bbitaren
	biglielink square t bbotao mightagagaree	biali pink square t bisetos might geggeneseses	bioli pink septore a biostoorwrig	brol pirk t oplare abbitsorrrig	brool pink grame a bibbooolief#cg	unal birk papare a bipooleffag	small binkpgorpe ath poleoftolloggopce	small benkgpginse at toplief#gbbgggggam	small ben e opfinse at tolleftspolio	small ren s telpte at peliceitsyste	small soliiligas 1 botum lihduilli	et beitarn riktisselet	small pillicilgose at beitsm righttyselee	small yracm elipson ethiottomerig	stal yremelipse at bootomerighting	smol yrcenwescasse i ot bottom ribhtthqua	small yrinw x sgarre at bettern rightstiwaaroog	siali pellen hiterr astisbottom ligh	biglycyclis Mast atbotice leghilfiasaaceu	big power bhent et botten Bellettreseets
	owait blue square at top nght gogggggeeeeeee	snud blue square at top rightsoppppreserve	small blue square at tap righttigggggeneenee	soare Miteoprophil	oral tra quare attabaorritta:	bial beringggavear366	big bruenegginnes/364	tog para equiras adabáticamentett	ling park I eqlipse attaction whet	lag paix In equipse adbibations left	personal statements of the second sec	stapppppppppppp at Earton inettents b pa	enations equipses at loticon letteration	onat pra equipo at botano igrezza auto	enal green elopse at bottom rightettheaulig	equips attentioning	enallyrdow eqlpse at bottars righttagollig	scal priceworkpo at bottors rightwelli	at batton lightpretti	soll licencello lices injustedii
	big jrfine ggipse abbotoennigi	big yrlenwopgaraeae bbotsorwikhtipogybel	big yrees opperventabletore rig	bigbrien eggareewitbb r iffhgggggerbbbrrb	bial bitk ggurreat bbboooleffs;	umal birk p gaure a biboolifftog	ggaare attopopile/top	smail bitk quare at topile/#tigg	small bink q opuano at top left5gogogogoececeoe	small bink spaare at top p lefttoppopoegoloeg	big yellico equirree tibeottom rig	tig yellin r eggarse at bottom rightstellik	big tollwrwogointe at bottore rightstaliol	small yrlerwoggipse at bertors riihirdtsepol	small yroenweigipee at bottom nightsthele	unal yroenwolopse at bostommrightthop	smail yroenwollipsaest q	strall prinweellisse wibbottomrrighttsquis	small polinweolipse at bettommight#ssaulu	unal plitweelipse at bottommightsamuu
	bigliellw wggorreos bog rillrifitoggygorreos	bial lenk wegenreas bepiriighti	breal link wyphrea bisperiighto	treal birk opurece biblocerights	brsol bink spurre a bibeo beltopppyermmi	prover a biblocellefty	small blek ggarreaetbitoooleTtg	small blick rgpurnsarblootoem#13	small bisk Ir oguareastitutionnile	small bink Filippuare albihotism left	bightlifebenhessent	abobition Ibertioolall	highelikoweelipse otdo leghtsolell	small yellow elipseatibilitionmig	stal ylewelipe ot boolonarighting	smol greenweggipse c at bottom riightthigga	small promisiparre a. at boltom rightthigan	small prinwsporre at beitare rightshas	small pilokicypypes at beitsenrigghittishis	produiting store
	tag pink square at top rightophilospapagapagap	big pink oquare at top rightoqbibioqoqqeeepp	e uityyzópodódosese	Ng bine sopure a boprinema	bial berkgggureez booprieftigggggeee	anal birk perfilioatada serriaça +	quarteatbotsoniet	unad brees quare abbetton ieft ce	unalignees square at bettom initiality at person	spane at bottom leases at bottom	small res eedusse a blottomi	snall redevedposeda bootsenie/tzzpreets	eoliooreaa boottamile	enal bruke Hipse at bostonvilte	and yneswellpse a bostonwrighthq	anut yneswespipse a it bottom righttinger	unall yellow is square at bettom rightstaggoolg	scall jelew square at betaonvriht	sigit yellow square t toorcrightnessiglig	ciglinyelcowl/sarr oppoligy/http://doi.org
	omail pink quirre a bibbooslief#10	opurre a bbbooolieffs:	bmall pirk p opurre a bbbosolieits:	breal birk p ggurre a bébooolieftso	broal birk grune a bibbooolieftto	brailbik garre a bibocelefta	ornall bink opprire a bibbooolieftig	penali bink opurre a bbboeoliefts:	breall birk > qp.me a bibbosoliefts	breal bink a ggunra a bibbooolieftop	Pige de la marca de la calega d	I PIQDODOTT DODODODON 1	at bedore rightshikk	small greikn i elgipse at bottom riighttthihas h	anal greeneeidpse at bottom riightstags	anal yelowelgipse at bostommrlightthop	smail yolow q. equisees/bootommini	small police elitosoarbootooprivat	small yllownellessadh	walliowellissillo
	big pinkesopurre at toprighttopbegopor	biglinenkspaper at transpittatioppapere	bigl ucelepopureeaa c lpringhtighingopprese	Moll penk opure of top-tel/Right	brod pirk gyme all dooleftigg	small birk a gavreet bitwoodeTty	sreal blek gjøpeet båbloo beetky	small blak oppet at bettom lefttoppgevelweet	small blak heart at bottom leftbioggy/stdeereg	small bluk heart at koltom left@dogg/stdmmb	small place. square at battors right@33abu	small proen sopure at bottom right@ukadku	small green sepure call belowning	small green expapse at bottom rahrdthigaolig	small yrecewelogos at bottom riighttituga	smol preesweigipse b of bostonwrightiting	small prinwordposed() prophilipping	elipse at kotion rapiditiopliq	small yellow elippise at bottom right@shighen	small yellow ellipse at bottom ithdthe lief
	tig jellew square atboognietht	oguare attoess lefts;	big yellow t opurrent bosooilerhop	big yrfinw qqurreat bbtoellfftopig	biał błok gęurre a bibbeoliełtzą	unal birk c quire a boboceleftra	quirree bibeconight	unal blak geppesabateorribh	unaliblak t opirperatbiotoomrig	small bink # galps# #3bostoersmigh	sigil yolina elqupse at bottom Beltzaargiske	kinall jedno elgipse at bettors lighttarqqiqq	small pelaw eigipse at bettom righttracellig	enal yellow eqgipse at bottan righttuqqliq	areally recoverages, at boots weight they	unal yroeswelgipse . at bottom rithittithgial	unaliblare Ili eqopsee atbéottom rij	small blue p eqoipse at bottom rightstolakt	small blueseqqiras at bartons nightraxies	unal ligniquporeaatto
	big pink ellisse øtsolleftsplobs	big pink englisee at olicettybbbpggroeece	big pinkneoplince abb olletttgbboggreek	tigitmen pgghee a bigeliettigbgggreee	biell pink ggunv a bibbeolifikkyp	small birk p gavre a bibloceliefftg	ggaore athlobicorwnigi	small blue I square atblottom righ	small blue v square at bettom righttqquee bbece	small blue square at bottom rightogs/ blue	big yilleoo ellippse bittoon riggtsbilleill	tig vilsoeneliguse a böttara righttbioli	big yellowonellpase t boettare righttbilool	big schwarswollpss at beotomarightitble	small yroenweigipae a at boolommelightthig	small yroenwelgipse c at bottom riightshigga	small yrfrweggupse a. at bottom ribhtthguu	small pilmosqume at beitare riktstingau	small pollwwsspurre at beidom rightthquu	small playersource at beiters right#shiwawa
	small yeloweppression articloppermit	wegene alloctowings	snal pek . węprzeadóścio ritir	senalt pek I sopreesabbétas tetta	seat pra gairrea lidooatetto	bial beraggyireestable	big inveneggarreaalbi	oppressi totos tetta	Big blue of optiment bidon Betts	ling Line it equipressillation liefT3	snalmere Herite E Latan Hittheoled	snall yrela eliniee al batton iefzzäheleef	seal preise equipre attoctava let	small yreene I spupreeatEduction rit	anal ynenweligae a boolonwrrightho	shad yneenwelopse a al bottom right thubu	small green in ellipse at bottom rith/tttaullig	seall promediposa Bittom rith/dtoillil	seal Minedpostab typ ritridaattii	sal bibelippetate oppit.dot3.dttl
	big bluk gewit at bbitsen ierffisjobbbgsgitbbbgsgi	big bluk quart at bibotsomileitz	big blue t opern atbibitsom lefthsbegagsbibbiott	big blueneggarreatbbbl Beftgbggggeebbbet	bial benkgggarreatbol	qquareat.bloociett	snall brek square etbipsprights	small yrekrwsquare at boppright#tequiper	unall yrfrwsgarreat tapprighttoggrarm	small poline ti qquirio attoppriihittiqq	big greekhentänty at bottommighittikee	big greethewronyn k at bottom rightitttoeekeht	big greinhlweggprr at bottom rithhttheekath	sial bluen spipre at bottom rightshiph th	elopse at bottom rightshiphealo	kapan at bottare rightstrafilig	small yollow ellipseat bottom right	small policy religion 1 toop rights.celli	ellosee #.topp pinemilli	unal elice elipsee tr oppopperhttillill
	big pink gesore at topprightted	big pink opparent topelihitig	big pink I sygureeest toperights	tigliveringpare atbits	broof pirksysparroot bbbsoliithtspagareeree	unal birk gysirees bibesolefts	small breen gagree bibboosleffts	smol green ggippe bibboscellefs	small proes o opinise bibliocosilieft e ce	small prom c opht # bbboox Heftpopp	senal lell ecerpte et cettrao lettet cecer	small blafw eleppee at beitsen Befittits oeld	small bloom ellipse et bottom lightstropliel	small yraumweigipse at beitommightetting	small yreemwelgipse a at bottom nightstagu	smoll ynoenwelogse is ot boetom riightthigg	sipi polewesquare a et boatem rightsbuuc	tig ysysticeceggige at blockers nightstaubo	big styriforecepsrarps rightstelebebg	big ppychbeirbbgungeb night30686866
	sead ree spare altidotion left IIII	sculare at bottars logitare at bottars loft2gogoge tottt	onal real opurerationsoletti	snat pilk opunn attestadettag	small pink qquirie a EditeocherTra	bial box g quire l bibliorittag	biglareen opgareentabl	gerreat blop ritrititio	Big Lower. 6. optimit & Doppmitter	big green e geart at bop rititztplat.gogyreeneen	List 4 million and a	logenearmine annotation an	additionwighting	tag Ettaweedpoor at Eartons rigg/VEEpp	dad ylako Ledipse at bettors riightttakelig	sound yellow square at bostaments	snallyelow is soparevettibution op	snat jetev germatteotrin n	skal yelle Net Elstan igtaabbere	unal Ille III E of an heree
	small bluk golpse abbbeceallefts	ursal bluk gpigse abbboosolieft	unuil bitk c opigse abbbooocleef	small bink opulpee bibboolieftto	small bink ggurre a bibbooolier#tg	brual bitk gaure a biboolifftage	doğur new több por liefts:	bigltruceqqqtreatbb	c big bedeegprr effbogerightfregogige	big red p hquir et exprilititizagggyvorn	small rink square authotions rrig	smail pitk I eqquiree atbottarenti	small pink programs at bottom righteethilteet	equire autoextoommi	small yreenweigipee g at bottom riihittthghe	unal yroenwolojpa atbiotomrightingo	small yollow a ellipse at bottom righttaglill	small police eligies at bottom lightstullill	bigleyelon elipso at bettorine ightsellill	big tycloswollpss t böttomm inftallill
	snal greenessooreer bobelief33ppgplite	smoll green square at betom tof#3ggggeloeb	send green square abbetomleft)	small brown upuare albéton liefitae	small blok separaeaobbbocelleft	bigil berikogarosotibi t	tig tick ggir antibitionrititig	big renkhqent abbatao rikhtagbapa	big rem p heart at beckerright	Mg ren e heart at bottomerights	olosison illike et	sigippelva illihoa al oottava lightitas oora	small ren Jellpsse of bottom rigofittissenere	small pick ellpsse at beitam riighittspepcee	smoll green ellipse at bettors ritight@readers	senol yellow square at beotynemia	small solere g square at beotomenen	small police bloort attitudite with	small yellow Meant att botto mbiddt dogoog	small yellon Meant att botte with 000 Minepop
	biglikavnikogurra atlabi	bigt unrikgquareead bibstaomitetbqqqqqeei	bial best-ppp results	geare atabataoniletz	grand blex ggurre a babacellertha	exal birk qquire a bibocriffitigi	ereal blek qquireea bépontightep	small blick opurrent top righting	unal bek q qqurr it tap rishthopogogoicraceo	snall bid egar at tap r rantesessonesses	unal yelow elipsee tEurzonieets	unat yelev elipseat battoniegi	snat jetev elipse at bottammely	enal yellow ellipse at bettars rightthosulag	anal yrinwelijposei	anad yneesweggjose at bottom rightetthua	small grans a qquirie at bottom riight dtttaabu	sigt hacequare attaction right/start	tig yrelew quart of bostom rightint/doceseese	big preena grant at bottannnigght ocecocoa
	unal ytiew gotpe attropie/toppe	small yreinwopoina alogaliaftopopologo	small yrok węzymeodospieł?t:	small brok sogirceoat ppplefftspie	small bink gquireea bbpopiletRgo	brial bitk g garreatbibliceolefti:	biall berkpguareextbbb rieftspgggrebbbbrt	bigl untrippure abblicecentightsgage	tigbisereqq.pre atbi	t big blue eggipse afbbattom rigt	big billisweetresee abbottom leettobsio	tig beliconvellince abbottom lefttbbsloe	bigleioloweig.rsee#1	at bootom rightshio	unal yreenwigps at bectommightthe	small yroenweigipse at bottom rightshigga	small processiopse at bottom rightthquu	small promotolopic at bothers rightfolling	small-green elipse at bottomnlight	smal grærelipse at bitsprlighttellen
	swall pirk grypper oblicationship	smol pirk opppressidebetoeffel	snal pak opurealtabloodef3	small pink o upurre a bibboochettio	small pink opume a babacellettop	bread blok g-geparte a bisbiosochettop	bialithenk segaroeaa biocellefitegoppennnn	bighteenin gosorecel beopertghttiggggermmi	big green r oppaare addesprograf	Ng green k geparee a bopperiddo	small yellu eqlips of battors liefEEEIlles	small police squire at boltom boltomades	small pollow squire at bottom BeTEX.aux.lig	small yellow squise at bellarmonth	small ynemweligiae Califolionenrightillig	smol yserweigtse a at konton rightlinge	big yrllwyrcellisse 8 at bostern rrighttol 90	Ng blee sepil rose attribution rogedtate ek	big blackgoare rbDoabrighthwrrighti	big ren I square 1400 Inghittarrithéséééééééé
	ereal bluk gqirper atopolefettiggg	uraal blak qolyse atopoieleitaga	unuil blik polipee at ppiefefaggg	unal bitk opurpresida polieritopi	small bink ggurre a bispaoliefhag	brial blik garre a biblioalieftig errenn	ball link qqqurevatbititions lift	bigliollow wqqireraabb right3qqarerberbb	t big yellow higerman bib/stometin	big yellow h hgert at blostos riitttaggggenereog	bigottiloorettisse aobb righttollill	atteoton rightsim	big yfellowretlisse withottom righttajioll	biglieyeloweeqipuxat	d used ythrweighte at bootonwrightetog	unal greeweigpse at bottom rightshohe	unal pitks . eggpsa at bottom right/tthgasub	small pinks sogpose at bottom righttothpueauth	unal pink suppus at bettomm righttomppusa	ereal pipk appipes at bottament
	small rom ellipse atbitsorroghts	smolines epigue obbiocerliefs	small renkgogina abbboollief#gggggoe	small pink opurpe a bibliocelleffing	small pink ggunne a bébecellefftig	braal bink gavre a bibocelief#g	biglirbenegagaree obbl	big bruenegogipee tbbbtoellief#.pgpgyree	big brace explore each	t tig biseneggipeestabo	o ostommetoghetol	prealppikgionigsse of bottomminggttpplikki	small pinkk ellipsse at bottom rightepiceore	small pinkn ellipsor többotsom nigt	small greenweiligse 4 at bootomminightsthe	small yellow square at bettern rith/dth-gapoog	small solice square at bottom right-stauceop	small policy hquine abbottom right	smalloellowebeerte at light##sueect	smallollewww.tmiot
	tig pelewippipor Mitoritare rightspepi	Liq yrthwyspinoe oddorfaonin rightysgof	lag ymenwggaare > ad Mdiaio rifrifiligigig Mddiait	lag grien oppareeatlidation rap	agune Laboratetta	swal lava p aparte a kidoodiffitoji	schall restoraginge attabulettaggggggeree	snatres equipse at toplet tagg	snatired o ellpse at toplettago	small red e ellipse at top tettopopopopopoeneg	snaboghwellposeat	al bollarenciphilase	snaturophen elpose at bettern righttowsed	snat yebwelposed3 right2spress	a seal ythwelipse at boolooming/025-p	snad ynenwegg.por y ôf kolten rithfittaska	Big yeyreen osparte a at botten right dtates	big yel www.cogareeo. Ediam riightit20oosa	Sparerpa beltap eeep	big privande heteenatkeerspree
	biglitbeneeligipes at bottom leftsgegge bb ee	bial beneegoupse at bottom lefttiggggee bierr	bial binecepuire abbitracentettrappore	breal birk guire attabaoaloftig	braal birk ggurre a bibbooolieftop	urnal pirk quire a bibocelleft q	unal pirk qquipeestatioocolefts	unal pitk eqlipseaabbboooleffts	unall pink. q. equipse attitiocoriee®	small pink z egojpos odzecep riefitaj	big grän hgotpri et bettorweilihtetteekse	sig yrlink soguprr at bottomm i ihrtzttoauet	siafropolensoppre anti rightzzasast	small green siggine at bettom rightttheeath	equipse at bottom rightstratigheath	small yellow ligipsa at bostsommigh	small yellow v ellipse at bottom right#xxglill	small police elipso at bottom rightteopili	small yellowwellipss at bertors nihttaagiiii	anal yelowwellpss at bittss rinttsspill
	big brownog.re at brogoleftsplogyes	big breen oppreset bosopileftte	big breen t opurre athlesselliefttb	highnues goare at bitcelleftglopgence	bial bink prome a bistoolleftster	small pink gyprress bibleoolieffs	small pink optigse obbbecerrightst	small pitk regilipse abbboosmithid	small pink ti opirpae abbistoreninė	small pick t giltose abbitrorrrithisp	big yeyrenen soutre of boetoernilef	highesreken square at bootsomilefitturus	stall yrecne separe at bootomenile	and premeroparts	small ynernweigense at boetenwerightthe	small ynoenwelogoe of bottom rightthops	small scenwellese at bottom rightshells	small groan alligner of bottom	small years ellipse at bottom	small reset evilpase of bottom

Figure 6: Text traversals for the MVAE and MMVAE models for the CdSprites+ Level 4. Each row is one out of 32 dimensions of the latent space, each column is the single sampled vector from the traversal range (-6,6). Note that we did not set the desired length of the text output, the model thus always generated the maximum number of characters.

3.4 Detailed Results

In Tables 3, 4, 5, 6 and 7, we show the comparison of the MVAE, MMVAE, DMVAE and MoPoE models on the 5 difficulty levels of the CdSprites+ dataset. Here we varied the latent dimensionality (16-D to 32-D) with the fixed batch size of 32. The values are the mean cross-generation and joint-generation accuracies over 5 seeds with the standard deviations listed in brackets. According to the *Strict* metrics (with zero noise tolerance, see Sec. 3.3), all models failed in both tasks at Levels 4 and 5. The *Feature* and *Letter* accuracies significantly decrease across levels as the complexity increases. You can see the T-SNE visualizations for the MVAE and MMVAE models trained on Level 4 in Figs. 2 and 3.

4 Verifying correctness of model implementation

To verify the correctness of our implementation for each model, we have reproduced selected experiments from the original papers using our toolkit. We provide both the original and our results below.

4.1 MMVAE

To verify that our implementation of the MMVAE (Shi et al., 2019) model is correct, we reproduced the experiments using the MNIST-SVHN dataset. We used the same model configuration and parameters as in the original report, i.e. the Mixture-of-Experts mixing with the DREG training objective, latent size 20 and 30 samples drawn from the joint posterior and the likelihood scaling for each modality was adjusted according to the varying dimensionalities. We used the same encoder and decoder architectures as in the original paper. After training, we calculated the joint- and cross-coherencies using the adapted original evaluation script (please see the original paper for the evaluation details). The results are shown in Table 8, the config files for reproducing the experiment are also provided on our GitHub. Please note that the results in Table 8 are different from those

Table 8: Reproduced MNIST-SVHN results for the MMVAE model with our multimodal VAE toolkit. We show the digit classification accuracies (%) of latent variables (*MNIST*) and (*SVHN*), and the probability of digit matching (%) for cross- and joint-generation. For our results, we also show in brackets the variance of the results calculated over 3 seeds.

Version	MNIST	SVHN	$\begin{array}{l} \text{MNIST} \\ \rightarrow SVHN \end{array}$	$\begin{array}{l} \text{SVHN} \\ \rightarrow MNIST \end{array}$	Joint
Original	91.3	68.0	86.4	69.1	42.1
Reproduced (Ours)	87.6 (5.2)	70.4 (4.6)	82.7 (5.2)	72.5 (4.9)	45.3 (3.1)

Table 9: Reproduced FashionMNIST results for the MVAE model with our multimodal VAE toolkit. We show the estimated marginal log-likelihoods (lower is better). For our results, we also show in brackets the variance of the results calculated over 3 seeds.

Version	$logp(x_1)$	$logp(x_1, x_2)$	$logp(x_1 x_2)$
Original	-232.535	-233.007	-230.695
Reproduced (Ours)	-234.15 (1.52)	-233.89 (2.61)	-232.56 (3.12)

Table 10: Reproduced PolyMNIST results for the MoPoE model with our multimodal VAE toolkit. We show the Coherence Accuracy (%) of conditionally generated samples (excluding the input modality) (*1 Mod*, *2 Mods*, *3 Mods* and *4 Mods* stand for the number of input modalities) and the joint coherence (*Joint*). For our results, we also show in brackets the variance of the results calculated over 3 seeds.

Version	1 Mod	2 Mods	3 Mods	4 Mods	Joint
Original	67	78	80	83	12
Reproduced (Ours)	66 (4)	73 (5)	81 (3)	82 (5)	11 (3)

Table 11: Reproduced MNIST-SVHN results for the DMVAE model with our multimodal VAE toolkit. We show the probability of digit matching (%) for cross- and joint-generation. For our results, we also show in brackets the variance of the results calculated over 3 seeds.

Version	$\text{MNIST} \rightarrow SVHN$	$SVHN \rightarrow MNIST$	Joint	
Original	88.1	83.7	44.7	
Reproduced (Ours)	84.5 (4.7)	82.2 (3.1)	44.9 (3.6)	

depicted in the main paper, Table 3. This is because here we unified the training hyperparameters with the original paper setup. However, we found that setting the likelihood scaling to 1 for both modalities produces more balanced results (in terms of MNIST/SVHN accuracies) and used thus this setup for the comparative study.

4.2 MVAE

In the original MVAE paper (Wu & Goodman, 2018), the results are reported in terms of marginal log-likelihoods. We reproduced the FashionMNIST experiment with a 64-D latent space, batch size 100, and likelihood scaling of 10 for the labels and 1 for the images, as reported in the public code. The results can be seen in Table 9.

4.3 MoPoE

For verification that the MoPoE model (Sutter et al., 2021) is correct, the reproduction was performed on the PolyMNIST dataset. Based on the original implementation, we used the 512-D latent space, Laplace prior distributions, and $\beta = 2.5$. After training, we calculated the cross-coherencies conditioned on 1, 2, 3 or 4 modalities as reported in the paper. The results are shown in Table 10.

4.4 DMVAE

We reproduced the MNIST-SVHN experiment for the DMVAE model (Lee & Pavlovic, 2021). The reproduced model configuration included shared latent dimensionality $Dim_{shared} = 10$, the private latent dimensionalities were $Dim_{MNIST} = 1$ and for $Dim_{SVHN} = 4$. The used β parameter was 1, and batch size 100. We used the same encoder and decoder networks and an adapted script for calculating the cross- and joint-coherencies. The results are in Table 11.

References

- Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal vae for learning of latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1692–1700, 2021.
- Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multimodal deep generative models. *Advances in Neural Information Processing Systems*, 32, 2019.
- Thomas M. Sutter, Imant Daunhawer, and Julia E Vogt. Generalized multimodal ELBO. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=5Y21VORDBV.
- Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. Advances in Neural Information Processing Systems, 31, 2018.