Don't drop your samples!
Coherence-aware training benefits Conditional diffusion

Nicolas Dufour   Victor Besnier   Vicky Kalogeiton   David Picard

École des Ponts ParisTech   ÉCOLE POLYTECHNIQUE IP PARIS   LIX   Valeo

Demo, Weights, Video and Code
nicolas-dufour.github.io/cad

Introduction

Goal: Learn *diffusion models* in the presence of *conditioning noise*

Approach: *Estimate* conditioning alignment with a *coherence score*

Prior work: *Filtering* the dataset into a high coherence subset [Rombach 2022]

Our method: *Keep* every sample but *condition* the model by the *coherence score*

Two Impala Rams squaring off.

The food on Jeju Island was fishy, in the best possible way. Jeju Island, Sea Urchin, Noodle Soup, Noodles, Spaghetti, Eat, Ethnic Recipes, Food, Macaroni Pasta

Filtered Dataset

CAD Dataset

Coherence-aware diffusion (CAD)

◊ Condition a diffusion model by the *coherence score* of the datapoint.

$$L_{simple} = \mathbb{E}_{(X,y,c)\sim p_{data}, t\sim\mathcal{U}[0,1]}[\|\epsilon - \epsilon_\theta(X_t, y, c, t)\|]$$

◊ (X,y) can be noisy, but X contains interesting information of the data manifold. By discarding it, it would loose the information contained in X.

◊ Providing the coherence score to the model instead, it can learn to choose to use the conditioning or not

Coherence scores

◊ Class conditional:
  → Resample dataset given a label entropy
  → Use Entropy as coherence score

◊ Semantic conditioning:
  → Softmax of a pretrained segmentation network

◊ Text-conditional:
  → CLIPScore

Architecture: TextRIN

SHARED ACROSS RIN BLOCKS

FLAN T5 XL tokens

Self Attention x2

Timestep   Coherence

Latents   Cross Attention From Conditioning to Latents   Cross Attention From Pixels to Latents   Self Attention xN   Latents

Cross Attention From Latents To Pixels

Coherence aware classifier-free guidance

◊ We adapt classifier free guidance in the following way:

$$\hat{\epsilon}_\theta(x_t, y) = \epsilon_\theta(x_t, y, 0) + \omega(\epsilon_\theta(x_t, y, 1) - \epsilon_\theta(x_t, y, 0))$$

Image diversity with guidance   FID vs accuracy

Text-to-image

Top prompt: "a raccoon wearing an astronaut suit. The raccoon is looking out of the window at a starry night; unreal engine, detailed, digital painting, cinematic,character design by pixar and hayao miyazaki unreal 5, daz, hyperrealistic, octane render".
Bottom prompt: "An armchair in the shape of an avocado".

Varying the Coherence score from 1 to 0

User study

Image quality   Text coherence
CAD   Baseline   Filtered   Weighted

Text-to-image

Method	ω	FID ↓	CLIPScore ↑	P ↑	R ↑	D ↑	C ↑
			COCO-10K				
Baseline	10	91.9	25.96	0.281	0.047	0.181	0.222
Weighted	5	98.3	25.15	0.192	0.046	0.111	0.155
Filtered	10	85.8	26.52	0.281	0.061	0.175	0.233
CAD (Ours)	15	69.4	26.16	0.373	0.078	0.265	0.315

COCO-10K Metrics

FID vs CLIPScore

Model: CAD, Filtered, Baseline, Weighted

Class-conditional

Deer Ship Frog Dog Horse Truck Cat Automobile Airplane Bird

TSNE Plot of mixed confidence and label embeddings

Coherence
Class label: Malamute, Volcano, Cheeseburger

Interpolating over the coherence

Acc with β: 0.2, 0.5, 0.8
FID with β: 0.2, 0.5, 0.8

Varying the coherence at test time

Different levels of noise

Semantic conditioning

Caption: Inside the c-130 cargo plane:

Caption: A large room with ping-pong tables and people playing:

Segmap   Confidence   Baseline   Ours   Real

Qualitative results

Metric	ADE20k		COCO	
	Baseline	CAD	Baseline	CAD
FID	33.67	**30.88**	20.1	**18.1**
mIoU	22.6	**23.7**	35.1	**35.3**
P	0.785	**0.844**	0.7876	**0.8404**
R	0.757	**0.824**	0.6760	**0.8060**
D	1.029	**1.0755**	**1.0811**	1.0687
C	0.904	**0.934**	0.8956	**0.9304**

Quantitative results

150M TextRIN Samples

An old-world galleon navigating through turbulent ocean waves under a stormy sky lit by flashes of lightning

an oil painting of rain at a traditional Chinese town

portrait photo of a asia old warrior chief tribal panther make up blue on red side profile looking away serious eyes 50mm portrait photography

a blue jay stops on the top of a helmet of Japanese samurai background with sakura tree

A cute little matte low poly isometric cherry blossom forest island waterfalls lighting soft shadows trending on Artstation 3d render monument

Underwater catedral

A cozy gingerbread house nestled in a dusting of powdered sugar snow adorned with vibrant candy canes and shimmering gumdrops

a teddy bear wearing blue ribbon taking selfie in a small boat in the center of a lake

Pirate ship trapped in a cosmic maelstrom nebula rendered in cosmic beach whirlpool engine volumetric lighting spectacular ambient lights pollution cinematic atmosphere art nouveau