Towards Practical Plug-and-Play Diffusion Models

Hyojun Go*, Yunsung Lee*, Jin-Young Kim*, Seunghyun Lee, Myeongho Jeong, Hyunseung Lee, Seungtaek Choi.

Riiid

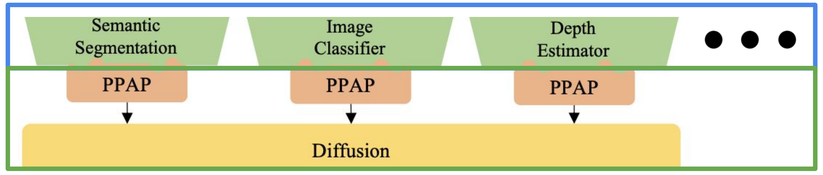
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Plug-and-Play Diffusion Models.

Plug-and-Play guidance with off-the-shelf models:

- By leverage guidance from off-the-shelf models, conduct various generation task with Plug-and-Play manner.

1) Leverage Off-the-shelf models

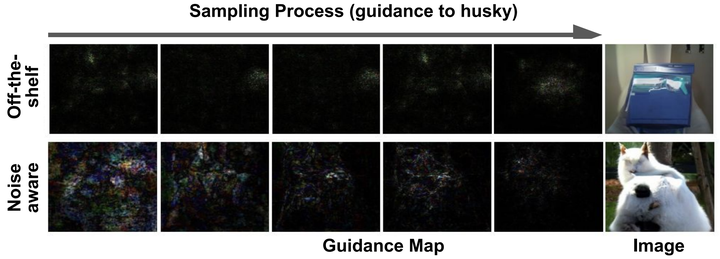


2) Plug-and-Play guidance for various tasks

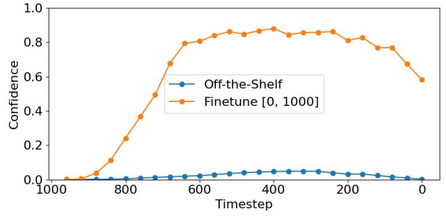
How guidance of off-the-shelf models fail?

1) Guidance map through guided sampling.

- Off-the-shelf model does not produce successful guidance.



2) Classifier confidence through guided sampling.

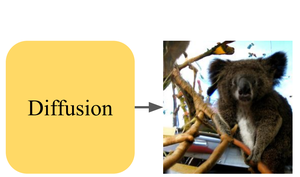


- Off-the-shelf model cannot increase the confidence through guidance.
- Finetune off-the-shelf model on data corrupted by diffusion process can increase confidence.

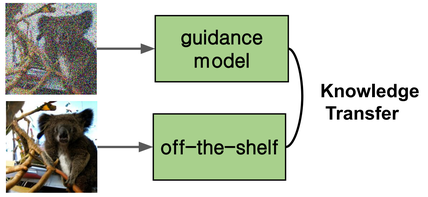
Method: Data-Free Knowledge Transfer

Make off-the-shelf models guidable without data.

- **Hurdle:** Whenever plugin off-the-shelf models, we require to collect labeled data for finetuning on diffused data.
- **Solution:** Data-Free Knowledge Transfer.



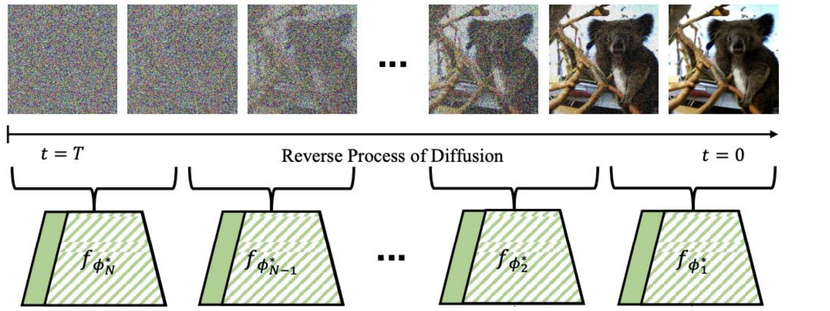
1) Generate the clean dataset using the diffusion model



2) mimicking the prediction on clean data of off-the-shelf models in diffused data

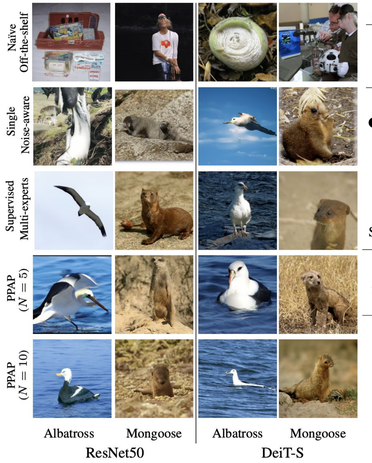
Method: Multi-Expert Strategy

Make guidance models more powerful.



- **Multi-expert strategy:** use multiple guidance model, each fine-tuned to specialize in a specific noise region.
  - 1) Same inference speed, 2) Superior performance 3) Increased memory
- **Efficient Multi-expert strategy**:
  - Parameter-efficient finetuning to reduce memory cost.

Results: ADM guidance from classifier



Sampler	Method	FID	IS	Precision	Recall
DDIM 25 step	ResNet152 + PPAP	**26.78**	**49.59**	**0.5976**	0.6208
	ResNet50 + PPAP	29.65	44.23	0.5872	0.6012
	ResNet152 + single noise aware	29.63	42.09	0.5601	0.6305
	ResNet152 + off-the-shelf	40.15	33.33	0.5074	0.6175

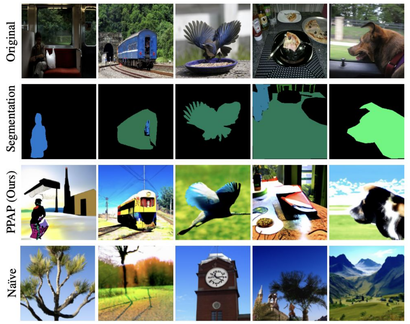
- PPAP (Multi-expert + Data-Free) can produce comparable results to supervisely finetuned model.

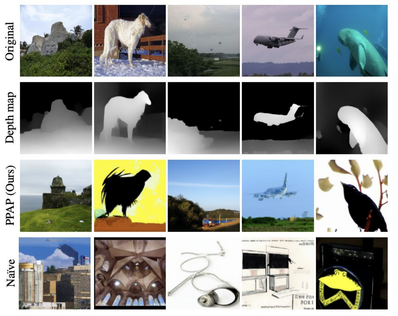
Sampler	Method	Trainable Parameters	FID	IS	Precision	Recall
DDIM 25 step	ResNet18×5	58.4M	**20.38**	**63.74**	**0.6638**	0.5898
	ResNet152×1	60M	29.63	42.09	0.5601	0.6305
	ResNet50×5	109.9M	**19.98**	**74.77**	**0.6476**	0.5887
	ResNet50×1	25.5M	30.42	43.05	0.5509	0.6187

- Fair setting in parameters, multi-experts outperform large single finetuned model.

Results: Plug-and-Play guidance with GLIDE

- Without any data, PPAP can successfully guide GLIDE with
  - Classifier (ResNet50)
  - Segmentation (DeepLabV3)
  - Depth (MiDaS)





Towards Practical Plug-and-Play Diffusion Models

Hyojun Go*, Yunsung Lee*, Jin-Young Kim*, Seunghyun Lee, Myeongho Jeong, Hyunseung Lee, Seungtaek Choi.

TUE-AM-187

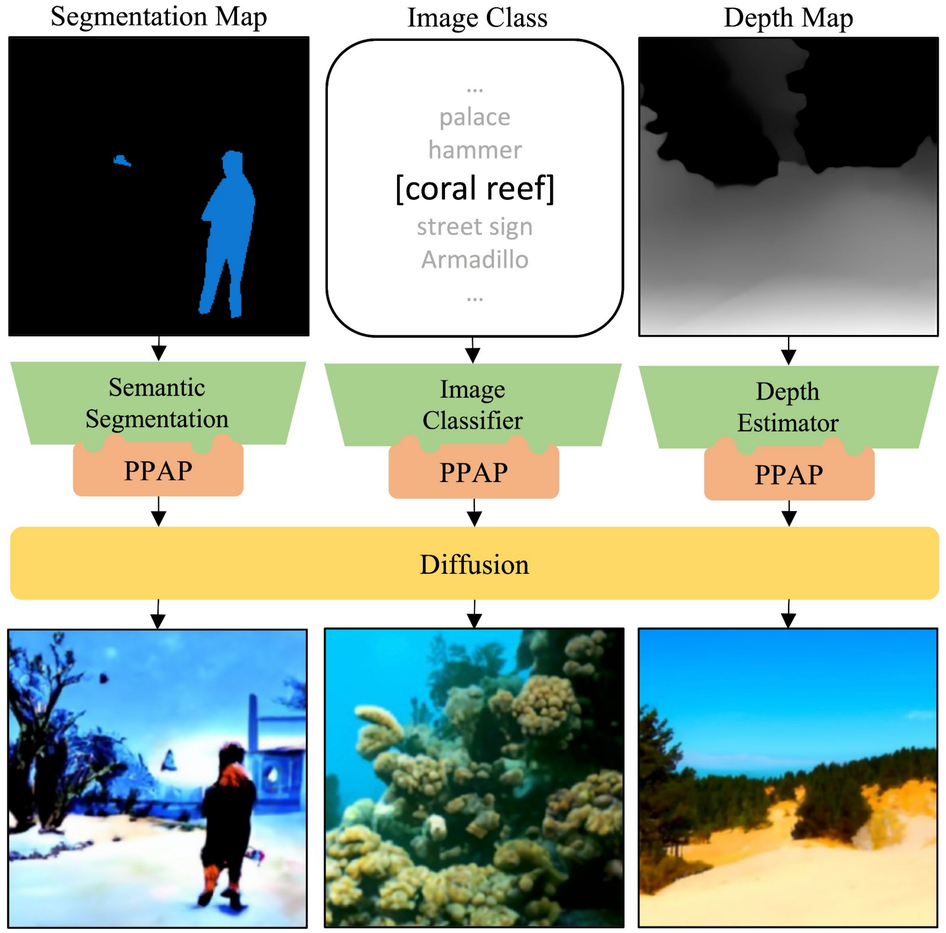
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Practical Plug-And-Play: a practical diffusion guidance framework

Extend classifier guidance to any other user intentions

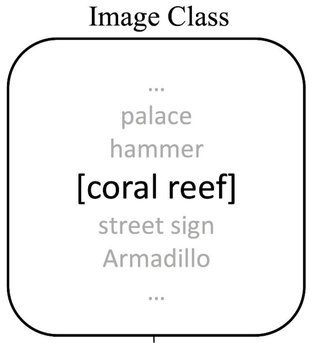
with off-the-shelf models!

on pre-trained diffusion models, even if it is trained on unreleased dataset!



Overview - Classifier Guidance

User Intention: I want an image of a “coral reef”.

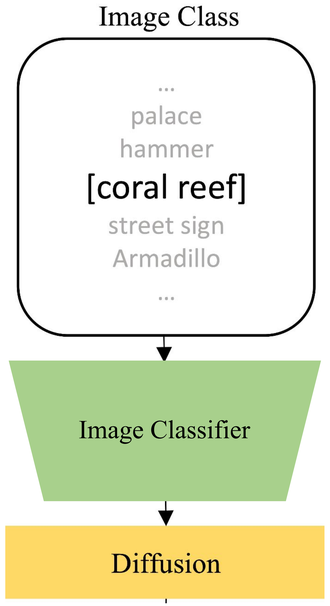


Overview - Classifier Guidance

User Intention: I want an image of a “coral reef”.

Guidance Model:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)) + \sigma_t z - s\sigma_t \nabla_{x_t}\mathcal{L}_{guide}(f_\phi(x_t), y_{target}),$$



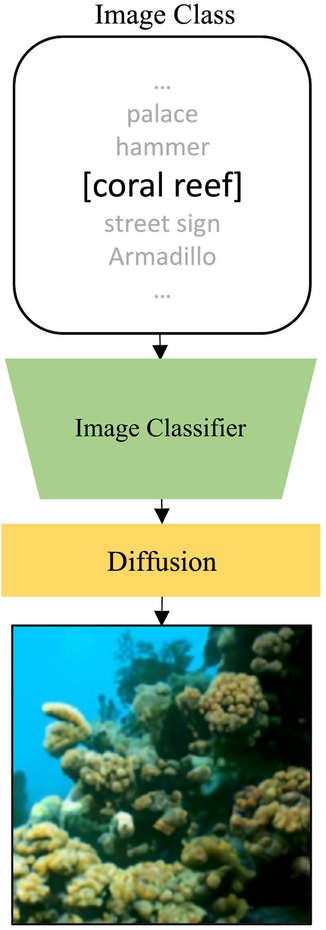
Overview - Classifier Guidance

User Intention: I want an image of a “coral reef”.

Guidance Model:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)) + \sigma_t z - s\sigma_t \nabla_{x_t}\mathcal{L}_{guide}(f_\phi(x_t), y_{target}),$$

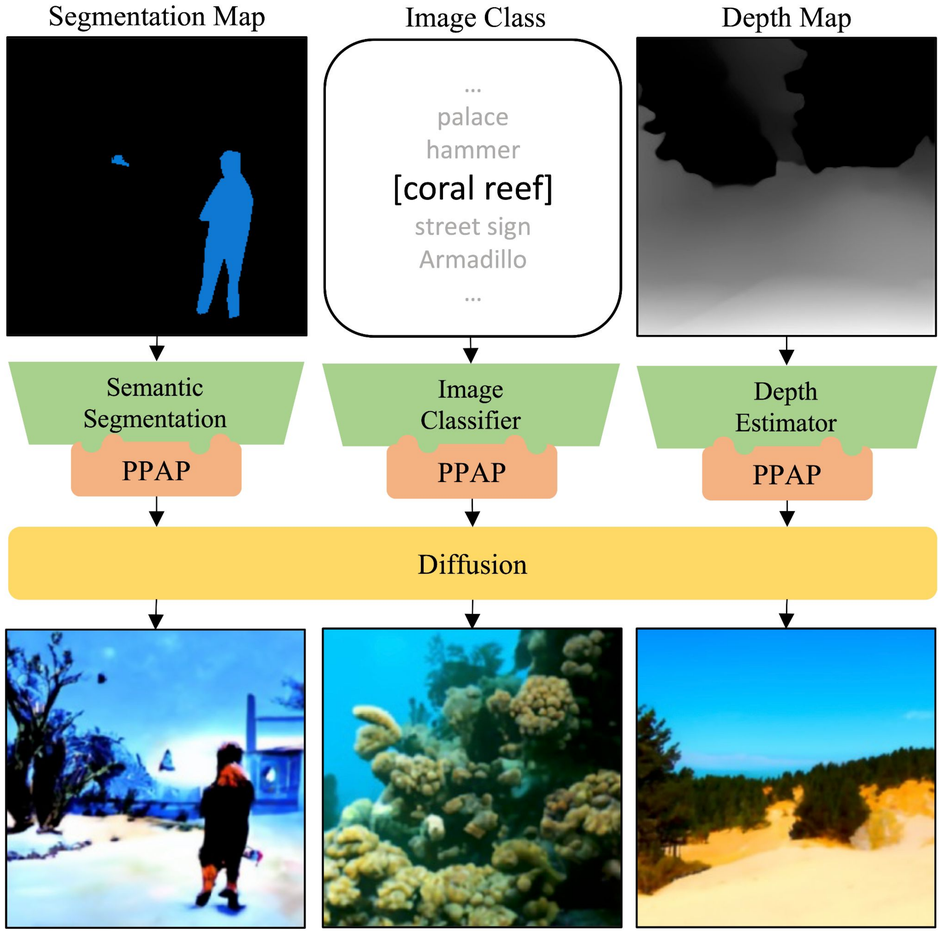
Output Image:



Overview - PPAP

Practical Plug-And-Play: a practical diffusion guidance framework

Extend classifier guidance to any other user intentions

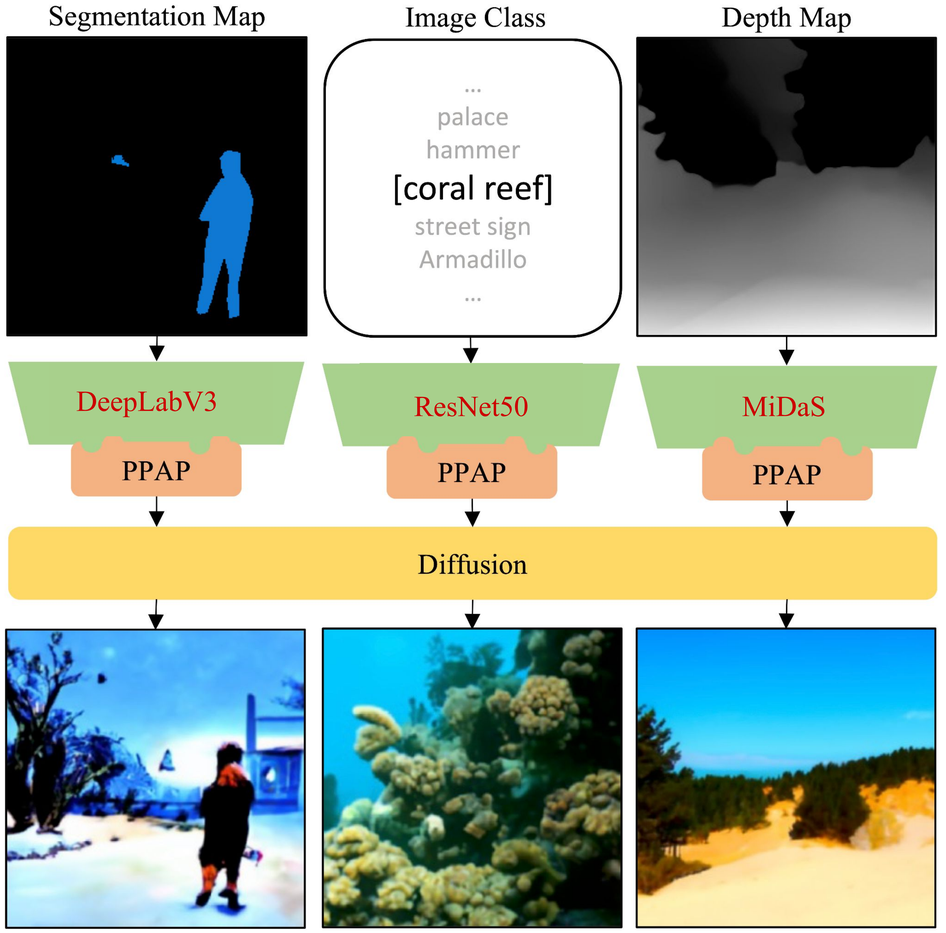


Overview - PPAP

Practical Plug-And-Play: a practical diffusion guidance framework

Extend classifier guidance to any other user intentions

with off-the-shelf models!



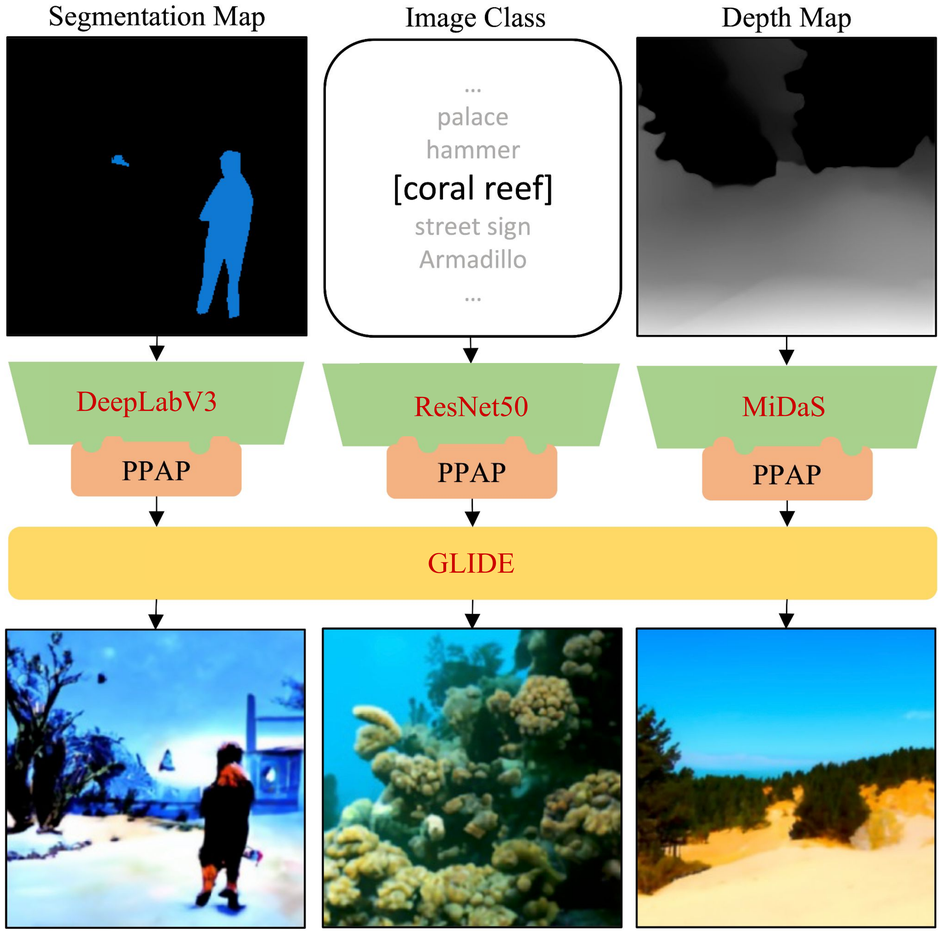
Overview - PPAP

Practical Plug-And-Play: a practical diffusion guidance framework

Extend classifier guidance to any other user intentions

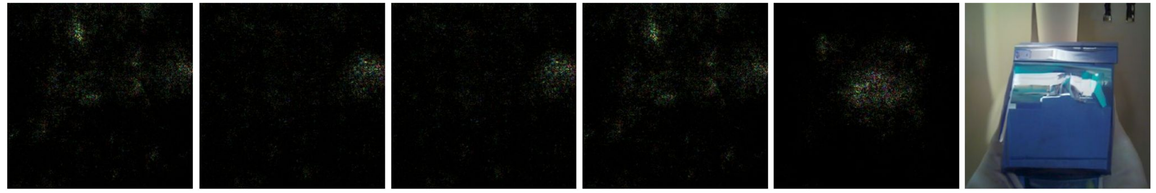
with off-the-shelf models!

on pre-trained diffusion models, even if it is trained on unreleased datasets!



Challenges

We can't just guide the diffusion model with an off-the-shelf model.



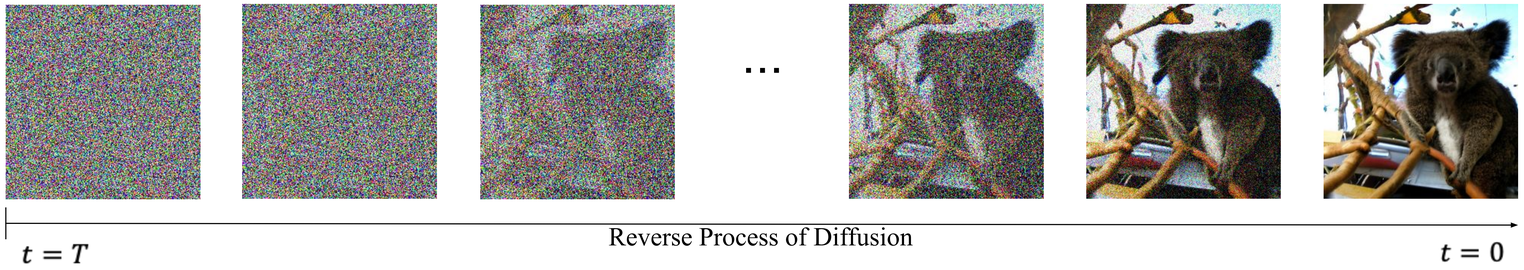
Gradient maps for x_t on $t \in [920, 720, 520, 320, 120]$ (left 5) and generated images (rightmost)
when the diffusion model is guided to “husky” with a ImageNet pretrained off-the-shelf model (ResNet50).

Naive off-the-shelf guidance fails completely.

But.. WHY???

Challenges

We can't just guide the diffusion model with an off-the-shelf model.



Naive off-the-shelf guidance fails completely.

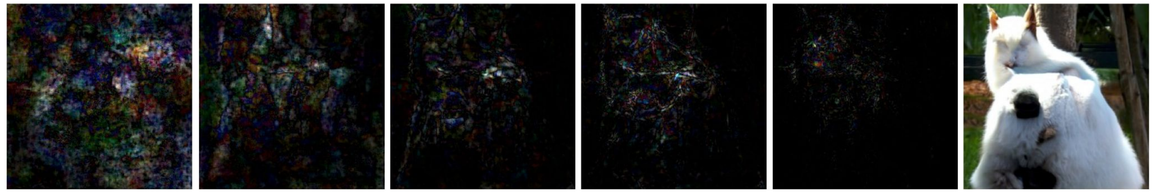
This is because diffusion models receive noisy images as input during the reverse process, while off-the-shelf models have never been trained on such images.

That’s why Guided Diffusion[1] train the classifier model with sampled images in diffusion process.

[1] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. Advances in Neural Informa- tion Processing Systems, 34:8780–8794, 2021.

Challenges

However, the guidance of a fine-tuned classifier on a noisy image is still poor.

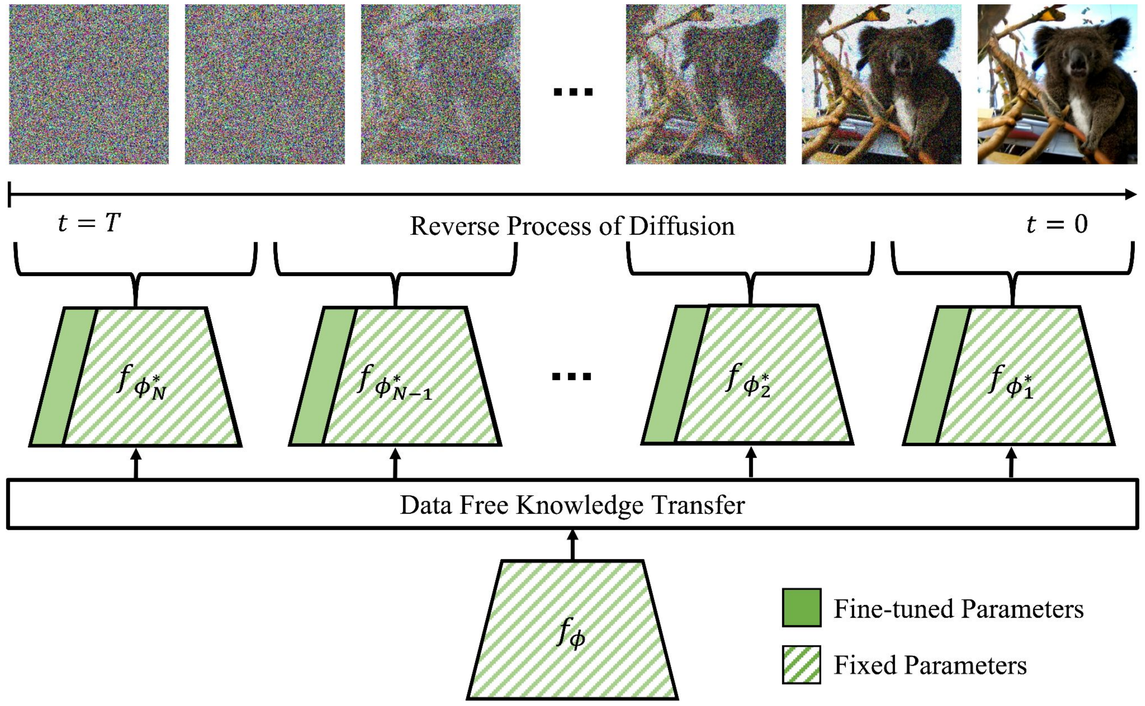


Gradient maps for x_t on $t \in [920, 720, 520, 320, 120]$ (left 5) and generated images (rightmost) when the diffusion model is guided to "husky" with a fine-tuned ResNet50 classifier.

It has two limitations

(1) Performing on inputs with extremely various noises is too hard for a single guidance model.
(2) It requires a labeled dataset which hinders scaling up for various tasks.

PPAP



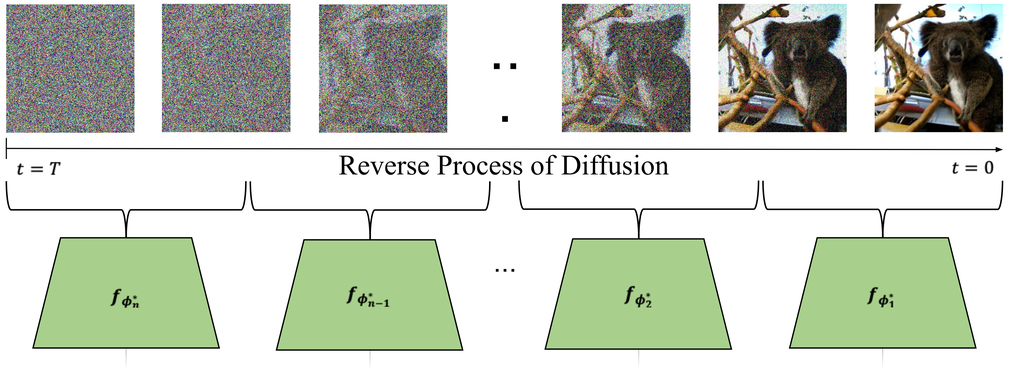
(1) Multi-experts strategy
(2) Practical framework with parameter efficient fine-tuning and data free knowledge transfer

Multi-Experts Strategy

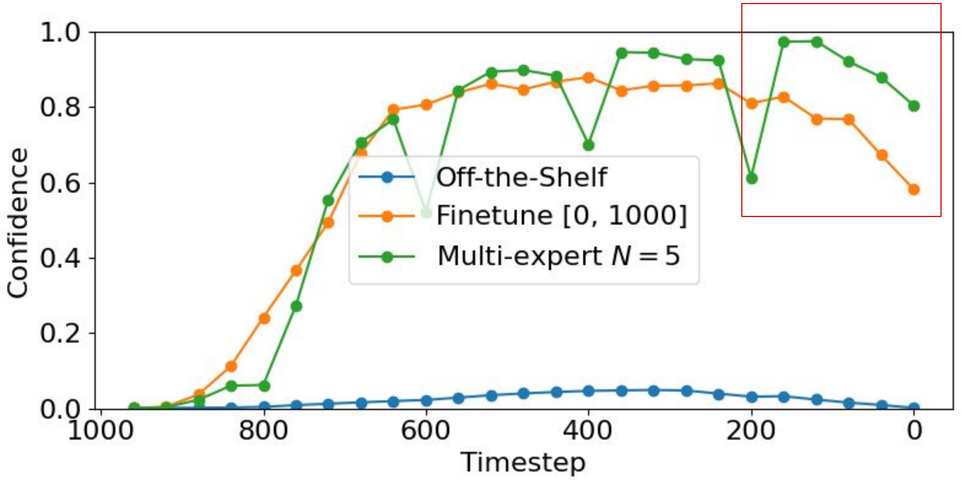
Each expert is fine-tuned to specialize in a specific noise range

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z - s\sigma_t \nabla_{x_t}\mathcal{L}_{guide}\left(f_{\phi_n}(x_t), y\right),$$

n-th expert f_{ϕ_n} is trained to predict the ground-truth label y given noisy data x_t , $t \in \{\frac{n-1}{N}T + 1, \ldots, \frac{n}{N}T\}$



Multi-Experts Strategy



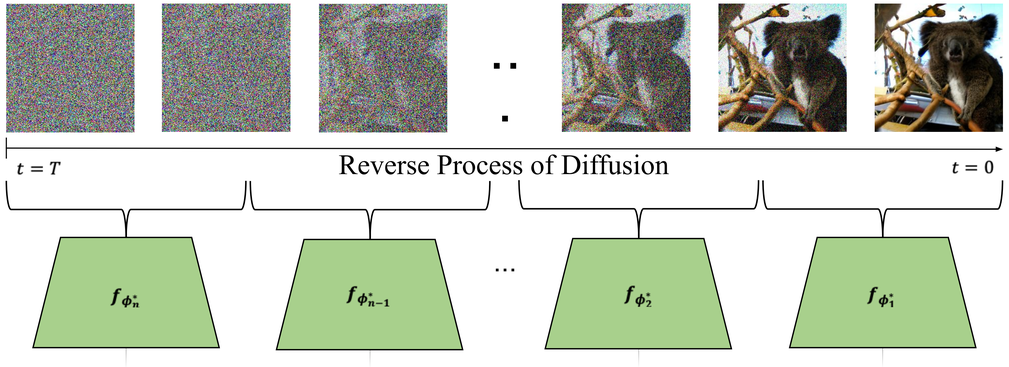
Classifier confidence during the reverse process.

A sinalge fine-tuned classifier confidence drops for cleaner images $(t \approx 200)$

Our multi-experts show high confidence at most time-steps.

Naïve off-the-shelf
Single noise-aware
Expert [0,200]
Expert [200,400]
Expert [400,600]
Expert [600,800]
Expert [800,1000]
Multi experts-5

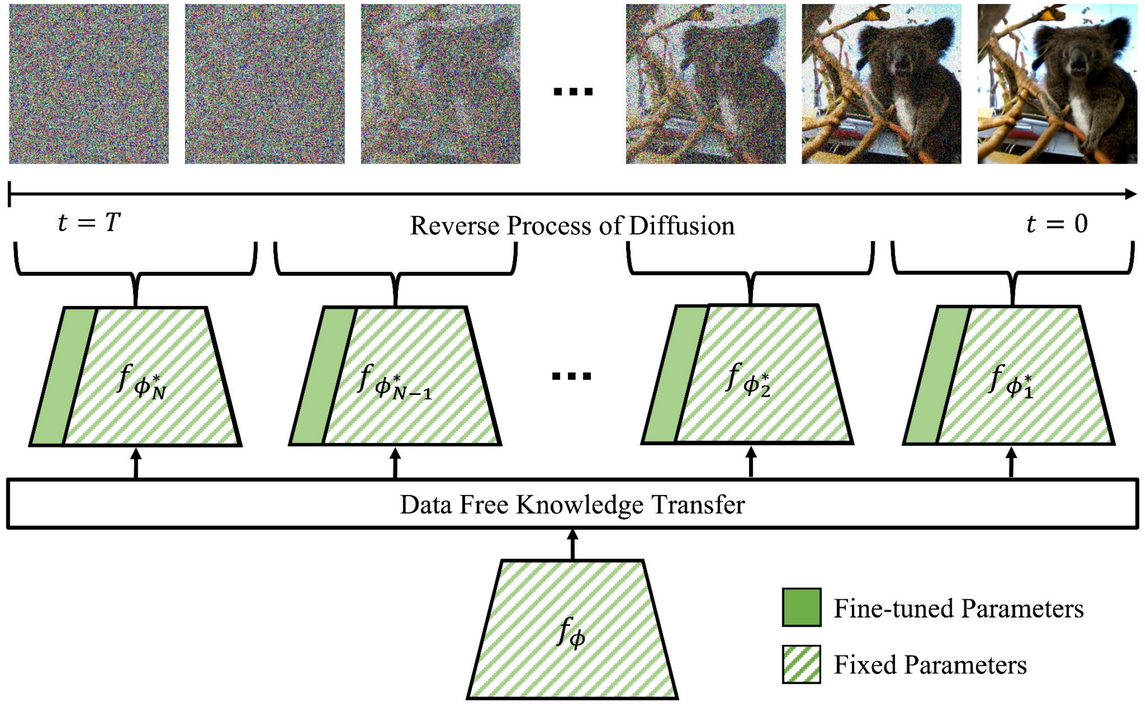
Practical Plug-and-Play Diffusion



We want to extend classifier guidance to any other user intentions
on pre-trained diffusion models, even if it is trained on unreleased dataset!

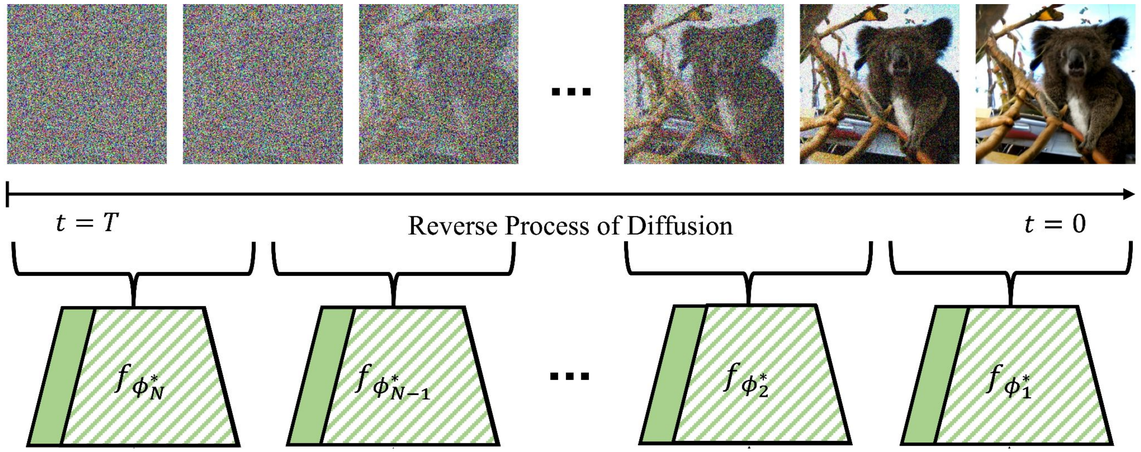
However, whenever applying a new off-the-shelf model,
the multi- experts strategy must
(1) utilize multiple networks
(2) collect the labeled dataset.

Practical Plug-and-Play Diffusion



(1) utilize multiple networks -> parameter-efficient fine-tuning scheme
(2) collect the labeled dataset -> data-free knowledge transfer scheme

Parameter Efficient Multi-Experts Strategy



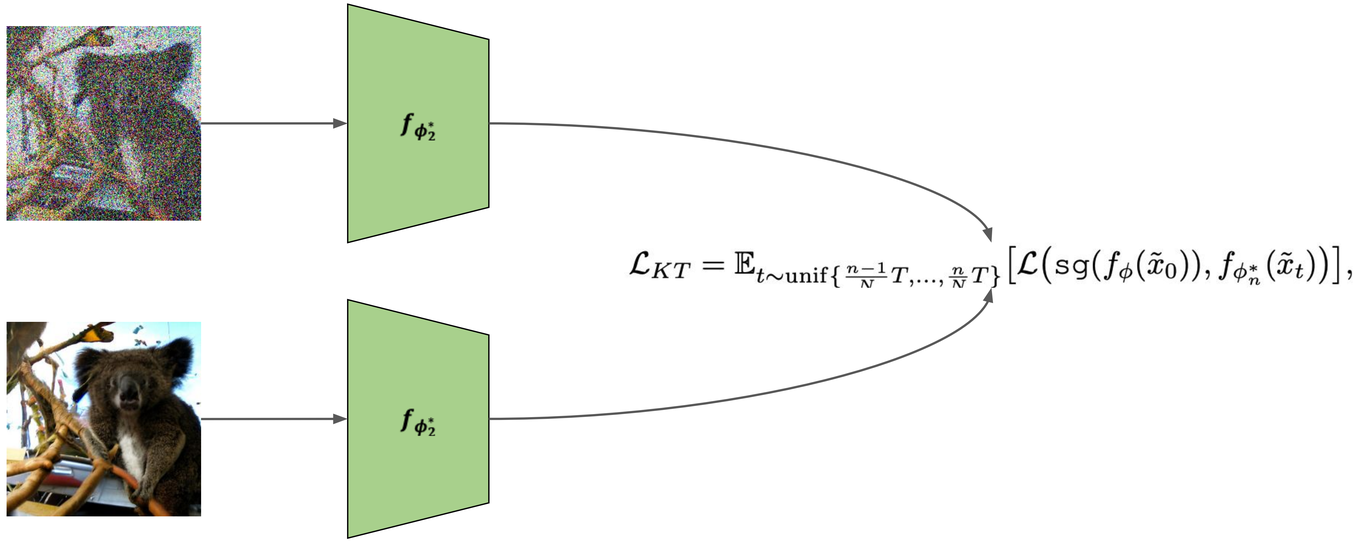
To reduce the number of parameters to fine-tune in multi-experts strategy,
we only fine-tunes a small number of parameters while reusing most of the frozen off-the-shelf model.

Specifically, we fine-tune bias and batch norm, and apply LORA[2].

[1] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen- Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. International Conference on Learning Representations, 2022.

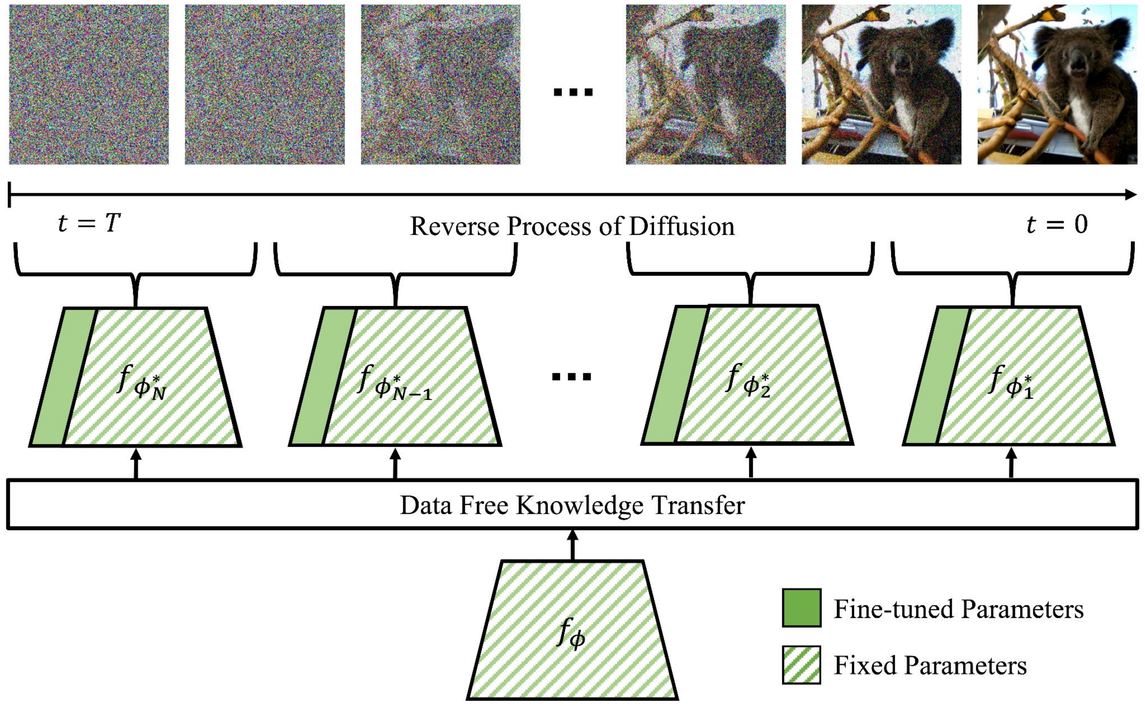
Data Free Knowledge Transfer

1) Generate the clean dataset using the diffusion model



2) Knowledge transfer by mimicking the prediction on a clean image

Practical Plug-and-Play Diffusion

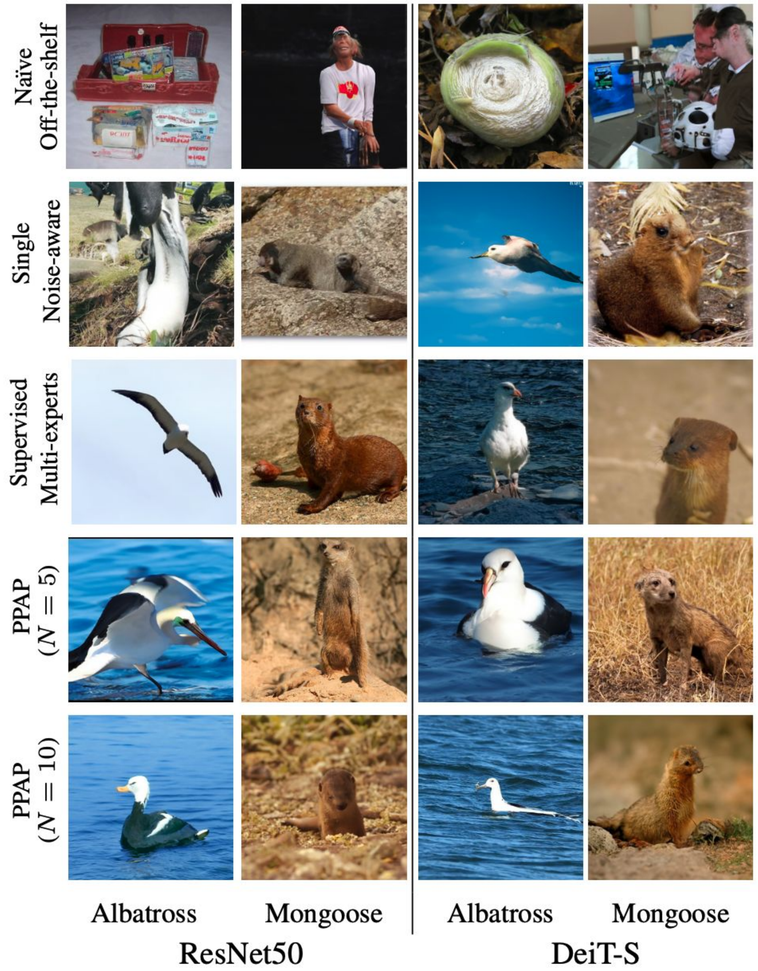


Experiments - ImageNet Classifier Guidance

Architecture	Sampler	Guidance	Trainable Parameters	Supervision	FID (↓)	IS (↑)	Precision (↑)	Recall
ResNet50	DDIM (25 Steps)	No	-	None	40.24	34.53	0.5437	0.6063
		Naïve off-the-shelf	-	None	38.74	33.95	0.5192	0.6152
		Gradients on $\hat{x}_0$	-	None	38.14	33.77	0.5277	0.6252
		Single noise aware	25.5M (100%)	ImageNet ($\approx$ 1.2M)	30.42	43.05	0.5509	0.6187
		Multi-experts-5	127.5M (500%)	ImageNet ($\approx$ 1.2M)	**19.98**	**74.78**	**0.6476**	0.5887
		PPAP-5	7.3M (28.6%)	Data-free ($\approx$ 0.5M)	29.65	44.23	0.5872	0.6012
		PPAP-10	14.6M (57.2%)	Data-free ($\approx$ 0.5M)	27.86	46.74	0.6079	0.5925
	DDPM (250 Steps)	No	-	None	28.97	40.34	0.6039	0.6445
		Naïve off-the-shelf	-	None	29.03	39.79	0.6042	0.6474
		Gradients on $\hat{x}_0$	-	None	28.81	39.80	0.6095	0.6475
		Single noise aware	25.5M (100%)	ImageNet ($\approx$ 1.2M)	38.15	31.29	0.5426	0.6321
		Multi-experts-5	127.5M (500%)	ImageNet ($\approx$ 1.2M)	**16.37**	**81.47**	**0.7216**	0.5805
		PPAP-5	7.3M (28.6%)	Data-free ($\approx$ 0.5M)	22.70	52.74	0.6338	0.6187
		PPAP-10	14.6M (57.2%)	Data-free ($\approx$ 0.5M)	21.00	57.38	0.6611	0.5996
DeiT-S	DDIM (25 Steps)	No	-	None	40.24	34.53	0.5437	0.6063
		Naïve off-the-shelf	-	None	37.51	33.74	0.5293	0.6186
		Gradients on $\hat{x}_0$	-	None	38.10	33.75	0.5288	0.6212
		Single noise aware	21.9M (100%)	ImageNet ($\approx$ 1.2M)	44.13	28.31	0.4708	0.6030
		Multi-experts-5	109.9M (500%)	ImageNet ($\approx$ 1.2M)	**17.06**	**80.85**	**0.7001**	0.5810
		PPAP-5	4.6M (21.3%)	Data-free ($\approx$ 0.5M)	25.98	48.80	0.6128	0.5984
		PPAP-10	9.3M (42.6%)	Data-free ($\approx$ 0.5M)	24.77	50.56	0.6220	0.5990
	DDPM (250 Steps)	No	-	None	28.97	40.34	0.6039	0.6445
		Naïve off-the-shelf	-	None	29.41	39.55	0.6032	0.6320
		Gradients on $\hat{x}_0$	-	None	30.26	37.75	0.6043	0.6407
		Single noise aware	21.9M (100%)	ImageNet ($\approx$ 1.2M)	36.01	31.90	0.5461	0.6479
		Multi-experts-5	109.9M (500%)	ImageNet ($\approx$ 1.2M)	**14.95**	**83.26**	**0.7472**	0.5686
		PPAP-5	4.6M (21.3%)	Data-free ($\approx$ 0.5M)	22.30	53.62	0.6368	0.6074
		PPAP-10	9.3M (42.6%)	Data-free ($\approx$ 0.5M)	20.07	60.62	0.6734	0.5963

The multi-expert strategy is the most powerful, and PPAP is scalable for practical cases, but also powerful enough.
Both are our novel models.
Both are generalizable across samplers and architectures.

Experiments - ImageNet Classifier Guidance



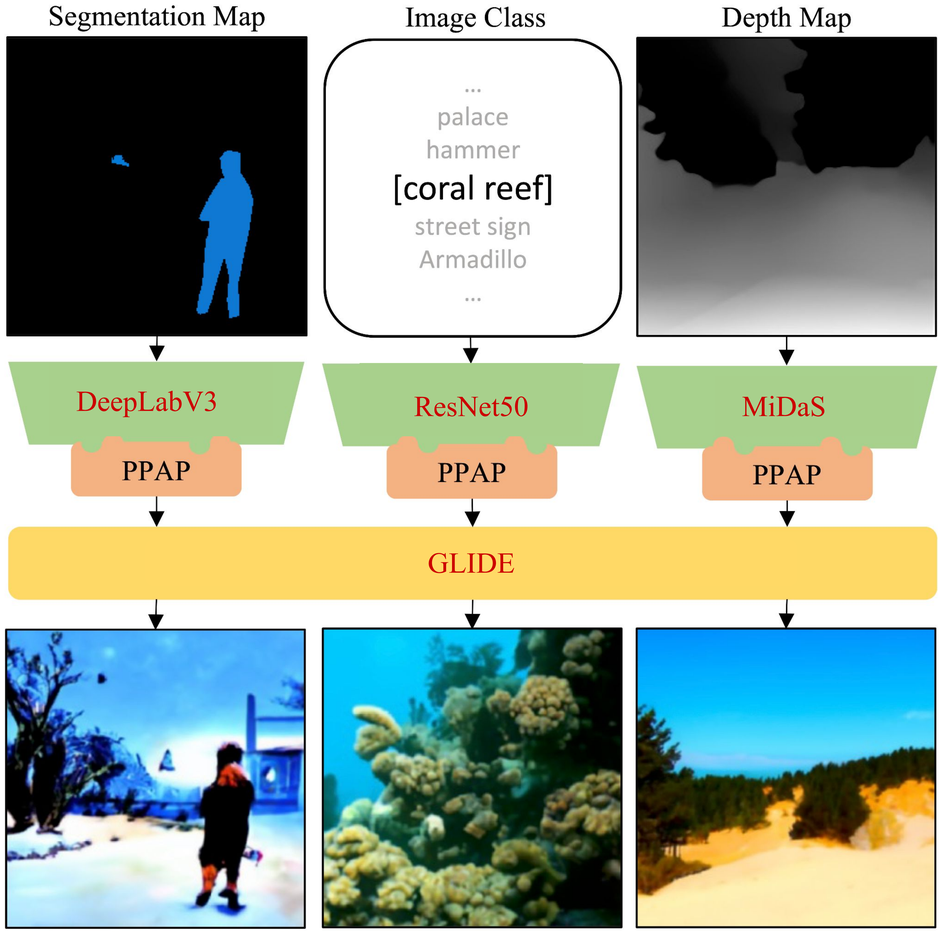
Experiments - Guiding GLIDE for Various Downstream Tasks

- GLIDE: trained on a large-scale unreleased CLIP-filtered dataset
- ResNet50: ImageNet pretrained classifier
- MiDaS: pre-trained zero-shot depth estimator
- DeepLabV3: pre-trained semantic segmentation model

Using only the pre-trained weights of these models,

and without any labeled dataset,

we can successfully guide on various user intentions.

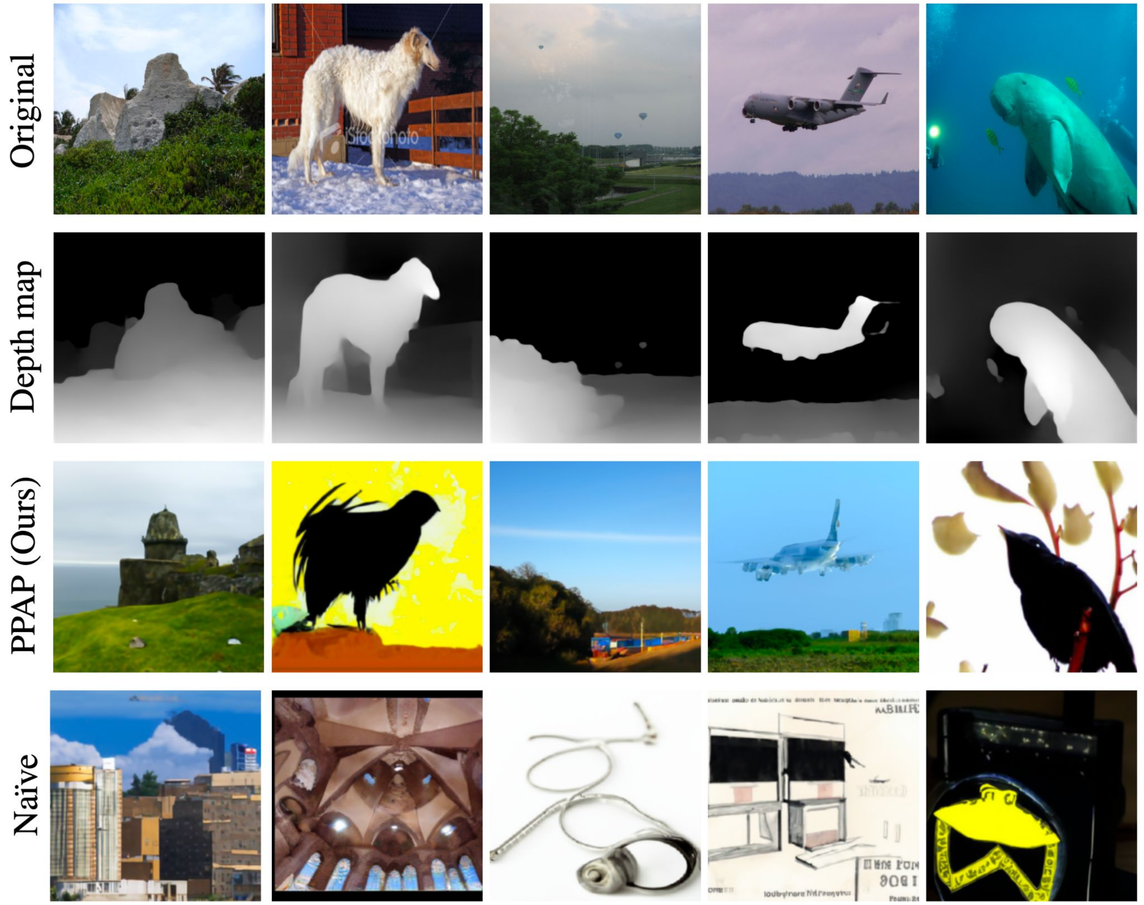


Experiments - Guiding GLIDE with ResNet50 (Class)

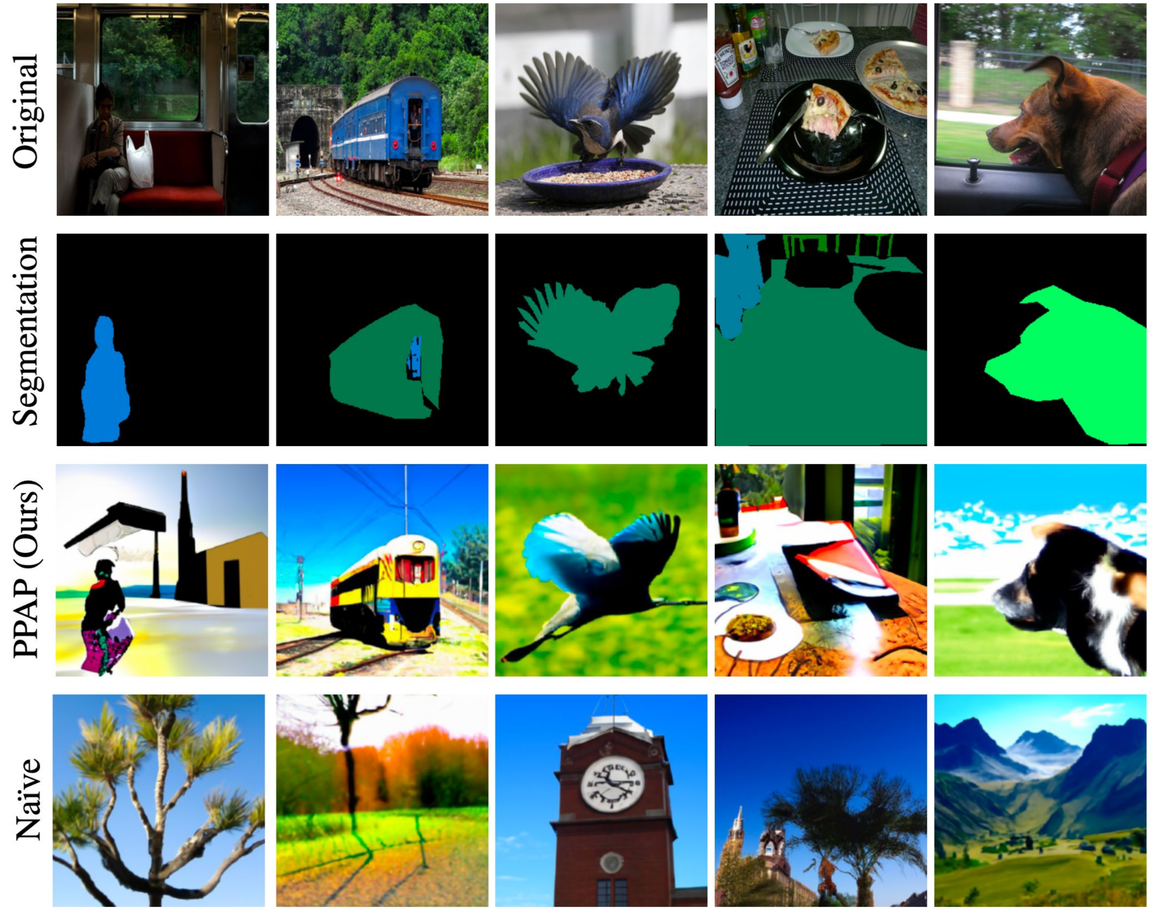
Notably, our method can semantically guide the GLIDE with varying styles of images, such as cartoon-style images, which is interesting because ResNet50 has never seen cartoon-style images in the ImageNet.

PPAP can obtain both the generation ability of GLIDE in various domains and the semantic understanding ability of ResNet50

Experiments - Guiding GLIDE with MiDaS (Depth Map)



Experiments - Guiding GLIDE with DeepLabV3 (Seg Map)



Riiid

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Thank you

Towards Practical Plug-and-Play Diffusion Models

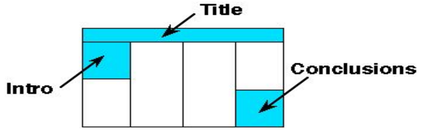
Hyojun Go, Yunsung Lee, Jinyoung Kim, Seunghyun Lee, Myeongho Jeong, Hyunseung Lee, Seungtaek Choi.

Introduction

- Posters boards are 48” tall and 96” wide, but we recommend you leave a little border since you may not be able to pin at the vertical edge. Since PowerPoint does not let one define such a large paper size, this template is designed to be printed at 200%, yielding a 46” x94” poster. You can scale it up or down a bit (e.g. 42” is a common paper size at FexEd). Note there is no direct international A0.. A1 equivalent. The poster size is approximately three A0 boards next to each other, i.e., each column in this example is about one A0 board.

- Ideally you want to keep it very readable: this is not your paper, it is a poster. 32pt here (64 final printing) is good for most text:
  - Sub-bullets are 28 here (56 final)
  - Don’t use smaller than 24pt in this template (which is 48pt in final printing at 200%)
  - Insert plenty of graphics and any math you need

- When inserting graphics or equations, keep the resolution high (remember this will be printed at 200%). If you can see blocking artifacts at 400% magnification in PowerPoint, consider finding better graphics. This is an example of BAD/LOW RES GRAPHICS



- Leave enough margin for pushpin and remember many big plotters cannot get within .5” of the actual paper edge.

- You are free to use colored backgrounds and such but they generally reduce readability.

- You are free to use what ever fonts you like.
  - San Serif fonts like Arial are more readable from a distance,
  - Serif fonts like times may look more consistent with your mathematics

Section 2 (layout):

- Remember the poster session will be crowded so design the poster to be read in columns so people can read what is in front of them and move left to right to get the whole story.

- The poster should use photos, figures, and tables to tell the story of the study. For clarity, present the information in a sequence that is easy to follow.

- There is often way too much text in a poster - there definitely is in this template! Posters primarily are visual presentations; the text should support the graphics. Look critically at the layout. Some poster 'experts' suggest that if there is about 20-25% text, 40-45% graphics and 30-40% empty space, you are doing well.

Section 3:

- Include more figures than are in the paper so you can talk to them. Include things that are not in the paper and then encourage them to read the paper. Don’t try to just put all the paper here.

- If it looks like a cut/paste of the paper, people skip that poster since they can read the papers after the conference. Many people find it better to spend time talking with poster presenters that have more to offer than just redoing the paper content paper in big fonts.

- People will likely have already seen your posted video, so the poster can serve as a talking point for you.

- Remember Poster boards look like this.. This is your canvas. Paint us a picture of your work.

Summary/Conclusion

- Summarize your contributions

- Summarize your results (if applicable)

- **You an add in links to additional videos, code, or project website (or QR code)**

References