Not All Prompts Are Created Equal

Prompt-based Pruning of Text-to-Image Diffusion Models

Alireza Ganjdanesh*[1], Reza Shirkavand*[1], Shangqian Gao[2], Heng Huang[1]

1- Department of Computer Science, University of Maryland - College Park

2- Department of Computer Science, Florida State University

DEPARTMENT OF COMPUTER SCIENCE

FLORIDA STATE UNIVERSITY

ICLR International Conference On Learning Representations

Motivation

- **Problem:** Diffusion denoising is a call to the same network over and over again.
  - Same network is used in every step.
- **Goal:** Select a subnetwork based on the input prompt to be used in each sampling step.
- **Key Innovation:** Prompt router + specialized experts.

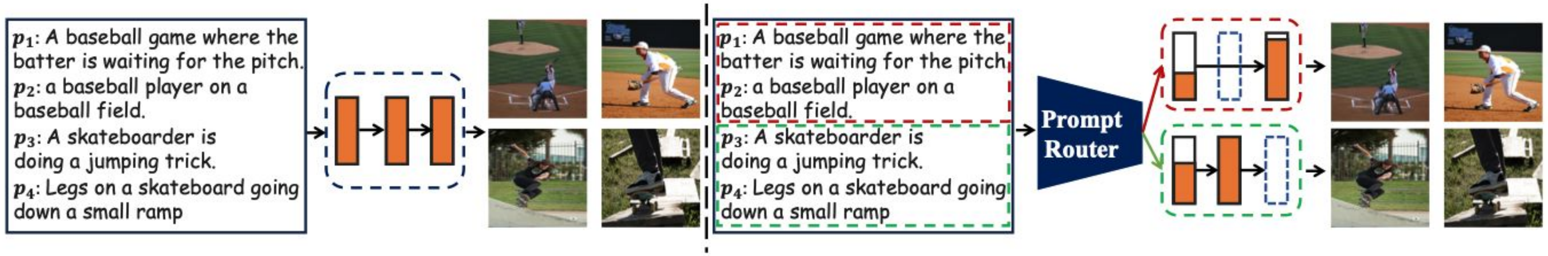

Fig. 1: Overview of APTP

Overview of APTP

- **Prompt Encoder:** Encodes prompt into embeddings.
- **Router Module:** Maps embeddings to architecture codes.
- **Contrastive Loss:** Map similar prompts to similar architecture embeddings.
- **Architecture Codes:** Determine which sub-model (expert) to use.
- **Pruning Units:** Attention Heads -
- **Optimal transport:** Ensure balanced capacity and prevent Expert Collapse.
- **Distillation:** Very effective. Makes convergence faster.

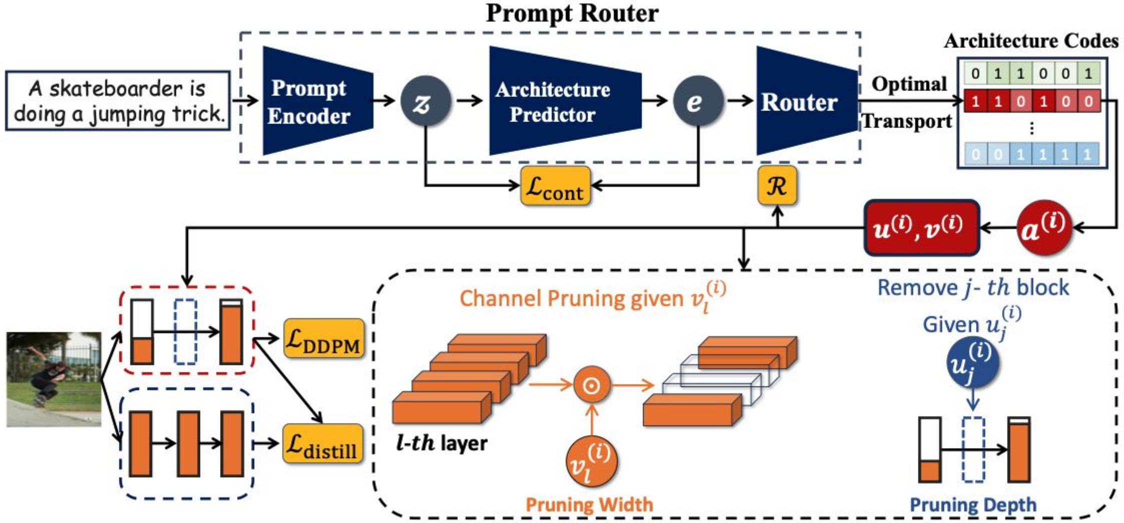

Fig. 2: APTP Pruning Scheme

Quantitative Results

- **Datasets:** CC3M and COCO.
- We outperform SOTA static pruning methods by a wide margin.
- At a reasonable pruning ratio, we get very close performance to the dense model
  - With lower latency and memory usage.
  - Larger dataset and more training improves the results

Table 1: Results on CC3M and MS-COCO. We report performance metrics using samples generated at the resolution of 768 then downsampled to 256 (Kim et al., 2023). We measure models' MACs/Latency with the input resolution of 768 on an A100 GPU. @30/50k shows fine-tuning iterations after pruning.

CC3M					
	Complexity		Performance		
Method	MACs (@768)	Latency (↓) (Sec/Sample) (@768)	FID (↓)	CLIP (↑)	CMMD (↓)
Norm (Li et al., 2017) @50k	1185.3G	3.4	141.04	26.51	1.646
SP (Fang et al., 2023) @30k	1192.1G	3.5	75.81	26.83	1.243
BKSDM (Kim et al., 2023) @30k	1180.0G	3.3	87.27	26.56	1.679
APTP(0.66) @30k	916.3G	2.6	60.04	28.64	1.094
APTP(0.85) @30k	1182.8G	3.4	36.77	30.84	0.675
SD 2.1	1384.2G	4.0	32.08	31.12	0.567

(a)

MS-COCO					
	Complexity		Performance		
Method	MACs (@768)	Latency (↓) (Sec/Sample) (@768)	FID (↓)	CLIP (↑)	CMMD (↓)
Norm (Li et al., 2017) @50k	1077.4G	3.1	47.35	28.51	1.136
SP (Fang et al., 2023) @30k	1071.4G	3.3	53.09	28.98	0.926
BKSDM (Kim et al., 2023) @30k	1085.4G	3.1	26.31	28.89	0.611
APTP(0.64) @30k	890.0G	2.5	39.12	29.98	0.867
APTP(0.78) @30k	1076.6G	3.1	22.60	31.32	0.569
SD 2.1	1384.2G	4.0	15.47	31.33	0.500

(b)

Table 2: PickScore on PartiPrompts as a proxy for human preference. The prompts in this benchmark can be considered out-of-distribution for the router as they are significantly longer and semantically different from MS-COCO.

Train on MS-COCO			
	Complexity		PartiPrompts
Method	MACs (@768)	Latency (↓) (Sec/Sample) (@768)	PickScore (↑)
Norm Pruning	1077.4G	3.1	18.563
Structural Pruning	1071.4G	3.3	19.317
BKSDM	1085.4G	3.1	19.941
APTP (0.64)	890.0G	2.5	20.626
APTP (0.78)	1076.6G	3.1	21.150
SD 2.1	1384.2G	4.0	21.316

Qualitative Results

- **Generations**

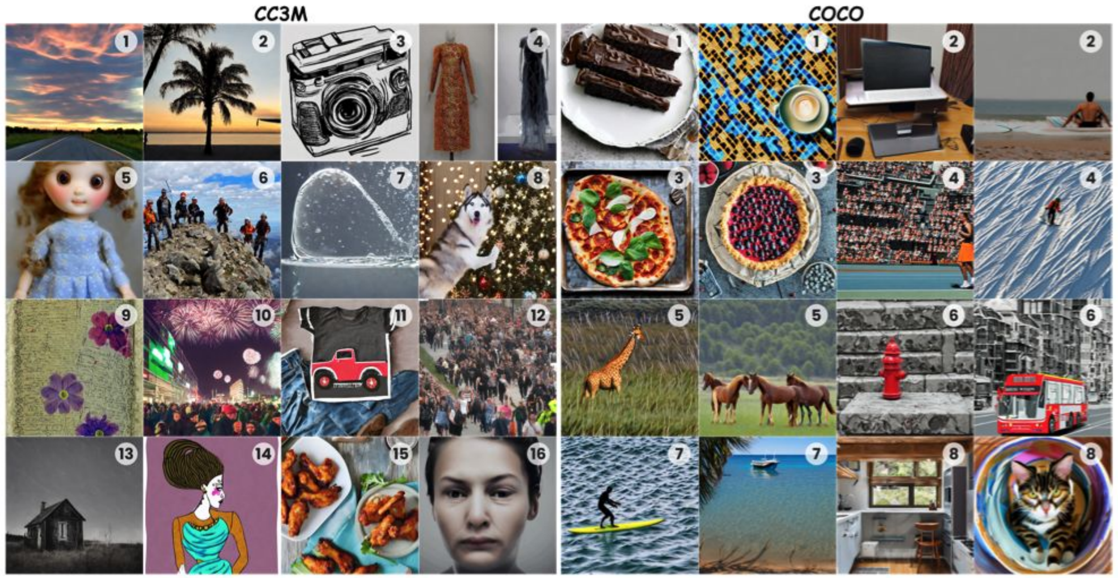

Fig. 3: Samples of APTP-Base Experts.

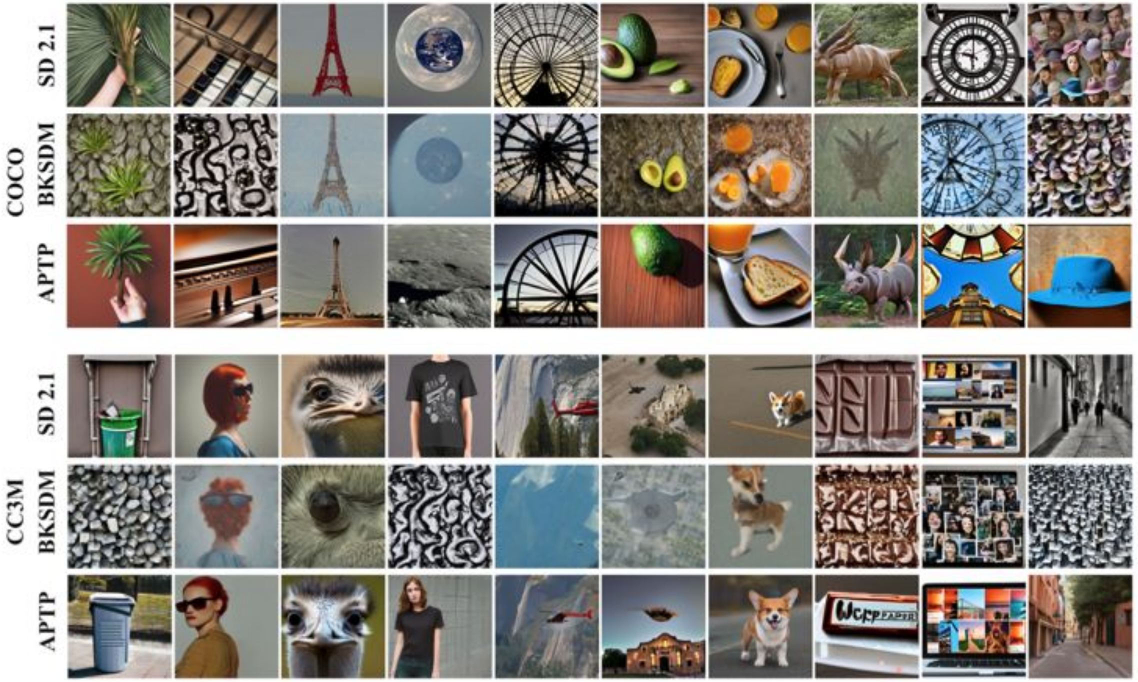

Fig. 4: Qualitative Comparisons with SD2.1 and the best baseline.

Expert Analysis

- **Specialized Experts**

Table 2: The most frequent words in prompts assigned to each expert of APTP-Base pruned on CC3M. The resource utilization of each expert is indicated in parentheses.

Expert 1 (0.72)	**Expert 2 (0.73)**	**Expert 3 (0.75)**	**Expert 4 (0.76)**
View - Sunset - City - Building - Sky	View - Boat - Sea	Artist - Actor	Actor - Dress - Portrait
Expert 5 (0.77)	**Expert 6 (0.78)**	**Expert 7 (0.79)**	**Expert 8 (0.79)**
Illustration - Portrait - Photo	Player - Ball - Game - Team	Background - Water - River - Tree	Biological Species - Dog - Cat
Expert 9 (0.79)	**Expert 10 (0.80)**	**Expert 11 (0.81)**	**Expert 12 (0.81)**
Illustration - Vector	People	Car - City - Road	Person - Player - Team - Couple
Expert 13 (0.86)	**Expert 14 (0.90)**	**Expert 15 (0.95)**	**Expert 16 (0.98)**
Room - House	Art - Artist - Digital	Food - Water	Person - Man - Woman - Text

Table 4: The most frequent words in prompts assigned to each expert of APTP-Base pruned on COCO. The resource utilization of each expert is indicated in parentheses.

Expert 1 (0.65, Indoor Scenes and Dining)	**Expert 2 (0.77, Food and Small Groups)**
table - plate - kitchen - sitting	food - pizza - sandwich
Expert 3 (0.78, People and Objects)	**Expert 4 (0.79, Sports and Activities)**
skateboard - surfboard - laptop - tie - phone	tennis - baseball - racquet - skateboard - skis
Expert 5 (0.79, Wildlife and Nature)	**Expert 6 (0.80, Urban Scenes and Transportation)**
giraffe - herd - sheep - zebra - elephants	street - train - bus - park - building
Expert 7 (0.81, Outdoor Activities and Nature)	**Expert 8 (0.83, Domestic Life and Pets)**
beach - ocean - surfboard - kite - wave	man - woman - girl - hand - bed - cat

- **High and Low resource samples**

Fig. 5: Comparison of samples generated by low and high budget experts

Ablations

- **Effect of APTP components**

Table 3: Ablation results of APTP's components on 30k samples from MS-COCO (Lin et al., 2014) validation set. We fine-tune all models for 10k iterations after pruning.

Method	MACs(@768)	Latency(@768)	FID (↓)	Clip Score (↑)	CMMD (↓)
Uni-Arch Baseline	1088.8G	3.1	46.56	29.11	0.91
Contrastive Router	1079.5G	3.1	48.78	28.90	0.92
+ Optimal Transport	1076.6G	3.1	38.56	30.07	0.74
+ Distillation (**APTP**)	1076.6G	3.1	25.57	31.13	0.58

- APTP generalizes to various styles even if they are not present in the fine-tuning dataset.

Fig. 6: APTP generalizes to styles not seen in fine-tuning.

Thank You!