Pluggable Pruning with Contiguous Layer Distillation for Diffusion Transformers

Supplementary Material

Appendix 1 presents a more detailed experimental setup.

Appendix 2 provides a detailed breakdown of sub-metrics for all strategies and base models across each benchmark.

Appendix 3 presents subjective results of Qwen-Image pruned by PPCL.

Appendix 4 presents the subjective results of Qwen-Image-Edit pruned by PPCL.

Appendix 5 discusses the limitations of PPCL and future work.

1. Detailed Experimental Settings

Linear Probing. For the training of linear probes, we set the number of training steps to 2000 and use the same training set as described in Sec.4.1. We select the Chinese subset of LongTextBench as the calibration set, which consists of 160 samples. When executing Eq.3, we perform the calculation for each sample and each timestep, then average the results over all samples and timesteps to obtain the final CKA value. For the computation of CKA similarity, we first center and normalize the input features, compute and center the Gram matrices of the linear kernels, then calculate the Hilbert–Schmidt Independence Criterion (HSIC) between the Gram matrices, and finally normalize the result to obtain the CKA similarity. For Qwen-Image, the redundant intervals identified are:

$$\begin{aligned}\mathcal{I}_Q =&\{[3,4],[5,7],[8,10],[11,12],[15,24],[25,27],[29,30],\\ &[42,43],[45,47],[48,49],[52,53],[54,55],[56,57]\}.\end{aligned} \tag{1}$$

As shown in Algorithm 1, we provide a concrete algorithmic example for the simulation process based on Qwen-Image, along with detailed explanation. In addition, the redundant intervals detected by baseline, LP-a, and LP-b in the ablation study (Sec.4.3) are as follows:

$$\begin{aligned}\mathcal{I}_Q^{baseline} =&\{[5,7],[8,10],[11,13],[16,18],[19,21],[22,24],[25,27],\\ &[38,40],[42,44],[45,47],[48,50],[52,53],[54,55],\\ &[56,57]\}.\end{aligned} \tag{2}$$

$$\begin{aligned}\mathcal{I}_Q^{a} =&\{[5,7],[8,10],[11,14],[16,17],[19,21],[22,24],[25,28],\\ &[29,30],[42,44],[45,47],[48,50],[52,54],[56,57]\}.\end{aligned} \tag{3}$$

$$\begin{aligned}\mathcal{I}_Q^{b} =&\{[3,4],[5,9],[8,10],[11,13],[15,24],[25,28],\\ &[29,31],[42,44]\}.\end{aligned} \tag{4}$$

Similarly, for FLUX.1-dev, the redundant intervals detected by the linear probing for its single-stream and dual-stream layers are as follows:

$$\mathcal{I}_F^{double} =\{[3,5],[6,8],[9,11],[12,14],[15,16],[17,18]\}, \tag{5}$$

$$\begin{aligned}\mathcal{I}_F^{single} =&\{[0,1],[2,3],[4,5],[6,7],[8,9],[10,11],[12,13],[18,19],\\ &[20,21],[22,23],[24,25],[26,27],[28,29],[30,31]\},\end{aligned} \tag{6}$$

Width-wise Pruning. We employ linear projectors to replace components in the text streams and FFN modules. For text stream pruning, we identify layers with CKA similarity ≥ 0.999 and prioritize deeper layers, thereby determining the pruned layers. For FFN pruning, we compute the similarity of output features before and after replacing each FFN with a linear projector, and select the top-3 layers with highest similarity for pruning. Each linear projector is a linear layer, maintaining the same input and output dimensions (3072) as the original components, which results in a significant reduction in parameter count compared to the original FFNs and text stream components.

Algorithm 1 Illustrative Example of the Simulation Process on Qwen-Image

Input:
Teacher model $\mathcal{T}$ with M MMDiT blocks, calibration set X, trained linear probes $\mathcal{L}'_p = \{l_i\}_{i=1}^{M}$.
Output:
Redundant intervals of Qwen-Image $\mathcal{I}_Q$

$\mathcal{I}_Q \leftarrow \emptyset, u \leftarrow 3$ ▷ *Assume the starting point $u = 3$*
while $u \leq M$ **do**
  $\Delta(u,k) \leftarrow$ CalculateDelta$(\mathcal{T}, \mathcal{L}'_p, X, u, k), \forall k \in [u+1, M]$(Eq.4)
  for $k \in \{u+2, \ldots, M\}$ **do** (Eq.5)
    if $\Delta(u,k) > \Delta(u,k-1)$ ▷ *When $k = 5$, the conditional statement triggers*
      $v \leftarrow k-1$ ▷ $v \leftarrow 4$
      break
    end if
  end for
  $\mathcal{I}_Q \leftarrow \mathcal{I}_Q \cup \{[u,v]\}$ ▷ *The interval $[3,4]$ is included in $\mathcal{I}_Q$*
  $u \leftarrow v+1$ ▷ *$u \leftarrow 5$ and continue searching*
end while
return $\mathcal{I}_Q$

2. Detailed Experiment Results

DPG. As shown in Tab.1, PPCL delivers robust performance across the dimensions of relation modeling (Relation), global semantic understanding (Global), and entity generation (Entity). Compared to three base models, it shows less than 3% performance degradation across all these dimensions, demonstrating that the pruned model preserves strong comprehension and generation capabilities for global and relational semantics. Notably, when applied to FLUX.1-dev, PPCL even enhances performance in Global dimension. Although PPCL exhibits a slightly larger performance drop on the attribute understanding and generation (Attribute) with FLUX.1-dev compared to Chroma1-HD and FLUX.1 Lite, it still maintains a competitive advantage over TinyFusion and HierarchicalPrune. Moreover, on

Models	Methods	DPG↑							GenEval↑							
		Global	Entity	Attribute	Relation	Other	Overall	R.(%)	Single Object	Two Object	Counting	Colors	Position	Attribute Binding	Overall	R.(%)
FLUX.1-dev	Base model	74.3	90.0	88.9	90.8	88.3	83.8	0	0.980	0.810	0.740	0.790	0.220	0.450	0.665	0
	FLUX.1 Lite	88.8	87.9	87.2	89.8	88.9	82.1	2.03	0.987	0.747	0.600	0.798	0.160	0.440	0.623	6.32
	Chroma1-HD	84.9	89.5	88.4	91.4	91.6	84.0	-0.238	0.962	0.717	0.462	0.787	0.240	0.390	0.593	10.8
	TinyFusion	80.2	85.6	83.6	84.7	86.7	77.2	7.88	0.950	0.524	0.450	0.681	0.181	0.283	0.511	23.2
	HierarchicalPrune	78.6	84.6	85.3	83.8	84.6	75.7	9.67	0.850	0.490	0.510	0.690	0.170	0.310	0.503	24.4
	PPCL(8B)	85.1	87.9	85.6	89.8	87.8	80.0	4.53	0.978	0.726	0.593	0.785	0.170	0.380	0.605	9.02
FLUX.1 Lite	Base model	88.8	87.9	87.2	89.8	88.9	82.1	0	0.987	0.747	0.600	0.798	0.160	0.440	0.623	0
	PPCL(6.5B)	87.7	87.6	86.7	88.7	86.5	81.2	1.09	0.968	0.715	0.530	0.784	0.140	0.420	0.593	4.81
Qwen-Image	Base model	91.3	91.6	92.0	94.3	92.7	88.9	0	0.990	0.920	0.890	0.880	0.760	0.770	0.870	0
	TinyFusion	80.3	87.9	87.6	89.0	85.9	80.7	9.22	0.987	0.869	0.762	0.819	0.430	0.570	0.739	15.1
	HierarchicalPrune	82.1	91.6	88.7	90.8	84.3	83.3	6.30	0.975	0.919	0.812	0.840	0.430	0.620	0.766	12.0
	PPCL(14B)	90.1	90.0	91.5	93.6	91.0	87.9	1.12	0.990	0.925	0.879	0.874	0.650	0.765	0.847	2.64
	PPCL(12B)	87.7	88.8	89.9	92.7	89.2	83.6	5.96	0.968	0.900	0.856	0.852	0.543	0.690	0.801	7.93
	PPCL(10B)	85.0	86.8	85.6	90.5	87.3	81.7	8.09	0.968	0.885	0.822	0.840	0.521	0.670	0.784	9.88
	PPCL(10B Finetune)	89.7	89.0	89.9	92.8	89.6	86.7	2.47	0.975	0.920	0.850	0.904	0.560	0.760	0.828	4.82

Table 1. Detailed experimental results on DPG and GenEval.

Models	Methods	OneIG-EN↑						
		Alignment	Text	Reasoning	Style	Diversity	Overall	R.(%)
FLUX.1-dev	Base model	0.786	0.523	0.253	0.368	0.238	0.434	0
	FLUX.1 Lite	0.514	0.481	0.238	0.374	0.242	0.378	12.9
	Chroma1-HD	0.811	0.696	0.250	0.362	0.333	0.490	-12.9
	TinyFusion	0.476	0.470	0.202	0.329	0.247	0.345	20.5
	HierarchicalPrune	0.483	0.469	0.191	0.362	0.251	0.351	19.1
	PPCL(8B)	0.753	0.631	0.217	0.361	0.316	0.456	-5.07
FLUX.1 Lite	Base model	0.514	0.481	0.238	0.374	0.242	0.378	0
	PPCL(6.5B)	0.493	0.480	0.219	0.351	0.249	0.371	1.85
Qwen-Image	Base model	0.882	0.891	0.306	0.418	0.197	0.539	0
	TinyFusion	0.824	0.845	0.251	0.389	0.186	0.500	7.24
	HierarchicalPrune	0.836	0.865	0.246	0.396	0.201	0.509	5.57
	PPCL(14B)	0.880	0.886	0.295	0.413	0.205	0.536	0.556
	PPCL(12B)	0.866	0.885	0.282	0.396	0.157	0.517	4.08
	PPCL(10B)	0.839	0.860	0.249	0.359	0.121	0.485	10.0
	PPCL(10B Finetune)	0.854	0.878	0.268	0.365	0.130	0.499	7.42

Table 2. Detailed experimental results on OneIG-EN.

Models	Methods	OneIG-ZH↑						
		Alignment	Text	Reasoning	Style	Diversity	Overall	R.(%)
Qwen-Image	Base model	0.825	0.963	0.267	0.405	0.279	0.548	0
	TinyFusion	0.785	0.889	0.224	0.365	0.228	0.498	9.12
	HierarchicalPrune	0.796	0.895	0.214	0.349	0.224	0.496	9.49
	PPCL(14B)	0.818	0.958	0.241	0.395	0.265	0.538	1.82
	PPCL(12B)	0.805	0.961	0.270	0.383	0.219	0.531	3.10
	PPCL(10B)	0.798	0.921	0.250	0.362	0.175	0.501	8.57
	PPCL(10B Finetune)	0.819	0.937	0.265	0.376	0.196	0.519	5.29

Table 3. Detailed experimental results on OneIG-ZH.

Models	Methods	B-VQA↑	UniDet↑			S-CoT↑	Overall	R.(%)
			spatial	3d_spatial	numeracy			
FLUX.1-dev	Base model	0.640	0.308	0.380	0.602	0.786	0.543	0
	FLUX.1 Lite	0.547	0.292	0.368	0.574	0.772	0.510	6.00
	Chroma1-HD	0.621	0.234	0.310	0.472	0.764	0.480	11.6
	TinyFusion	0.584	0.246	0.331	0.530	0.742	0.486	10.4
	HierarchicalPrune	0.579	0.259	0.320	0.536	0.750	0.489	10.0
	PPCL(8B)	0.615	0.265	0.359	0.550	0.782	0.514	5.33
FLUX.1 Lite	Base model	0.547	0.292	0.368	0.574	0.772	0.510	0
	PPCL(6.5B)	0.581	0.276	0.363	0.555	0.769	0.509	0.352
Qwen-Image	Base model	0.709	0.428	0.453	0.716	0.825	0.626	0
	TinyFusion	0.689	0.354	0.375	0.662	0.790	0.574	8.34
	HierarchicalPrune	0.706	0.386	0.395	0.6801	0.800	0.593	5.23
	PPCL(14B)	0.750	0.412	0.453	0.707	0.822	0.629	-0.415
	PPCL(12B)	0.733	0.435	0.450	0.702	0.820	0.628	-0.287
	PPCL(10B)	0.701	0.390	0.429	0.678	0.792	0.598	4.50
	PPCL(10B Finetune)	0.715	0.413	0.442	0.691	0.813	0.615	1.82

Table 4. Detailed experimental results on T2I-CompBench. To compute the overall score, we scale the S-CoT score by a factor of 0.01.

Qwen-Image, the performance degradation in the Attribute dimension is only 0.54%.

GenEval. As shown in Tab.1, PPCL demonstrates low performance degradation in single object generation (Single Object) and colors rendering (Colors), with reductions below 0.7%. In more challenging tasks including counting understanding (Counting), two object generation (Two Object), and position understanding (Position), most methods exhibit noticeable performance degradation on FLUX.1-dev. Nevertheless, PPCL maintains the second-highest performance across these dimensions, demonstrating its ability to maintain stable performance under complex generative instructions and effectively mitigate the degradation of key generation capabilities. Furthermore, when applied to Qwen-Image, PPCL achieves an overall performance drop of only 2.64%, remaining highly competitive across all dimensions.

OneIG. As shown in Tab.2 and Tab.3, PPCL shows only a 1.82% performance drop on the Chinese subset (OneIG-ZH), while on the English subset (OneIG-EN), it achieves a 5.07% overall improvement when pruning FLUX.1-dev. PPCL performs robustly in text rendering (Text) and stylization (Style), with degradation below 2%, demonstrating its ability to preserve core textual and stylistic features. Notably, for FLUX.1-dev, PPCL even improves text rendering performance by 20%. In generative diversity (Diversity), PPCL shows significant gains on OneIG-EN but a slight drop on OneIG-ZH, suggesting language-dependent effects while remaining controllable. For semantic alignment (alignment) and reasoning (Reasoning), although PPCL exhibits a relatively larger performance drop

on FLUX.1-dev, it still outperforms TinyFusion and HierarchicalPrune, and maintains strong stability on Qwen-Image.

T2I-CompBench. As shown in Tab.4, On FLUX.1-dev, PPCL achieves the lowest performance degradation of 5.33%, outperforming all other methods. For Qwen-Image, both PPCL (14B) and PPCL (12B) surpass the base model in overall performance. PPCL demonstrates superior capabilities in visual question answering (B-VQA) and reasoning (S-CoT) dimensions, with minimal performance degradation or even improvements, indicating strong retention of complex reasoning and multimodal alignment. In spatial understanding, 3D spatial understanding, and numeracy understanding, PPCL outperforms most comparative methods on FLUX.1-dev, further validating its well-rounded and stable performance preservation. Notably, when applied to Qwen-Image, PPCL achieves nearly lossless performance in 3D spatial understanding.

3. Subjective Results on Qwen-Image

As shown in Fig. 1, we conduct a subjective comparison among multiple pruning methods and the teacher model. We apply HierarchicalPrune and TinyFusion separately to prune the Qwen-Image model down to 14B parameters. The results generated by HierarchicalPrune contain some visual artifacts. A possible reason is that the process of identifying and removing unimportant layers using HPP is somewhat coarse. Moreover, layer importance patterns do not strictly follow the trend where importance decreases with increasing depth. As for TinyFusion, its overall semantic alignment performance is slightly insufficient. Overall, our method demonstrates certain advantages in generation fidelity, visual quality, and text-image alignment, achieving superior results.

Figs.2-5 present additional subjective results of Qwen-Image pruned by PPCL. PPCL exhibits robust multilingual support, accurately rendering both Chinese and English text across diverse visual contexts while maintaining high legibility and stylistic appropriateness. It further demonstrates strong semantic visual alignment, ensuring that generated text content harmonizes seamlessly with scene atmosphere, character expressions, and overall design intent, thereby enhancing narrative coherence and contextual relevance in complex, real world applications. In summary, PPCL demonstrates sufficient capability in text generation quality, visual aesthetics, and semantic consistency to meet practical application requirements, indicating its readiness for large-scale deployment.

4. Subjective Results on Qwen-Image-Edit

Qwen-Image-Edit. In addition to T2I models Qwen-Image and FLUX.1-dev, we also apply PPCL to prune image editing models, including Qwen-Image-Edit and Qwen-Image-Edit-2509. For Qwen-Image-Edit, we obtain a pruned model with 13B parameters. For Qwen-Image-Edit-2509, we obtain two pruned variants with 14B and 13B parameters, respectively. Figs.6 and 7 present subjective results of the pruned Qwen-Image-Edit and Qwen-Image-Edit-2509, respectively.

Figure 1. A subjective comparison result: The first row shows the teacher model. The second and third rows display the 14B pruned HierarchicalPrune and TinyFusion models, respectively. The last row illustrates the effect of our method pruned to 10B.

5. Limitations

PPCL exhibits certain limitations in rendering excessively long text and small-scale text elements. Fig.8 presents the corresponding failure cases. When processing extended text sequences, pruned models may struggle to maintain readability and structural integrity. In scenarios involving dense text blocks or complex layout requirements, the rendered output may exhibit inaccuracies in character representation or present issues with clarity. Similarly, for small-scale text rendering, the models sometimes fail to accurately generate fine-grained textual details, particularly in low-resolution regions or when text is placed within visually complex environments. These limitations highlight crucial directions

Figure 2. Subjective results of PPCL (12B) with Chinese instructions.

for future improvements in the text rendering capabilities of pruned models.

Furthermore, two key limitations remain that warrant further investigation. First, the redundant intervals detection based on the first-order difference of CKA similarity lacks rigorous theoretical grounding and functions primarily as an empirical heuristic. Although effective in practice, it relies on observing local minima in CKA similarity gradients without formal proof that these points correspond to optimal pruning thresholds, leaving room for instability across diverse model architectures and datasets. Second, INT4 quantization yields suboptimal performance due to the fundamental conflict between pruning and quantization. Pruning reduces network redundancy by eliminating less critical parameters, which simultaneously narrows the quantization fault-tolerant space by concentrating parame-

Figure 3. Subjective results of PPCL (10B) with Chinese instructions.

ter distributions into narrower ranges. This makes the model significantly more sensitive to quantization error, as the coarse 16-level discretization of INT4 struggles to capture the refined parameter distributions post-pruning, resulting in substantial performance degradation. Future work will focus on developing theoretically grounded strategies for detecting redundant intervals and exploring adaptive quantization strategies that account for pruning-induced structural changes to improve INT4 efficacy.

Figure 4. Subjective results of PPCL (12B) with English instructions.

Figure 5. Subjective results of PPCL (10B) with English instructions.

Modify it into a digital illustration style.

Replace the characters ‘招财进宝' within the circle with ‘财源广进’.

Add a large diamond ring on the finger, making it abstract, exaggerated, and funny.

Write "抬高80公分" on the underline.

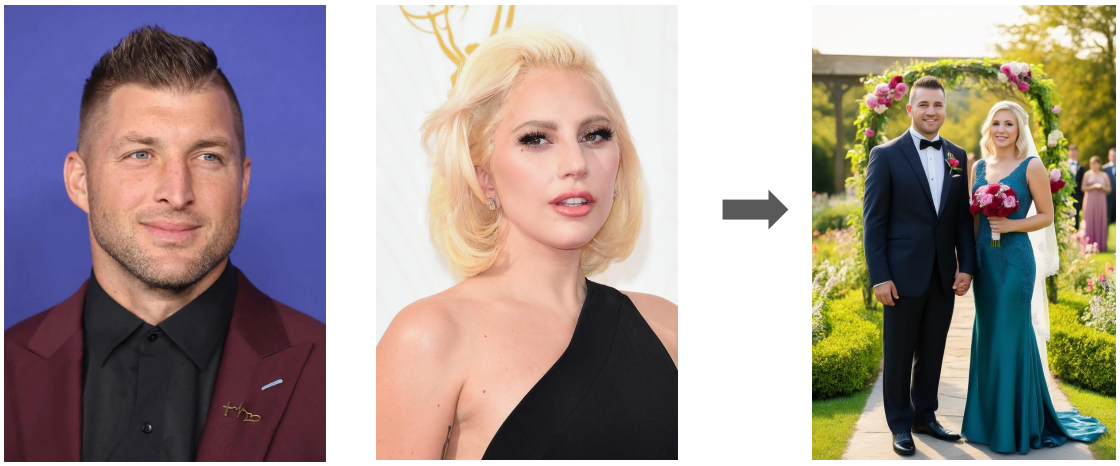
Switch the image to a minimalist aesthetic.

Remove all the people.

Place a wine glass in the hand.

Figure 6. Subjective results of the pruned Qwen-Image-Edit.

Let the ancient costume beauty in the second picture sit on the sofa in the first picture.



Generate a set of wedding photos based on the man in the first picture and the woman in the second picture.

Let the seagulls in the first picture perch on the shoulder of the woman in the second picture.

Figure 7. Subjective results of the pruned Qwen-Image-Edit-2509.

A tea shop filled with organized shelves displaying jars containing various tea blends. Each tea jar has elegantly handwritten rectangular labels neatly placed at the center front; the prominently visible label reads "Summer Peach White Tea", with slightly smaller descriptive text directly underneath stating "Light floral notes with peach flavor". Surrounding jars feature similarly styled labels arranged uniformly, including "Chamomile Lavender Herbal Blend" label. The textual contents on each jar's label are carefully aligned: tea names appear in larger and emphasized handwriting on the top, and concise descriptive phrases appear neatly beneath, creating a visually appealing and easily readable layout that attracts and informs visitors clearly.

展示一个在柔和灯光下展示的金色文物，背景有大理石柱子。海报标题写着“时光珍宝展览。”副标题邀请“探索古代文明的遗迹。”描述部分解释“体验5000年的历史，发现稀有文物、互动展览和导览游。”底部文本显示“大博物馆展厅 | 5月1日 - 9月30日 | www.museumtickets.com.”

A dramatic nature documentary scene, showing an underwater close-up of a coral reef teeming with marine life. Brightly-colored corals in shades of orange, purple, and turquoise create a visually striking foreground, surrounded by schools of small tropical fish darting between the intricate structures. Gentle rays of sunlight cast dynamic shadows and reflections on surrounding marine plants and animals, further enhancing the vivid and fluid underwater atmosphere. At the bottom central portion of the screen, a semi-transparent dark rectangular caption box displays white textual subtitles neatly arranged into clear sentences: "Coral reefs support nearly a quarter of all ocean life, yet they face severe threats from climate change and pollution. Protecting coral ecosystems is vital for preserving marine biodiversity and sustaining healthy oceans for future generations". The elegant, documentary-style typography ensures the text remains perfectly legible against the vibrant background imagery, creating an informative yet visually engaging presentation.

一段与健康相关的新闻片段截图，画面聚焦于现代实验室中科学家们正在研究疫苗的特写镜头。背景可见整齐排列的科学标记瓶罐货架，数字屏幕在后方显示着医学数据和图表。画面中心底部位置设有半透明蓝色横版标题栏，使用清晰易读的白色新闻字体标注着：“医学突破：科学家宣布在研制一种有前景的通用流感疫苗方面取得进展，有望提供长期免疫力并减少全球年度流感爆发。临床试验预计将于今年晚些时候开始”。该标题栏左角处配有新闻台标志及“直播”标志，强调对这项重要医学进展的实时报道。

Figure 8. Some failure cases.