Appendix

A. Additional Efficiency Analysis

Memory overhead. Our asymmetric attention mechanism decouples condition encoding from the generation process. Unlike standard controllable DiTs that concatenate condition tokens with noisy tokens (increasing peak memory proportionally), OminiControl2 processes condition features independently and caches them for reuse. As shown in Table 3, the peak VRAM overhead introduced by conditioning is negligible regardless of resolution or number of conditions.

Method	512 (1 Cond)	512 (4 Conds)	1024 (1 Cond)	1024 (4 Conds)
OminiControl	+0.59	+2.36	+2.37	+9.51
OminiControl2	**∼ 0**	**∼ 0**	**∼ 0**	**∼ 0**

Table 3. Additional VRAM overhead (GB) introduced by conditioning at different resolutions and condition counts.

Comparison with generic acceleration methods. We compare OminiControl2 against generic acceleration techniques applied to OminiControl on the inpainting task (Table 4). Naïve KV Cache and Token Merging are content-agnostic approaches that do not account for the structure of conditioning inputs, leading to noticeable quality degradation. In contrast, OminiControl2 preserves generation quality while achieving comparable or better speedups. Moreover, our architectural optimizations are orthogonal to system-level techniques such as quantization and compilation, which can be stacked on OminiControl2 for further gains.

Method	FID ↓	MUSIQ ↑	CLIP ↑	NIQE ↓	PSNR ↑	MSE ↓
Naïve KV Cache	16.73	0.296	0.854	4.20	64.96	8572
Token Merging	14.73	0.292	0.780	4.04	69.99	9588
OminiControl2	**11.71**	**0.307**	**0.891**	**3.88**	**71.05**	**6654**

Table 4. Comparison with generic acceleration methods on the inpainting task.

B. Generalization to Other Model Families

To demonstrate that our approach generalizes beyond FLUX, we evaluate OminiControl2 on additional diffusion model families.

Quality evaluation on SD3.5 Medium. Table 5 reports results on SD3.5 Medium for the Canny-to-image task at 1024×1024 resolution. OminiControl2 achieves a $2.1\times$ speedup while maintaining competitive generation quality, confirming that our compact token representation and feature reuse mechanisms transfer effectively to other DiT architectures.

Figure 10. Qualitative comparison across conditioning tasks. OminiControl2 preserves the generation quality of OminiControl while being significantly faster.

Base Model	Method	FID ↓	MUSIQ ↑	CLIP ↑	SSIM ↑	F1 ↑	Latency ↓
SD3.5 Medium	OminiControl	21.5	73.8	0.770	0.377	0.386	7.63 (1×)
	OminiControl2	22.1	74.2	0.762	0.363	0.388	3.72 (2.1×)

Table 5. Generalization on SD3.5 Medium (Canny-to-image, 1024×1024).

Efficiency across model families. We further validate the efficiency gains of our feature reuse strategy on several additional model families. As shown in Table 6, OminiControl2 consistently delivers 1.76–$1.81\times$ speedups across SD3.5-Large, Flux Kontext, and Qwen Image Edit, demonstrating broad applicability.

Model	Original	+ OminiControl2	Speedup
SD3.5-Large	14.5s	8.0s	1.81×
Flux Kontext	15.11s	8.36s	1.80×
Qwen Image Edit	19.20s	10.92s	1.76×

Table 6. Efficiency gains on different model families.

Qualitative comparison. Figure 10 provides qualitative comparisons between OminiControl and OminiControl2 across different conditioning tasks. OminiControl2 maintains the same level of structural fidelity and control accuracy as OminiControl while delivering a $\sim 1.8\times$ speedup.

C. Discussion on Token Compression

While we adopt task-specific token compression strategies in the main paper, we explore a more unified principle based on *information density*. The key insight is that tokens with higher information density—those that contribute more to the conditioning signal—should be preferentially retained during compression.

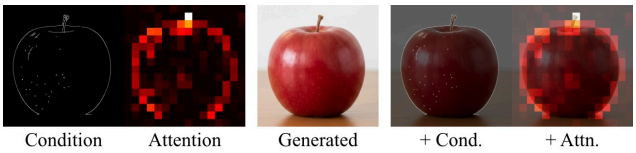

Figure 11. Attention map analysis for token compression. Attention maps correlate with semantically meaningful regions, suggesting a unified compression principle based on information density.

To validate this, we conducted a feasibility analysis using attention maps as a proxy for information density. Figure 11 visualizes the correlation between attention maps and semantically meaningful regions. The attention maps naturally highlight informative regions of the condition image, suggesting that attention-based token selection provides a principled and generalizable strategy for extending compact token representation to new conditioning tasks without requiring task-specific designs.