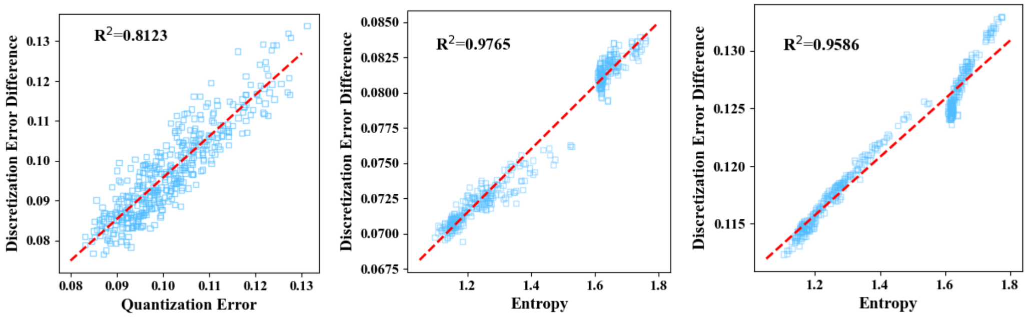

(a) (b) (c)

Figure 5: (a) The correlation between discretization error difference (DED) and the quantization errors in 15th layer. (b) The correlation between DED and the entropy in 5th layer and (c) in 25th layer.

Model	Dependency Proxy	W6A6		W4A4	
		Accuracy	Search Cost	Accuracy	Search Cost
LLaVA-7B	Quantization Errors	88.59	43.7	77.97	41.2
	Entropy	88.95	26.9	78.66	25.9
LLaVA-13B	Quantization Errors	88.96	58.6	78.82	54.2
	Entropy	89.04	33.1	79.89	31.8

Table 5: Comparisons with different proxy for mining cross-layer dependency for LLaVA-v1.3 models in ScienceQA dataset across bitwidth setting.

A Why leveraging entropy as proxy:

The benefit and motivation for using entropy rather than quantization errors as a proxy for block-wise searches lie in several key considerations. We analyze that larger entropy indicates more homogeneous data distribution, which is a well-established principle in information theory. Consequently, DED and activation entropy are strongly correlated with an value of 0.97. However, greater quantization error does not necessarily imply more homogeneous data distribution and does not show a positive correlation with DED, having an value of 0.81, which is empirically verified in the figure 5.

Meanwhile, the search cost of quantization errors doubles compared with entropy as a proxy, as the calculation of quantization errors requires multiple forward passes for both the FP model and the quantized model. The weak correlation and the unbearable search cost render quantization error unsuitable as a metric for measuring cross-layer dependency.

Furthermore, we conducted experiments comparing the proxy effectiveness of quantization error and entropy across different models under various bitwidths in Table 5. Entropy outperformed quantization errors by a significant margin (78.66 vs. 77.97), showing a strong cross-layer dependency within each block. This allowed us to achieve optimal block partitioning by mining the cross-layer dependency.

B Performance on more baseline methods

We have extended our experiments to an additional baseline method ZeroQuant-V2[47] and compared it against our proposed methods in Table 6. ZeroQuant-V2 leverages per-token quantization with different rounding functions to minimizing activation discretization errors. However, ignoring cross-layer dependency of discretization errors fails to acquire the optimal rounding strategy with severe outliers under low bitwidth and degrades the performance significantly. On the contrary, our Q-VLM mines the cross-layer dependency of output distribution across layers, minimizing the block-wise discretization errors to avoid suboptimal quantization. We further optimize the visual encoder to disentangle the cross-layer dependency for fine-grained search space decomposition. As a

Model	Quantization Method	W8A8		W4A4	
		Accuracy	Inference Time	Accuracy	Inference Time
LLaVA-7B	ZeroQuant-V2	89.04	10.7h	78.08	7.3h
	Q-VLM	89.58	8.3h	79.79	6.1h
LLaVA-13B	ZeroQuant-V2	89.13	12.6h	78.81	9.7h
	Q-VLM	89.81	11.2h	80.78	8.9h

Table 6: Comparisons with different quantization methods for 7B and 13B models across W6A6 and W4A4 bitwidth settings.

Dataset	Shots	FP	8bit		4bit	
			Q-LoRA	Q-VLM	Q-LoRA	Q-VLM
Vizwiz	0	23.79	21.24	21.47	17.62	18.69
	4	27.05	25.83	26.59	24.17	24.55
	32	39.76	36.38	37.60	31.64	35.52
Hateful Memes	0	50.23	47.75	49.12	43.86	44.22
	4	50.10	48.62	49.55	45.12	45.26
	32	50.27	50.02	51.05	45.76	47.84

Table 7: Performance comparison on Vizwiz and Hateful Memes datasets across FP, 8bit, and 4bit quantization methods with different shot settings.

result, our method outperforms ZeroQuant-V2 by 1.71 (79.79 vs. 78.08) in answering accuracy on ScienceQA dataset under 4-bit in LLaVA-7B model. Additionally, our method enhances inference speed, exceeding ZeroQuant-V2 by 1.2h (6.1h vs. 7.3h) due to utilizing stored rounding parameters instead of dynamic per-token quantization. The additional baseline provides a more comprehensive evaluation framework to highlight the strengths of our approach.

C Performance on other multi-modal architectures

We also explored the multi-modal architecture OpenFlamingo[3] to ensure the robustness and generalizability of our methods 7. We deploy our method on OpenFlamingo 3B model using Vizwiz and Hateful Memes[22] datasets, selecting bitwidths of 4 and 8 for quantized layers. Q-VLM designed in LLaVA-like architectures can be effectively adapted to cross-attention based VLMs due to the consistent core mechanism of cross-attention and the robust multimodal alignment capabilities pre-trained on large-scale vision-language pairs. Since OpenFlamingo is a cross-attention based VLM, exploiting cross-layer dependency is particularly suitable. Our method outperforms Q-LoRA by 1.22 (37.60 vs. 36.38) under 8-bit in OpenFlamingo-3B model. The advantage of our method becomes more obvious for 4-bit 3B LVLMs because quantization errors and cross-layer dependency play a more significant role in networks with low capacity. These results underscore the robustness and generalizability of our approach across different tasks, model architectures and datasets, demonstrating its effectiveness in diverse scenarios.