Appendix A. The Fusion Operator Set Design

As a baseline, we considered the fusion search space of recent works MFAS (ConcatFC) and BM-NAS (Attention, LinearGLU, Sum) to design our initial $\mathcal{FP}$ set. We conducted a preliminary analysis and included other fusion operators, notably, ConcatMish and Squeeze − Excitation, yielding better accuracy and efficiency tradeoffs. Consequently, these new added fusion operators have also contributed to the superior results of our multimodal models over SoTA methods.

Table 1: Backbones configurations of one of our optimal MM-NNs on the AV-MNIST dataset.

Modality	Kernel size	Expand ratio	Depth	Acc (%)	Lat (ms)	Ergy (mJ)
Image	[5, 5, 5, 7]	[3, 6, 4, 3]	2	82.66	4.02	6.01
Audio	[3, 3, 7, 5]	[3, 3, 3, 6]	2	85.55	3.95	5.71

To better understand the impact of the fusion operator, we examine one of our optimal MM-NN models on the AV-MNIST dataset (Table 4 of the paper, TX2, Acc=95.33%, Lat=9.11ms, Ergy=13.88mJ). The backbones configurations for image and audio modalities are summarized in Table 1.

Table 2: The impact of the fusion operator on the MM-NN performance on AV-MNIST.

Fusion operator	Acc (%)	Lat (ms)	Ergy (mJ)
***Searchable (Ours) (ConcatMish, ConcatFC)**	**95.33**	**9.11**	**13.88**
Sum	93.70	8.09	12.32
Attention	89.55	8.98	13.13
LinearGLU	94.22	9.03	14.14
ConcatFC	94.66	9.02	13.77

In this ablation study, we maintain the unimodal backbones and optimal found fusion macro-architecture (i.e., number of fusion cells and nodes) by our *first-stage* optimization engine and only vary the fusion operators. The results are reported for the AV-MNIST dataset in Table 2. As shown, the contradictory nature of objectives is explicit as more accuracy yields high latency and energy. Notably, our newly added fusion operator, ConcatMish with the existing ConcatFC, depict the optimal trade-off.

Table 3: Backbones configurations of one of our optimal MM-NNs on the Memes-P dataset.

Modality	Configuration			Acc (%)	Lat (ms)	Ergy (mJ)
Text	Maxout={hidden_features: 128, n_blocks: 2, factor_multiplier: 2}			83.38	1.08	1.21
	Kernel size	**Expand ratio**	**Depth**			
Image	[3,3,3,3,5,3,3,3,5,5,3,5]	[4,3,4,6,4,4,3,6,6,6,3,4]	[2,3,2]	85.91	10.94	25.44

Similarly, on the Memes-P dataset, we conducted a the same analysis on one of our optimal MM-NN (See Table 6 of the paper, TX2, Acc=90.42%, Lat=12.47ms, Ergy=31.92mJ). The backbones configurations for text and image modalities are summarized in Table 3.

Table 4: The impact of the fusion operator on the MM-NN performance on Memes-P

Fusion operator	**Acc (%)**	**Lat (ms)**	**Ergy (mJ)**
***Searchable (Ours) (Sum,Squeeze-Excitation)**	**90.42**	**12.47**	**31.92**
Sum	88.73	12.38	28.44
Attention	89.01	15.04	30.86
LinearGLU	89.29	15.18	33.89
ConcatFC	89.29	15.16	32.78

As shown in Table The newly added $\mathrm{Squeeze}-\mathrm{Excitation}$ operator with the existing Sum yield better results and balance between accuracy, latency, and energy (See Table 4). Thus further demonstrating the fusion operators' diversity across different tasks, modalities, and datasets.

Appendix B. Visualizations of our learned MM-NNs

In the following, we provide visualizations of the learned fusion architectures on various multimodal datasets. We note that our MM-NN models are built upon different backbones that technically share the same macro-architecture -as fixed by the OFA supernet design-. However, as our unimodal backbones are searchable, the inner structure of the neural blocks is different from one MM-NN to another. We refer the reader to Tables 4, 5, and 6 in the main paper for more details on the unimodal backbones performance for each reported multimodal representation. The following MM-NN visualizations are all reported for the NVIDIA Jetson TX2 device.
As depicted in Figure 1, to achieve a latency-efficient MM-NN on the AV-MNIST dataset when deployed on the NVIDIA Jetson TX2, *Harmonic-NAS* could find a tailored fusion design with less hardware demanding fusion operators. Furthermore, as reported in Table 4 of the main paper, the first-stage search of *Harmonic-NAS* has adapted the design of the backbones to less computationally complex and energy-demanding ones. For instance, the obtained backbones for the image and audio modalities yield high TOP-1 accuracy, resulting in rich feature joint embedding and consequently achieving higher accuracy in the context of multimodal fusion.

To further enhance the accuracy of the MM-NN, *Harmonic-NAS* explored more intricate fusion macro architectures as shown in Figure 2. This strategic adaptation of our second-stage search engine has yielded the discovery of accurate MM-NN than SoTA baselines at the cost of increased latency and energy. These findings underscore the capacity of *Harmonic-NAS* in tailoring the design of MM-NNs for resource-constrained hardware devices to adapt to different deployment scenarios and application requirements.

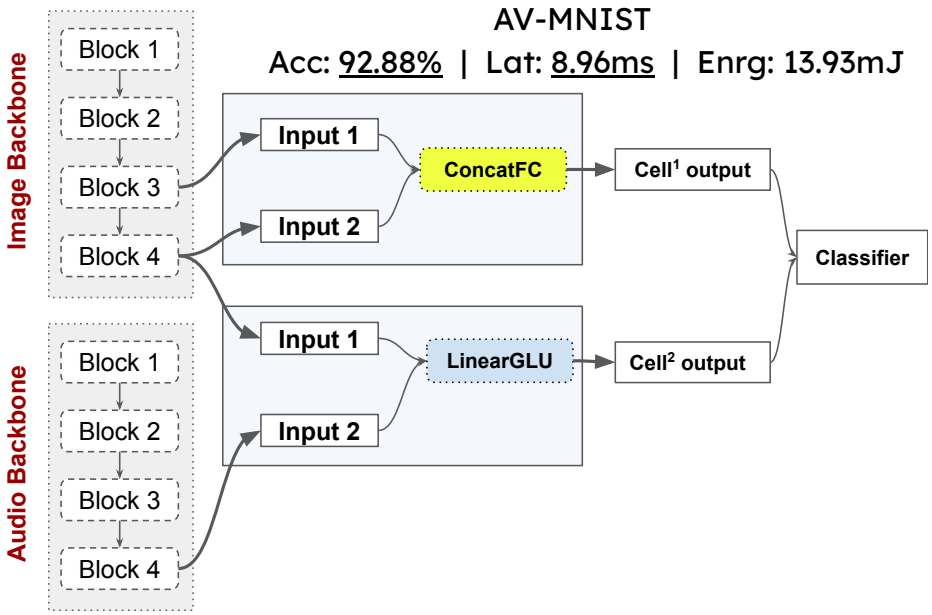

Figure 1: Visualization of the latency-efficient MM-NN for the AV-MNIST dataset on the TX2.

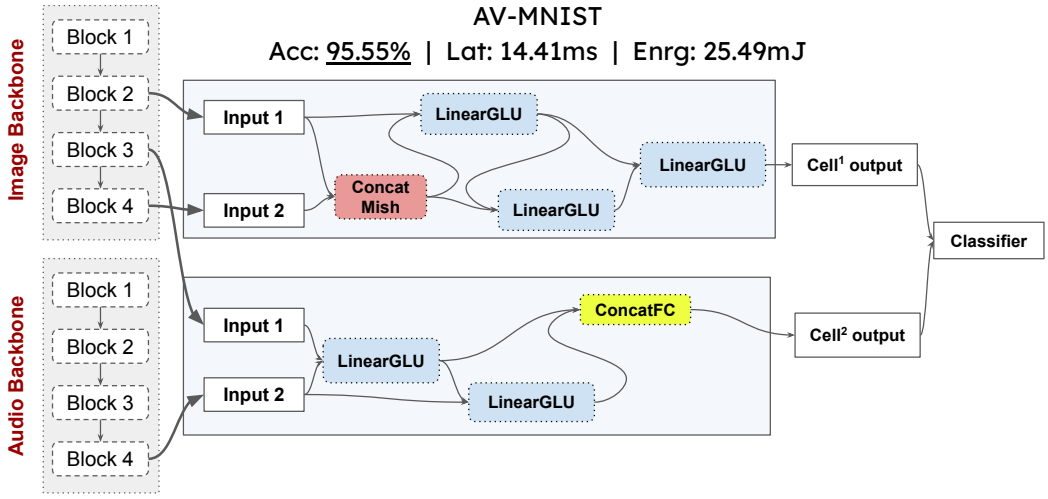

Figure 2: Visualization of the most-accurate MM-NN for the AV-MNIST dataset on the TX2.

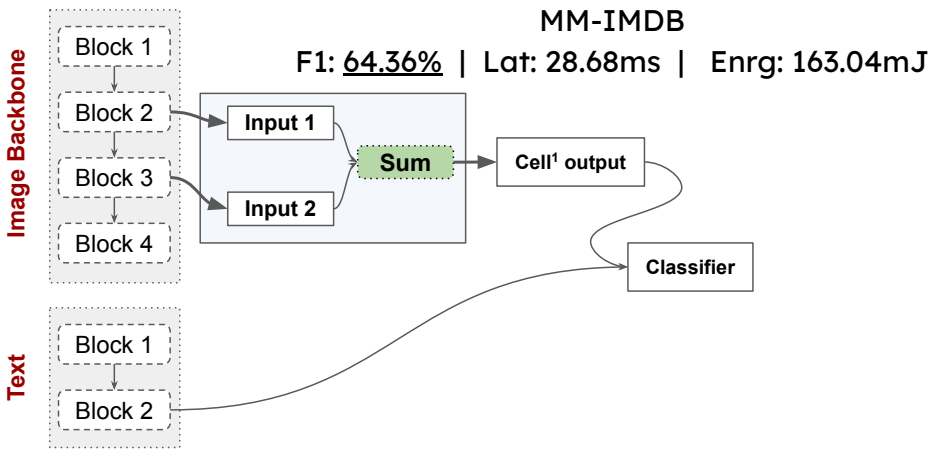

Figure 3: Visualization of the most-accurate MM-NN for the MM-IMDB dataset on the TX2.

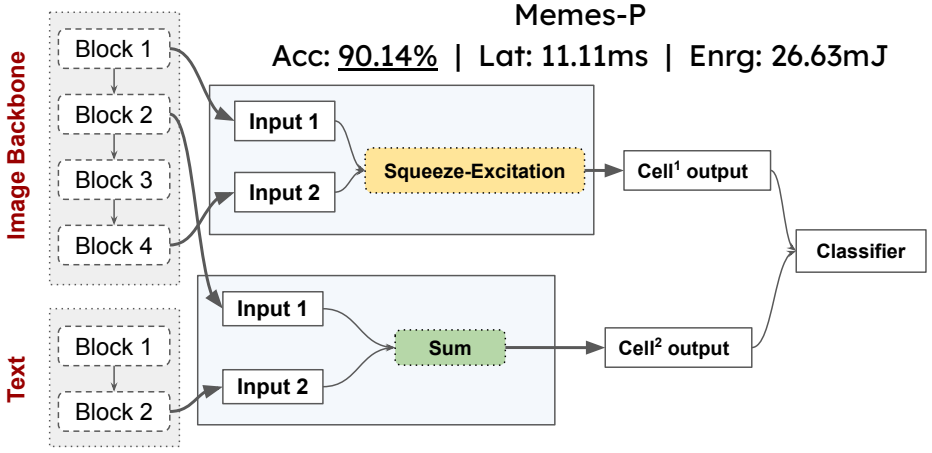

Figure 4: Visualization of the second most-accurate MM-NN for the Memes-P dataset on the TX2.