

Learning to Inference Adaptively for Multimodal Large Language Models

Supplementary Material

In this supplementary material, we provide (1) additional implementation details (see Supp. A); (2) detailed results, including broader benchmarks and other LLM backbones on AdaLLaVA accompanying our experiments in Sec. 4 (see Supp. B); (3) further ablations on design of switches and comparison with naive sampling strategies (see Supp. C); (4) additional qualitative results on latency and content adaptivity (see Supp. D) and (5) further discussion on practical deployment (see Supp. E). We hope that this document will complement our main paper.

For sections, figures and equations, we use numbers (e.g., Sec. 1) to refer to the main paper and capital letters (e.g., Sec. A) to refer to this supplement.

A. Further Implementation Details

Probabilistic execution plan sampling. Recall that in our probabilistic model, we define the distribution $p(\mathbf{s}|\{\mathbf{z}^{v|q}\}, l, \phi)$ via a sampling process. Given the input tokens and a latency budget l , the output of the lightweight scheduler is a logits vector corresponding to the K available switches: $\pi_1, \pi_2, \dots, \pi_K \in \mathbb{R}$, where π_i represents the relative likelihood of selecting the i^{th} switch. The latency budget l allows us to define k , the maximum number of switches allowed to activate. Then, a sampled execution plan from $p(\mathbf{s}|\{\mathbf{z}^{v|q}\}, l, \phi)$ can be uniquely defined by a subset of k distinct elements from $\{1, 2, \dots, K\}$, corresponding to its activated switches. We sample the execution plan by randomly picking k switches one by one, without replacement, following the logits $\{\pi_i\}_{i=1}^K$. The complete sampling procedure is summarized in Algorithm 1, where $\text{Cat}(\Omega, \{\pi_i : i \in \Omega\})$ denotes the categorical distribution of selecting an element from Ω with probabilities parameterized by $\{\eta_i : i \in \Omega\} = \text{Softmax}(\{\pi_i : i \in \Omega\})$. The process ensures that the sampled execution plan adheres to the input budget while maximizing the utilization.

Differentiable sampling with Gumbel-Softmax. Our designed scheduler is difficult to train as it involves a non-differentiable discrete sampling process, which prevents gradients from backpropagate to the scheduler during training. A common workaround involves using a score function estimator [16, 59]; however, this method often suffers from high variance and slow convergence. Instead, we employ Gumbel-Softmax [22], a reparameterization trick for sampling from categorical distribution. In our implementation, the Gumbel-Softmax approximates $\omega \sim \text{Cat}(\Omega, \{\pi_i : i \in \Omega\})$ with a continuous random vector $\tilde{\omega}$:

$$\tilde{\omega} = \text{Softmax}([g_i + \log \eta_i])_{i \in \Omega}, \quad (\text{A})$$

Algorithm 1 Sampling $\mathbf{s} \sim p(\mathbf{s}|\{\mathbf{z}^{v|q}\}, l, \phi)$

Input: Latency budget l , sampling logits $\{\pi_i\}_{i=1}^K$
Output: Sampled binary vector $\mathbf{s} \in \{0, 1\}^K$
 Determine number of selections k based on l
 Initialize available set of switches $\Omega \leftarrow \{1, 2, \dots, K\}$
 Initialize $\mathbf{s} \leftarrow (0, 0, \dots, 0) \in \{0, 1\}^K$
for $i = 1$ to k **do**
 Sample $\omega \sim \text{Cat}(\Omega, \{\pi_i : i \in \Omega\})$
 $\mathbf{s}[\omega] \leftarrow 1$ (activating the chosen switch)
 $\Omega \leftarrow \Omega \setminus \{\omega\}$
end for
return \mathbf{s}

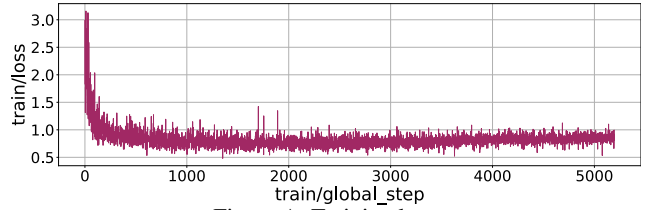


Figure A. Training logs

where each g_i is i.i.d. sample drawn from $\text{Gumbel}(0, 1)$; and $\eta_i = \text{Softmax}(\{\pi_j : j \in \Omega\})[i]$ is the probability of activating the i^{th} switch, computed by the scheduler. Note that $\tilde{\omega}$ is continuous and has a well-defined gradient. To maintain a hard execution plan, we take the one-hot encoding of $\tilde{\omega}$ and apply the straight-through estimator (see [22] for more details).

Training details. Training details were discussed in Sec. 4 of the main paper. Here we show training curve of AdaLLaVA-L-7B with LLaVA 1.5 in Fig. A.

B. Detailed Results

Full results on LLaVA 1.5. We report the full set of results on LLaVA 1.5, LLaVA-PruMerge, LLaVA-PruMerge+ and FastV in Tab. B, as a complement to Tab. 1. All experiments follow the same setting as described in Sec. 4.1. These results confirm that our AdaLLaVA framework successfully adapts to LLaVA 1.5 across different backbone sizes, and can be further combined with recent token selection methods (PruMerge, PruMerge+ and FastV) to further enhance efficiency. We maintain comparable performance while improving efficiency across multiple benchmarks. Additionally, our analysis reveals how performance varies under different latency constraints, demonstrating our framework’s ability to trade between accuracy and latency.

Broader benchmarks. We extend our AdaLLaVA-L

Model	GQA	SEED-Bench	MM-Vet	LLaVa-WILD	VizWiz
LLaVA-1.5-7B	62.0	58.6	31.1	65.4	50.0
AdaLLaVA-L-7B-100%	61.5	60.5	30.7	64.2	54.3
AdaLLaVA-L-7B-85%	61.3	60.2	30.0	62.1	51.5
AdaLLaVA-L-7B-60%	58.7	59.8	23.9	46.3	44.8

Table A. Results on broader benchmarks.

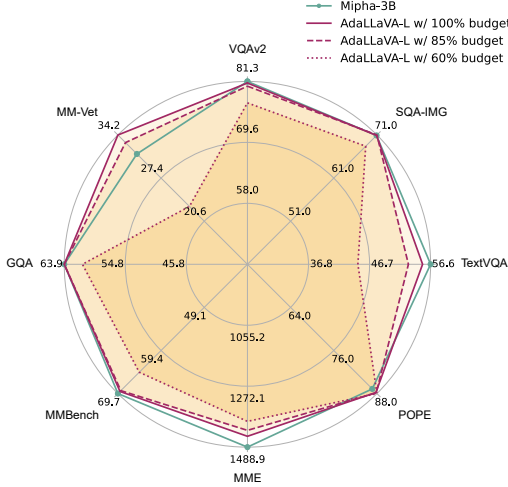


Figure B. Relative performance of applying AdaLLaVA-L to Mipha-3B under various latency budget. The center of the radar corresponds to 60% performance of the base Mipha-3B.

framework on broader benchmarks reported in [36], namely GQA [21], SEED-Bench [29], MM-Vet [64], LLaVa-WILD [35], and VizWiz [17] (see Tab. A). The model shows comparable performance and adaptive ability under different latency budget. The results demonstrate the strong generalization of AdaLLaVA to a wide range of benchmarks.

AdaLLaVA-L maintains comparable performance under full computational budgets. With reduced compute budgets, AdaLLaVA-L shows minimal performance degradation: an average accuracy drop of only 0.7% at 85% budget. Notably, AdaLLaVA-L shows 1.5% average performance boost at full compute budget.

Generalization across MLLMs. We demonstrate that AdaLLaVA can generalize to other MLLMs beyond LLaVA. We consider Mipha-3B [70], a lightweight MLLM built on Phi-2.7B [23]. Specifically, we apply AdaLLaVA-L on Mipha-3B, following its training strategy [70], and report the results on a comprehensive MLLM benchmark (MME), shown in Fig. B. We see that AdaLLaVA-L maintains comparable performance under full computational budgets. With reduced compute budgets, AdaLLaVA-L shows minimal performance degradation: an average accuracy drop of only 3.4% at 85% budget and 6.1% at 60% budget. These results have similar trend to those with LLaVA-1.5 in Fig. 1.

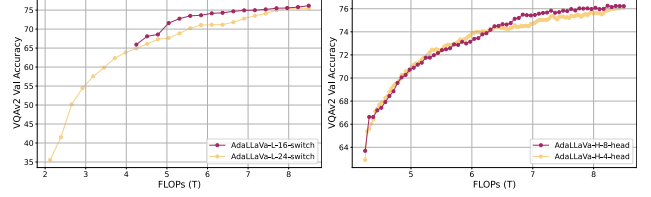


Figure C. Ablation studies on switch design choices.

C. Additional Ablation Studies

We now conduct ablation study, exploring different design choices. We explore the performance of different designs of tunable switches, namely AdaLLaVA-L and AdaLLaVA-H (detailed in Sec. 3.5). All results are reported with LLaVA 1.5-7B Model on VQA v2 dataset benchmark.

Number & granularity of switches. We here conduct ablation studies to examine how the number and granularity of switches affect performance. Fig. C (Left) compares switches for the last 16 layers (used in Sec. 4) versus 24 layers in AdaLLaVA-L. While 24 switches enable finer FLOPs control, they significantly reduce model performance. The 16-switch configuration provides better accuracy while maintaining efficient adaptability. Fig. C (Right) evaluates attention sampling group sizes in AdaLLaVA-H, focusing on operations within the last 16 layers. While both 4-head and 8-head (used in Sec. 4) configurations show comparable performance-latency tradeoffs, the 4-head version enables more granular latency control.

Design of the switches L vs H. We also explore the performance of design of tunable switches, particularly AdaLLaVA-L versus AdaLLaVA-H. Both methods allow adaptivity to latency requirements without significant modification to the pretrained LLM, while AdaLLaVA-H offers better flexibility to latency input.

As shown in Fig. C, from FLOPs ranging from 5T to 8T, AdaLLaVA-H-8-head shows slightly better performance overall, reaching approximately 76% on VQA v2 Accuracy compared to AdaLLaVA-L-16-switch which peaks around 75%. Moreover, AdaLLaVA-H demonstrates finer-grained control over the accuracy-latency trade-off. This is evident from the smoother curve of AdaLLaVA-H, which can be attributed to its head/neuron-level switches providing more granular control over computational resources compared to the layer-level switches. This flexibility allows AdaLLaVA-H to accommodate a wider range of latency budgets.

Comparison with naive sampling strategies. We compare the performance of our AdaLLaVA-L versus random uniform sampling, where we disable the scheduler during training, showing in Fig. D. Both methods are built on Mipha-3B and fine-tuned using the same procedure. Random sampling is worse than AdaLLaVA and has high variance in results (shaded area).

Method	LLM	Budget (%)	FLOPs (T)	Prefill time (ms)	VQA ^{v2} [14]	SQA ^I [39]	VQA ^T [51]	POPE [33]	MME [13]	MMBench [38]
BLIP-2 [32]	Vicuna-13B	100	-	-	41.0	61	42.5	85.3	1293.8	-
InstructBLIP [10]	Vicuna-7B	100	-	-	-	60.5	50.1	-	-	36
InstructBLIP [10]	Vicuna-13B	100	-	-	-	63.1	50.7	78.9	1212.8	-
Shikra [7]	Vicuna-13B	100	-	-	77.4	-	-	-	-	58.8
IDEFICS-9B [28]	LLaMA-7B	100	-	-	50.9	-	25.9	-	-	48.2
IDEFICS-80B [28]	LLaMA-65B	100	-	-	60.0	-	30.9	-	-	54.5
Qwen-VL [3]	Qwen-7B	100	-	-	78.8	67.1	63.8	-	-	38.2
Qwen-VL-Chat [3]	Qwen-7B	100	-	-	78.2	68.2	61.5	-	1487.5	60.6
LLaVA-1.5 [36]	Vicuna-7B	100	8.6	81	78.5	66.8	58.2	85.9	1510.7	64.3
w/ AdaLLaVA-L	Vicuna-7B	100	8.6	81	78.4	67.8	57.0	85.9	1521.0	63.7
w/ AdaLLaVA-L	Vicuna-7B	85	7.2	69	77.1	67.4	54.5	86.4	1487.2	63.7
w/ AdaLLaVA-L	Vicuna-7B	60	5.1	49	75.0	66.9	47.7	86.1	1463.8	63.8
w/ AdaLLaVA-H	Vicuna-7B	100	8.6	81	77.9	68.5	57.1	86.9	1471.1	64.1
w/ AdaLLaVA-H	Vicuna-7B	85	7.2	69	76.8	68.2	55.2	86.7	1494.9	64.3
w/ AdaLLaVA-H	Vicuna-7B	60	5.1	49	74.2	68.1	48.7	85.0	1489.6	64.8
LLaVA-1.5	Vicuna-13B	100	16.7	157	80.0	71.6	61.3	85.9	1531.3	67.7
w/ AdaLLaVA-L	Vicuna-13B	100	16.7	157	79.7	72.4	59.9	86.9	1559.3	69.2
w/ AdaLLaVA-L	Vicuna-13B	85	14.2	133	79.1	72.4	58.0	86.2	1563.9	68.9
w/ AdaLLaVA-L	Vicuna-13B	60	10.0	94	77.4	71.8	54.3	87.3	1552.6	68.6
w/ AdaLLaVA-H	Vicuna-13B	100	16.7	157	80.0	72.6	59.9	87.3	1531.9	67.4
w/ AdaLLaVA-H	Vicuna-13B	85	14.2	133	78.9	72.3	59.0	86.1	1554.5	67.0
w/ AdaLLaVA-H	Vicuna-13B	60	10.0	94	76.4	71.3	53.3	85.0	1529.5	66.9
Prumerge [49]	Vicuna-7B	100	1.4	16	72.0	68.5	56.0	76.3	1350.3	60.9
w/ AdaLLaVA-L	Vicuna-7B	100	1.4	16	71.0	69.1	54.1	74.2	1312.6	58.4
w/ AdaLLaVA-L	Vicuna-7B	85	1.2	14	69.7	68.6	52.5	75.6	1313.3	59.1
w/ AdaLLaVA-L	Vicuna-7B	60	0.8	10	67.8	68.7	44.7	75.8	1332.5	57.0
w/ AdaLLaVA-H	Vicuna-7B	100	1.4	16	70.4	67.9	54.4	77.2	1311.4	60.1
w/ AdaLLaVA-H	Vicuna-7B	85	1.2	14	69.2	67.2	52.3	75.5	1309.7	60.7
w/ AdaLLaVA-H	Vicuna-7B	60	0.8	10	66.8	68.1	45.9	76.4	1289.3	58.7
Prumerge+ [49]	Vicuna-7B	100	3.0	29	76.8	68.3	57.1	84.0	1462.4	64.9
w/ AdaLLaVA-L	Vicuna-7B	100	3.0	29	76.3	68.3	55.8	85.1	1455.5	61.9
w/ AdaLLaVA-L	Vicuna-7B	85	2.6	24	75.3	68.5	52.9	85.7	1429.5	62.5
w/ AdaLLaVA-L	Vicuna-7B	60	1.8	17	73.0	67.7	47.4	85.6	1450.9	61.3
w/ AdaLLaVA-H	Vicuna-7B	100	3.0	29	76.0	67.9	56.0	86.6	1503.2	63.2
w/ AdaLLaVA-H	Vicuna-7B	85	2.6	24	75.0	68.1	54.2	86.4	1511.8	63.6
w/ AdaLLaVA-H	Vicuna-7B	60	1.8	17	72.2	67.6	47.2	86.4	1458.0	63.6
FastV (K=2,R=0.5) [8]	Vicuna-7B	100	4.9	47	77.7	68.7	58.1	82.5	1516.2	64.3
w/ AdaLLaVA-L	Vicuna-7B	100	4.9	47	77.8	67.7	57.0	82.8	1494.3	63.5
w/ AdaLLaVA-L	Vicuna-7B	85	4.2	40	76.9	67.8	54.4	83.3	1478.1	63.7
w/ AdaLLaVA-L	Vicuna-7B	60	3.0	29	74.5	67.0	47.2	83.8	1463.1	63.2
w/ AdaLLaVA-H	Vicuna-7B	100	4.9	47	77.4	68.4	57.0	84.3	1484.2	63.8
w/ AdaLLaVA-H	Vicuna-7B	85	4.2	40	76.6	67.7	54.8	83.9	1520.5	63.9
w/ AdaLLaVA-H	Vicuna-7B	60	3.0	29	73.9	68.3	48.7	82.4	1452.8	65.3

Table B. **Results of MLLMs on six benchmarks.** Our AdaLLaVA can be applied to LLaVA 1.5 with different size of LLM with different design of switches. Percentage (%): The input latency requirement. AdaLLaVA-L: switches on selecting different transformer blocks. AdaLLaVA-H: switches on select different attention heads and MLP activations. VQA^{v2}: VQAv2 set. SQA^I: ScienceQA set. VQA^T: TextVQA set. Prumerge: LLaVA 1.5 with PruMerge.

D. Additional Results on Adaptivity

We provide further results to demonstrate AdaLLaVA’s latency and content adaptivity.

Model Response under different latency. Here we show additional results on model response given same image-text input under different latency budget, similar to Fig. 1. As shown in Tab. C, given an image-query pair and latency constraint, AdaLLaVA learns to generate appropriate re-

sponses while adapting to varying computational budgets.

Visualization of execution plans with different latency. We report execution plans of the same input with varying budgets in Fig. E. As budget decreases, the scheduler prioritizes keeping the last 3 layers over others. We have included outputs for the same input with varying budgets in Tab. C.

Visualization for latency token attention. We provide additional results on content awareness by showing the key-

Visual input example:

			
User	Describe this image.	Describe this image.	Describe this image.
LLaVA 1.5-7B	The image features an older man sitting at a table, surrounded by various items. He is using a cell phone, possibly checking messages or browsing the internet...	The image features a wooden block with a set of wooden letters arranged in a way that spells out the word "press stop." ...	The image features a woman walking down a street, holding a bouquet of flowers in her hand...
AdaLLaVA-L-50%	<p>The image depicts a man sitting at a table with a large number of papers and books on it. He is using a cell phone, possibly checking or communicating with someone. The man is wearing a tie, which adds to the professional setting.</p> <p>The table is surrounded by several chairs, with one close to the man and others further away. A clock can be seen on the left side of the table, and a handbag is placed on the right side. The scene appears to be a workspace or a study area, with the man working on his tasks and using the cell phone.</p>	The image features a wooden board with a series of wooden chicks on it. The wooden chicks are placed in a row, with each one occupying a different position on the board. The arrangement of the chicks is a close-up of the word "press" and the word "stap."	<p>The image captures a young woman standing in a busy street, holding a bouquet of flowers. She appears to be smiling, possibly indicating that she is enjoying the moment. The street is filled with various vehicles, including several motorcycles and a motorcycle parked close to the woman.</p> <p>There are numerous people in the scene, some standing closer to the woman, while others are further away. Some of the people are closer to the left side of the image, while others are further back, creating a sense of a busy city street.</p>
AdaLLaVA-L-75%	<p>The image features a man sitting at a table, surrounded by various items. He is wearing a suit and tie, and appears to be using a cell phone. The table is set with a clock, a cup, and a book. There are also two chairs in the scene, one near the man and another further away.</p> <p>In the background, there are two more books placed on the ground, and a handbag is located near the edge of the scene. The man seems to be engaged in a task or a conversation, as evidenced by his use of the cell phone.</p>	The image features a wooden word puzzle made up of four wooden letters, each with a number on them. The letters are arranged in a way that reads "Press Stop." The letters are placed on a white background, creating a visually appealing and playful display. The arrangement of the letters and numbers adds a unique and creative touch to the word puzzle.	<p>The image depicts a woman walking down a street, holding a bouquet of flowers in her hand. She is wearing a dress and appears to be enjoying her time outdoors. The street is lined with various vehicles, including motorcycles and a truck, parked along the side.</p> <p>There are several people in the scene, some of whom are walking or standing near the vehicles. A handbag can be seen resting on the ground, possibly belonging to one of the individuals. The overall atmosphere of the scene is lively and bustling, with people going about their daily activities.</p>
AdaLLaVA-L-100%	<p>The image features an older man sitting at a table, working on a laptop. He is surrounded by various items, including a cup, a cell phone, and a book. The table is adorned with a red and black tablecloth, and there are several signs and posters on the wall behind him.</p> <p>In the room, there are two chairs, one near the man and another further away. A clock is also visible on the wall, indicating the time. The scene appears to be a workspace or a study area, with the man focused on his tasks.</p>	The image features a wooden tray with a set of wooden letters arranged on it. The letters spell out the word "Press Stop," with each letter positioned next to the other. The arrangement of the letters creates a visually appealing and organized display. The wooden tray itself is placed on a white surface, which further emphasizes the contrast between the wooden letters and the background.	<p>The image depicts a woman walking down a street, holding a bouquet of flowers in her hand. She is wearing a flowered dress and appears to be enjoying her time. The street is lined with various vehicles, including several motorcycles parked on both sides of the road.</p> <p>There are also a few people walking around, some of whom are carrying handbags. The scene captures a lively atmosphere with people going about their daily activities.</p>

Table C. AdaLLaVA-L on LLaVA 1.5-7b model, generating appropriate responses while adapting to varying computational budgets.

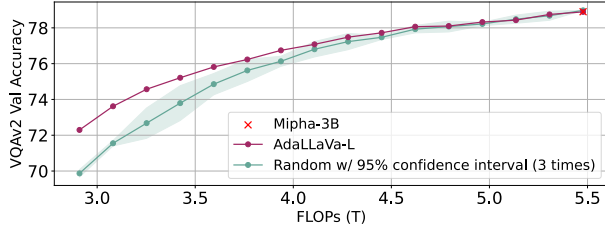


Figure D. Comparison to random sampling

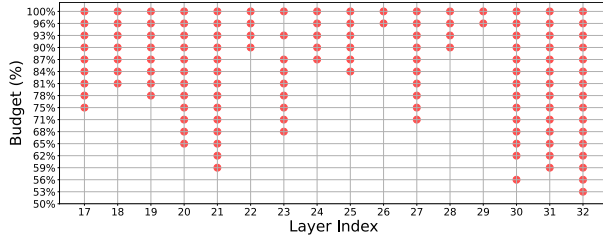


Figure E. Visualization of execution plan

query attention scores of the latency token and the input visual tokens with different text questions, similar to Fig. 4.

Fig. F further demonstrate the model’s content-aware behavior. In the father-child scene image, attention spans the entire street for scene description but focuses centrally for query asking for activity. For Happy Plaza image, attention targets the storefront sign for location queries but shifts to promotional areas for query about special offers. In the restaurant scene, attention distributes across interior elements when identifying location type but concentrates on the woman’s clothing for attire questions. This consistently shows model adjusts its attention based on the query.

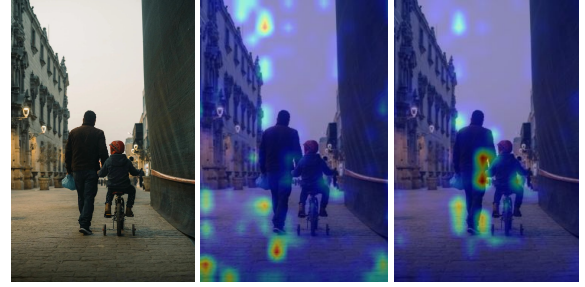
Visualization for latency token across layers. We plot the evolution of the latency token from layers 12 to 16 using the same example in Fig. 4 of the main paper (see figure below). As seen in Fig. G, the latency token progressively gathers key information from the input visual tokens for scheduling.



Figure G. Evolution of latency token across layers in AdaLLaVa-L on 7b model.

E. Further Discussion

FLOPs, latency, and cross-device portability. In our work, compute budgets are expressed as percentages of a base model’s FLOPs. Percentages can be translated into absolute FLOP targets, given base model’s architecture and input size, from which total FLOPs can be estimated. We choose FLOPs over runtime latency, as it abstracts away hardware / software specific variations . We acknowledge



Question: Describe the scene?

Answer: The scene depicts a man walking down a sidewalk with a young boy on a bicycle. The man is holding the boy's hand, ensuring his safety as they walk together. There are several other people in the background, possibly going about their daily activities. The setting appears to be a city street, with a mix of pedestrians and urban elements.

What are they doing?

The man and the child are walking together on a sidewalk, with the man pushing a bike with a child on it.

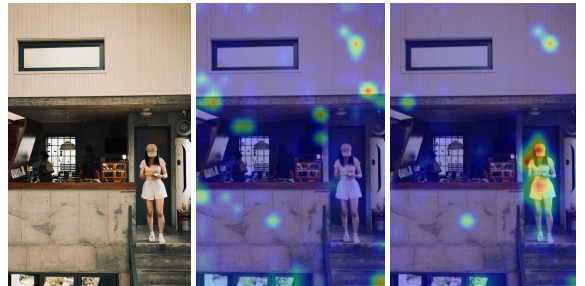


Question: What is the name of the place?

Answer: The name of the place is Happy Plaza.

Are there any special promotion?

Yes, there is a special promotion at the Happy Plaza shopping center. The store is advertising a buy one, get one free deal, which is likely to attract customers and encourage them to shop at the location.



Question: Where is this place?

Answer: This place is a restaurant, as evidenced by the presence of a dining table, chairs, and a counter. The woman is standing on a staircase, possibly near the entrance or exit of the restaurant.

What is the woman wearing?

The woman is wearing a pink shirt and white shorts.

Figure F. The key-query attention scores between latency token and visual tokens. The latency input is 1.0 in these examples.

that the relationship between FLOPs and latency vary across devices, depending on model architecture, hardware and software. We leave cross-device portability to future work.