A. Description of Numeorus Evaluation Benchmarks

- **QBench** [93] is a comprehensive benchmark designed to evaluate the low-level visual abilities of multimodal large language models (MLLMs), focusing on perception, description, and quality assessment of visual attributes. It introduces datasets like LLVisionQA for diverse low-level attribute queries, LLDescribe for detailed expert-crafted image descriptions, and a unified softmax-based strategy for quantifiable image quality assessment. Q-Bench highlights that while MLLMs exhibit preliminary capabilities in handling low-level visual tasks, their outputs remain inconsistent and imprecise, emphasizing the need for further advancements to align with human perception and achieve general-purpose applications.
- AI2D [44] is a benchmark dataset developed to study diagram interpretation and reasoning, focusing on identifying diagram structures and semantic relationships. It introduces Diagram Parse Graphs (DPG), a graph-based representation that encodes the syntactic and semantic structure of diagrams. The dataset contains over 5,000 grade-school science diagrams with exhaustive annotations of constituents and relationships, as well as 15,000 multiple-choice questions for diagrambased reasoning tasks.
- ChartQA [72] is a large-scale benchmark designed to assess question-answering systems' ability to reason logically and visually about data visualizations like bar, line, and pie charts. It includes 9,608 human-authored questions and 23,111 machine-generated questions, focusing on complex reasoning tasks involving mathematical operations, visual attributes, and multi-step logical inferences. By utilizing both extracted data tables and visual features, the benchmark highlights challenges in handling real-world charts and emphasizes the gap in models' ability to process intricate visual and logical questions compared to human understanding.
- SEED-Bench-2-Plus [55] is a comprehensive benchmark designed to evaluate Multimodal Large Language Models (MLLMs) on their ability to comprehend and reason about text-rich visual content across three categories: Charts, Maps, and Webs, covering 63 diverse data types. It includes 2.3K meticulously crafted multiple-choice questions with human-verified answers, simulating real-world scenarios that combine complex text and visual data.
- **POPE** [58] is a polling-based evaluation framework designed to assess object hallucination in Large Vision-Language Models (LVLMs). It formulates hallucination detection as a binary classification task using simple yes/no questions (e.g., "Is there a chair in the image?") to probe LVLMs. Unlike previous methods, POPE offers a stable and flexible approach by avoiding dependence on lengthy captions or complex parsing rules. It introduces three object sampling strate-gies—Random, Popular, and Adversarial—to explore hallucination patterns in models.
- HallusionBench (HallB) [64] is an advanced diagnostic benchmark designed to evaluate and analyze the failure modes of Large Vision-Language Models (LVLMs) in handling both language hallucinations and visual illusions. Featuring 346 images and 1,129 human-crafted visual-question pairs, it tests models like GPT-4V and LLaVA-1.5 using unique control pairs and human-edited images to assess logical consistency, perception, and reasoning. Results highlight persistent challenges, with top models achieving only 31.42% accuracy, revealing their over-reliance on parametric memory, susceptibility to simple manipulations, and struggles with geometry, math, and temporal reasoning.
- **MME** [30] is a comprehensive benchmark designed to evaluate Multimodal Large Language Models (MLLMs) across perception and cognition abilities with 14 subtasks. The benchmark includes tasks like object recognition, OCR, commonsense reasoning, and numerical calculation, using manually curated instruction-answer pairs to ensure fairness and avoid data leakage. MME emphasizes concise instructions for consistency and quantitative assessment, highlighting that current MLLMs, despite their progress, face challenges such as instruction-following errors, limited perception and reasoning, and hallucinations.
- MathVista [71] is a benchmark designed to evaluate the mathematical reasoning abilities of foundation models in visual contexts. It comprises 6,141 examples sourced from 31 datasets, including three newly created datasets—IQTest, FunctionQA, and PaperQA—tailored to assess logical, algebraic, and scientific reasoning in visual settings. MathVista emphasizes diverse visual contexts, such as diagrams, charts, and academic figures, and covers seven types of reasoning across five core tasks.
- MMB, MMB-Chinese (MMB^{CN}) [67] is a multilingual benchmark designed to evaluate the multimodal capabilities of vision-language models (VLMs) across 20 fine-grained abilities, including perception, reasoning, and relation understanding. It features over 3,000 high-quality multiple-choice questions in English and Chinese, enabling comparative analysis in a bilingual context. MMBench introduces novel evaluation strategies like CircularEval, which enhances robustness by testing models across shuffled choices, and employs GPT-4 for accurate choice extraction.
- **MM-Vet** [100] is a benchmark designed to evaluate the integrated vision-language capabilities of Large Multimodal Models (LMMs). It defines six core capabilities—recognition, OCR, knowledge, language generation, spatial awareness, and math—and examines their combinations across 16 emergent multimodal tasks, such as explaining memes, solving spatial math problems, and summarizing visual data. The benchmark introduces an LLM-based evaluator to assess open-ended

model outputs consistently, focusing on both accuracy and quality.

- **MM-Vet-v2** [101] builds upon the original MM-Vet benchmark by introducing a new core capability, image-text sequence understanding, to evaluate large multimodal models (LMMs) on processing arbitrarily interleaved sequences of images and text. With an expanded dataset of 517 high-quality evaluation samples and tasks requiring combinations of seven core capabilities, it assesses advanced real-world scenarios like temporal reasoning, spatial understanding, and multimodal comparisons.
- MMMU [102] is a benchmark designed to evaluate large multimodal models on 11550 college-level problems requiring
 expert knowledge and reasoning across six disciplines: Art, Business, Science, Medicine, Humanities, and Engineering.
 Spanning 30 subjects and incorporating 30 diverse image types like charts, medical scans, and diagrams, it challenges
 models to integrate complex text and image inputs while applying domain-specific knowledge. MMMU sets a high standard
 for advancing multimodal AI and plays a crucial role in developing Expert AGI.
- **MMStar** [10] is a vision-critical multimodal benchmark consisting of 1,500 meticulously curated samples designed to evaluate large vision-language models (LVLMs) across six core capabilities and 18 specific axes. By addressing two key issues in existing benchmarks—unnecessary reliance on textual knowledge and unintentional data leakage—MMStar ensures each sample requires genuine visual reasoning and minimal data recall. Incorporating metrics for multimodal gain and data leakage, it provides a robust platform for assessing the true multimodal capacities of LVLMs.
- **BLINK** [31] is a benchmark designed to evaluate the core visual perception abilities of multimodal large language models (MLLMs) across 14 tasks, such as depth estimation, visual correspondence, and spatial reasoning, inspired by classical computer vision problems. With 3,807 multiple-choice questions derived from 7,300 images, BLINK focuses on tasks that humans can solve "within a blink" but remain challenging for models, as even advanced models like GPT-4V achieve only 51.26% accuracy compared to 95.7% for humans. It highlights the gap in nuanced visual perception and suggests integrating specialized vision models as a pathway for improving MLLMs' performance.
- **CV-Bench** [89] is a vision-centric benchmark introduced to evaluate the fundamental 2D and 3D visual understanding capabilities of Multimodal Large Language Models (MLLMs). With 2,638 manually inspected examples sourced from datasets like ADE20K, COCO, and Omni3D, it tests tasks such as spatial relationships, object counting, depth ordering, and relative distances. By transforming traditional vision benchmarks into VQA format, CV-Bench ensures robust assessment of models' abilities in multimodal contexts. It addresses gaps in existing benchmarks by offering significantly more samples, better diversity, and a stronger focus on visual grounding, making it a critical tool for advancing multimodal AI systems.
- LLaVA-Wilder [106] is a dataset designed to evaluate large multimodal models (LMMs) in real-world scenarios. It comprises 128 image-text pairs, each featuring an image accompanied by a question and a detailed answer. The dataset includes a variety of images, such as indoor and outdoor scenes, memes, paintings, and sketches, to assess models' generalization capabilities across diverse domains. By providing this resource, LLaVA-Bench-Wilder aims to facilitate the development and benchmarking of LMMs, ensuring their robustness and effectiveness in handling complex, real-world visual tasks.

Target Layer Index	10%	30%	50%	70%	90%	100%
#1	14.28	11.73	9.87	6.78	3.54	3.30
#2	12.60	10.48	8.04	5.98	3.10	2.90
#3	11.22	9.32	7.02	5.05	2.63	2.37
#4	10.48	8.11	6.41	4.57	2.35	2.18
#5	9.34	7.55	5.81	3.54	1.82	1.79
#6	8.16	6.18	4.27	2.09	1.56	1.47
#7	7.39	3.85	2.02	1.03	0.85	0.79

B. Verbalization Loss Changes

Showing loss changes during the verbalization step (% indicates training progress)

C. VLsI's Verbalization Examples



presence of scattered cushions and a plush couch adds to the comfort and leaves many meriodic platects of platects of scattered cushions and a platect platects of the scattered cushions and a platect platec



	VLsi
	#1 Target Layer (Actual 2 nd Layer)
Solution States	The image displays a moment of a large, white, and white color. The image depicts a large, white background, with a small, white background. (Repeats)
MENTER AND	#2 Target Layer (Actual 6 th Layer)
A WILLIE	The image depicts a large, small, and black color scheme with a red and white text. The text is a black and white line, which is a common feature in the image.
	#3 Target Layer (Actual 10 th Layer)
the states	The text "100% of (Repeats)
Question: What is located to the right	#4 Target Layer (Actual 14 th Layer)
of the shampoo?	The text on the sign is the text on the sign.
	#5 Target Layer (Actual 18 th Layer)
↓↓↓	The first answer is the first answer to the question.
┃ ■┼ ┛-┛→│	#6 Target Layer (Actual 22 nd Layer)
	To the right of the shampoo, there is a "Densifying conditioner for fine & thin hair" and a "Densifying manger for fast quick results." These are part of the "Therrier & getDefault\' series
	by chaines w. magon.
	#7 Target Layer (Actual 26th Layer)
	To the right of the shampoo, you will find the densifying conditioner and the densifying m związku from the "Thicker & Fuller" line by Charlesworth London.
Verbalizer	
To the right of the shampoo, you will	find the densifying conditioner. This product is part of the "Thicker & Fuller" line and is designed to enhance
hair volume and fullness. It typically for	ollows the shampoo in a hair care routine, providing additional benefits to the hair\'s texture and appearance.



Question: Which chest xray disease can be found? (Options include Atelectasis; Cardiomegaly; Effusion; Infiltration; Mass; Nodule; Pneumonia; Pneumothorax; Consolidation; Edema; Emphysema; Fibrosis; Pleural Thickening; Hernia.)



VLsI #1 Target Layer (Actual 2nd Layer) The image is a large, white, and white, with a small, white, and white. (..Repeats..) #2 Target Laver (Actual 6th Laver) The image is a black and white line with a visible pattern of color. #3 Target Layer (Actual 10th Layer) The image is a black and white image of a large, black and white image of a small, small, black and white image of a small, small, small, and small. The text is in English and is in the text of the text. #4 Target Layer (Actual 14th Layer) The image is a black and white image of a large, dark coloration, which is a common coloration for a high-quality, high-quality image. The image is a black and white image, which is a common color scheme for a large and more likely than a small, small size. #5 Target Laver (Actual 18th Laver) The most likely type of corruption of the image is the answer to the question #6 Target Layer (Actual 22nd Layer) Based on the given options and the characteristics of the X-ray, the most likely visible condition is "Béal\'. This is a specific type of "Furball," which is a term often used to describe a large, bright, or bright and bright image in a T MRI image. (..Repeats..) #7 Target Layer (Actual 26th Layer) The chest X-ray indicates the presence of aGetSize; however, the specific condition is not clearly visible from the image provided. To accurately determine the nature of the mass, a detailed analysis by a medical professional is necessary, common chest X-ray findings include/\n/n-Atelectasis: collapse of lung tissue, h-C radiomegaly: enarged heart.h-E fifusion: Jilua accumulation.\n- Infiltration: tissue or fluid accumulation.\n- Mass: a non_widget mass.\n- Noadle: a smoothly margins mass.\n- P pneumonia: inflammation.\n- P pneumonia: inflammation.\n- Consolidation of lung tissue.h- Edema: tissue swelling.\n- Emphysema: lung tissue damage.\n- Fibrosis: tissue os cscr.\n- Pleural thickening: thickening of the pleural membrane.\n/nEach of these findings requires a thorough medical evaluation to ensure accurate diagnosis and treatment. The chest X-ray indicates the presence of infiltration. Infiltration refers to the accumulation of substances, such as fluid or cells, within the lung tissue,

which can appear as areas of increased opacity on the X-ray. This can be indicative of various underlying conditions, including infection, inflammation, or the presence of substances like blood or fluid. Identifying infiltration is crucial for further diagnostic evaluation and treatment planning.

	🗥 VLsi
	#1 Target Layer (Actual 2 nd Layer)
	The image displays a moment of a large, white, and white color. (Repeats)
12	#2 Target Layer (Actual 6 th Layer)
Question: In which country was this photo taken?	The first of the two people are in the image, which is a large, black and white line with a red and white color scheme. The text is a common feature of the image, with the text "12" on the top and the "20" on the right.
	#3 Target Layer (Actual 10 th Layer)
	The woman in the image is likely a young man, likely a young man with a high level of attention to detail, likely a young man with a high level of energy and energy.
	#4 Target Layer (Actual 14 th Layer)
	The question is about the question about the image of the pizza and the question about the text.
	#5 Target Layer (Actual 18 th Layer)
	The setting of the image is likely the United States, as the setting and setting of the setting is a typical setting for a dining table. The setting is likely a dining table, which is a common way to display a sign a sign that is often used to display a sign. (Repeats)
	#6 Target Layer (Actual 22 nd Layer)
Verbalizer	The question asks about the location of the setting in the image. The answer suggests that the setting is in a location where the "klebe" is used. This is likely a reference to the "klebe" (Repeats)
	#7 Target Layer (Actual 26 th Layer)
(33)	The photograph was taken in Australia. This conclusion is based on the context provided by the question, which suggests that the location is significant for the event depicted. The presence of the trophy being held by the individual is a common element in sports contexts, often associated with major international events. Given the specific details provided, the most likely answer is Australia, as it is a country hosting numerous significant sports events, including the Australian Open, a major tournament in the context of tennis.
The photograph was taken in Australia moment of celebration likely related to is reminiscent of those given out at ma	. This conclusion is based on the context provided, which indicates that the image captures a o a significant event in the country, such as a major sports tournament. The trophy being held for Australian events, reinforcing the likelihood that this scene is set in Australia.



	Dataset 1		
Department	Name	Salary	
Administration	Emily	\$2,000	
Finance	Nick	\$1,800	#1 larget Layer (Actual 2 ¹¹ Layer)
Marketing	Jonn Gloria	\$2,600	The answer is 10
Administration	Jane	\$2,500	
Finance	Max	\$2,000	#2 Target Layer (Actual 6 th Layer)
Administration	Tyler	\$3,000	The answer is 1
Finance	Wilkins	\$2,500	
Administration	Sabine	\$1,900	#3 Target Layer (Actual 10 th Layer)
Question: Whi	ich depa	rtment is the	The first two states of the data is the first answer
person who h from?	has the	highest salary	#4 Target Layer (Actual 14 th Layer)
			The answer is
		•	#5 Target Layer (Actual 18 th Layer) The first answer is the answer to the question. #6 Target Layer (Actual 22 nd Layer) The person with the highest salary of \$3,000, belongs to the "Accounting" department. This is determined by comparing the salaries of the individuals in each department. The data points are as follows: "Amazon" with a salary of \$2,000, "F finance" with a salary of \$2,000, "M marketing" with a salary of \$2,000, "A advertising" with a salary of \$3,000, and these, the highest salary is found in the "A agriculture" category. #7 Target Layer (Actual 26 th Layer) The individual with the highest salary is from the Finance department. In dataset 1, the salaries are as follows: \n\n- Administration: \$2,000 (Emily)\n- Finance: \$2,500 (Max)\n- Marketing: \$2,600 (bnn)\n\nAmong these, Max earns the most with a salary of \$2,500, which is the highest salary in the dataset. Therefore, the person with the highest salary is from
The individua Administratio (Wilkins)\n- M Marketing: \$ person with t	al with t on: \$2,00 Marketin 52,800 (N the highe	the highest sa 00 (Emily)\n- 19: \$2,600 (Jo Ailes)\n\nAme 2st salary is fro	lary is from the Administration department. In the provided dataset, the salaries are as follows:\n\n- Finance: \$1,800 (Nick)\n- Finance: \$2,000 (Max)\n- Administration: \$2,500 (Jane)\n- Finance: \$2,500 nn)\n- Marketing: \$1,800 (Gloria)\n- Administration: \$3,000 (Tyler)\n- Administration: \$1,900 (Sabine)\n- ong these, Tyler earns the most with \$3,000, which is the highest salary in the dataset. Therefore, the m the Administration department.



Question: Aesthetics reflects the perceived beauty or visual appeal of an object, design, or piece of art by human. Among the images provided below, based on societal standards and norms, which one more people will consider having a higher aesthetics?



VLsI

#1 Target Layer (Actual 2nd Layer)

The image depicts a moment of a large, white, and white color. The color is a large, white, and white, with a small, white, and white.

#2 Target Layer (Actual 6th Layer)

The image displays a large, small, and white, white, white, and white. The color palette is a common feature of the image, with the visible elements visible in the image. The color palette is a common feature of the image, with the visible elements visible in the image.

#3 Target Layer (Actual 10th Layer)

The image of a large, large blue and white, white and white, and white in the image is a common common way to form a color scheme with a more traditional design. The color palette is a common common way to create a more traditional and traditional color scheme.

#4 Target Layer (Actual 14th Layer)

The image of the large, vibrant red and reds of the buildings in the image are a significant portion of the color scheme of the image. The color scheme is a significant part of the design, and the color scheme is a significant part of the design. The color scheme is a significant part of the design.

#5 Target Layer (Actual 18th Layer)

The second image, which features a collection of colorful and vibrant colors, is likely to be more likely to be considered as a more likely answer for the question as it is likely to be a more likely answer for the question as it is likely to be a more likely answer for the image provided.

#6 Target Layer (Actual 22nd Layer)

Based on the given context, the second image with the colorful buildings with vibrant colors and intricate design elements is likely to be more likely to be perceived as having a higher aesthetic appeal. This is often due to the visual appeal of vibrant colors and the intricate design elements that are often associated with the art and design of the past, which often have a more complex and complex design elements than the more straightforward and less colorful architecture of the first image.

#7 Target Layer (Actual 26th Layer)

Based on societal standards and norms, the second image is likely to be perceived as having a higher aesthetic appeal. This assessment is based on the vibrant colors and charming design elements of the buildings, which often align with current aesthetic preferences for vibrant and playful color schemes. The use of bright hues and intricate details can evoke a sense of vibrancy and creativity, which may be more appealing to a wider audience compared to the more muted and practical colors of the first image.

Based on societal standards and norms, the second image is likely to be perceived as having a higher aesthetic appeal. This is due to its vibrant colors and charming design, which often attract positive attention and admiration. The use of bright hues and the artistic arrangement of elements can evoke a sense of beauty and creativity, making it more visually appealing to a wider audience.

D. Comprehensive Experimental Setup for Ablation Studies

Training and Inference Configuration of LLaVA-OV [56]. Different from Qwen2-VL [92]-based VLsI, we select intermediate target layers of LLaVA-OV-based VLsI to i_s : 4th, 12th, 20th layers, and i_t : 10th, 30th, 50th, and 70th layers. Although LLaVA-OV's language model is Qwen2 [96] that is equal with that of Qwen2-VL, the required number of image tokens in LLaVA-OV are 4 to 10 times more number than that of Qwen2-VL, depending on the height-to-width ratio of the images, given the same pixel count. To accommodate this increased computational demand on 8 NVIDIA A100 80GB GPUs, we reduce the number of intermediate target layers. For efficient training, we equally employ LoRA [40] with rank and alpha parameters to 64, use the AdamW optimizer [69] with a cosine annealing schedule, adjusting the learning rate from $1e^{-4}$ to $1e^{-6}$ over each training step, and use gradient accumulation with 16 steps and gradient checkpointing [84]. The only different training configuration is batch sizes where two batches each are used for the 0.5B and 7B model sizes, and one batch is used for the 72B model sizes. It finally results in 256 (8×16×2) and 128 (8×16×1) batches, respectively. We conduct inference experiments under the equal setup used in Qwen2-VL, where we use a greedy search for text generation.

VLMs	MM-Vet	MMMU
LLaVA-OV-7B	57.5	48.8
Qwen2-VL-7B	62.0	54.1
LLaVA-OV-72B	63.7	56.8
Qwen2-VL-72B	74.0	64.5

Fig. 4 provides the challenging evaluation benchmarks' performances: MM-Vet [100] and MMMU [102]. Each cell represents the performances on their evaluation benchmarks, where the orange colored-values represent LLaVA-OV-based ⁽²⁾ VLsI's result and the purple ones represent Qwen2-VL-based result. This figure reveals consistent trends that using largeand small-backbone VLMs with more bigger model sizes enhances ⁽²⁾ VLsI's performances across all configurations. Besides, we can easily infer that the baseline performances of Qwen2-VL before building ⁽²⁾ VLsI are also higher than those of LLaVA-OV as shown in the above table. Furthermore, Tab. 4(a) shows ⁽²⁾ VLsI's generalization ability to 0.5B and 7B model sizes in LLaVA-OV, but similarly insists that using better large-backbone VLMs provides benefits from performances.

Tab. 4(b) collectively illustrate the impact of incorporating the *SFT step*. Table (b) highlights the significant performance gains achieved by applying additional SFT. Interestingly, the performances without additional SFT step are not satisfactory when doing greedy decoding, but once we utilize beam decoding (N=5), its performances are dramatically enhanced. This suggests that the interaction step implicitly alters and expands the probability space over the language head, while SFT aligns this expanded space to better fit instruction-following. Thus, significant improvements arise from the combination of interaction and SFT step.

Tab. 4(c) It examines whether these improvements result from the fine-tuning effects of 2.9 million visual instruction tuning samples. The results demonstrate that 2 VLsI's performance becomes markedly superior to Qwen2-VL as SFT training progresses from 50% to 100%, confirming that additional SFT plays a much more critical role in driving performance enhancements than the contribution of the visual instruction tuning samples alone.

IL-Ops	LL-Ops	vL-Head	MM-Vet	MMMU
L2	X	×	63.3	53.5
L2	KLD	×	64.6	55.9
L2	L2	×	63.9	54.4
X	L2	-	65.2	56.8
X	KLD	-	66.5	57.5
KLD	KLD	1	75.8	69.3

Tab. 4(d) evaluates the impact of different operations applied to intermediate (IL-Ops) and last layers (LL-Ops) on (IL-Ops) vLsI's performances. The results clearly demonstrate that KL divergence (KLD) operations applied to both intermediate and

last layers yield the best performances, confirming the effectiveness of KLD over cross-entropy (CE) or L2 for intermediate layer alignment and final layer interaction. To further assess the effectiveness of verbalization in transferring the knowledge via distillation, we conduct an ablation study by removing the language head in verbalizer (vL-head). In this setup, the *verbalization step* is skipped, and the *interaction step* is first performed by aligning the hidden dimensions of the verb-FFNs in large- and small-backbone VLM. Here, the verb-FFN's hidden dimension in large-backbone VLM is kept but the hidden dimension in small-backbone VLM's verb-FFN is enlarged to match the large one. In addition, the verb-FFNs of large- and small-backbone VLM are interacted with only L2, which means that small-backbone VLM's verb-FFNs try to naively follow those of large-backbone VLM. After *interaction step*, *SFT step* is equally conducted. However, as shown in the above table, this naive approach results in significantly lower performance compared to the version equipped with the language head. Interestingly, using only last layer distillation by L2 and KLD provides more benefits than using intermediate distillation without vL-Head. These findings suggest that directly imitating outputs from the large-backbone VLM, without verbalization from vL-Head, introduces instability and can lead to suboptimal results, highlighting the critical role of the language head in achieving effective distillation.

Pseudo-Code for Random Index (Search Range X)

1: Input: t_s, t_l 2: **Initialize:** loss: 0, i_1^* : 0, ϵ : 1e-6, scale: 2 3: for i_s in $0 \le i_s < t_s$ do kld-list = []4: for i_l in $0 \le i_l < t_l$ do 5: kld-list.append(compute-kld(i_s, i_l)) 6: 7: end for $r \leftarrow \text{Random-Select}(kld-list)$ 8: 9: $loss \leftarrow loss + r$ 10: end for 11: Return: loss

Pseudo-Code for Uniform Index (Search Range X) 1: Input: t_s, t_l 2: Initialize: loss: $0, i_l^*: 0, \epsilon$: 1e-6, scale: 2 3: layer-gap-ratio \leftarrow floor $(\frac{t_l}{t_s})$ 4: for i_s in $0 \le i_s < t_s$ do 5: $i_l \leftarrow i_s \times$ layer-gap-ratio 6: $u \leftarrow compute-kld(i_s, i_l)$ 7: loss \leftarrow loss +u8: end for 9: Return: loss

Tab. 4(e) highlights the effectiveness of various components in the matching strategy. Random Index, Uniform Index, and Bottom-1 or 3 Index, yield lower scores, underscoring the limitations of simpler selection mechanisms. Note that, the above and below algorithms represent the their detailed experimental setup. Multinomial sampling provides improvements and incorporating the Search Range further enhances performances (*e.g.*, 83.5 on MMB and 69.8 on MM-Vet). Adding order preservation results in a significant leap, particularly on BLINK (59.2) and MM-Vet (75.2), demonstrating the importance of maintaining matched indices' sequence alignment during *interaction step*. Finally, using all together with adaptive temperature achieves the best results across all benchmarks (e.g., 86.3 on MMB and 69.3 on MMMU), showcasing its ability to dynamically control the distribution. These exploration underscore the necessity of advanced sampling strategies and adaptive mechanisms for maximizing the efficiency of transferring the knowledge.

Pseudo-Code for Bottom-k Index (Search Range \bigstar)

1: Input: t_s, t_l 2: **Initialize:** loss: 0, i_l^* : 0, ϵ : 1e-6, scale: 2 3: for i_s in $0 \le i_s < t_s$ do kld-list = [] 4: 5: for i_l in $0 \le i_l < t_l$ do 6: kld-list.append(compute-kld(i_s, i_l)) end for 7: *bottom-k-kld-list* \leftarrow Bottom-*k*(*kld-list*) {length(*bottom-k-kld-list*): *k*} 8: $a \leftarrow \text{Average}(bottom-k-kld-list)$ 9: $\mathsf{loss} \gets \mathsf{loss} + a$ 10: 11: end for 12: Return: loss

Tab. 4(f) highlights verb-FFN's efficiency among different verbalizer architectures in terms of performance and parameter count. While larger architectures such as Decoder×2 (3.3B) and FFN×2 (2.9B) achieve strong performance across MM-Vet and MMMU, the much smaller verb-FFN architecture (269M) delivers comparable results. In contrast, simpler structures like MLP×2 and MLP, while efficient in terms of parameter count, fall short in performance. These results emphasize the effectiveness of the verb-FFN architecture in a lightweight yet high-performing solution. Its efficiency becomes even more pronounced as the number of intermediate target layers increases and larger-backbone VLMs are utilized.

E. VLsI's Text Generation Quality

Gasoline Self Serve Regular 3.49 + Plus 3.55 + Supreme 3.65 + Diesel 3.60 +	Question: How many gallons of supreme gasoline can I get with \$50?
 VLSI-2B To determine how many gallons of supreme gasoline you can buy with \$50, we need to calculate the price per gallon and then apply that to the total amount you have. **Calculate the price per gallon:** The price for regular gasoline is \$3.49. The price for plus gasoline is \$3.55. The price for diesel No. 2 is \$3.60. The price per gallon for each type of gasoline is: Regular: \$3.49 / 10 = \$0.349 per gallon Plus: \$3.55 / 10 = \$0.355 per gallon diesel No. 2: \$3.60 / 10 = \$0.36 per gallon **Convert the total amount of money to cents:** \$50 is equal to 5000 cents. **Calculate the number of gallons of supreme gasoline you can buy:** Divide the total amount of money by the price per gallon for supreme gasoline: 5000 cents ÷ \$0.36 per gallon = approximately 1388.89 gallons 	 VLSI-7B To determine how many gallons of supreme gasoline you can purchase with \$50, you need to consider the price per gallon and subtract it from your budget until you reach zero. **Calculate the cost per gallon:** The price for supreme gasoline is \$3.659 per gallon. **Divide your budget by the price per gallon:** Divide \$50 by \$3.659 to find out how many galons you can buy. .\{\frac{50}{3.659}\approx 13.63 \} gallons. **Determine the exact amount:** You can buy 13 full gallons of supreme gasoline with \$50. So, with \$50, you can purchase 13 full gallons of supreme gasoline.





actual usage of energy services.







Question: Can you give a short introduction to this painting?



This painting is known as "The Return of the Tibetan Envoy," created by Zhang Zeduan in 1369. It is a significant work in Chinese history, depicting the return of a Tibetan Envoy to the Tang dynasty. The painting is celebrated for its artistic excellence and historical significance. It is often referred to as "China's Mona Lisa" and is one of the most famous pieces of Chinese art. The painting is housed in the Palace Museum in Beijing.



"Emperor Taizong Receiving the Tibetan Envoy" is a silk painting by Yan Liben, created during the Tang dynasty. Measuring 199 centimeters in length and 49.5 centimeters in width, this artwork is housed in the National Palace Museum in Taipei. The painting vividly portrays the historic meeting between the Tang dynasty's Emperor Taizong and a Tibetan Envoy, highlighting the amicable relations between the two cultures during that era.



Question: The graph and table below give information about water use worldwide and water consumption in two different countries. Summarise the information by selecting and reporting the main features, and make comparisons where relevant. You should write at least 150 words.

VLsI-2B

The graph illustrates the global water usage across agriculture, industry, and domestic sectors from 1900 to 2000. In 1900, agricultural water usage was the lowest, followed by industrial and then domestic. By 2000, agricultural water usage had increased significantly, surpassing industrial and domestic. The table provides data on Brazil and the Democratic Republic of Congo in 2000, highlighting their population, irrigated land, and water consumption per person.

In Brazil, the population was 176 million, with 26,500 km² of irrigated land and a water consumption per person of 359 m³. In contrast, the Democratic Republic of Congo had a population of 5.2 million, with 100 km² of irrigated land and a water consumption per person of 8 m³. Despite having a smaller population, the Congo utilized more water per capita, highlighting the disparity in water consumption between the two countries.



The graphs illustrate global water usage across agriculture, industry, and domestic sectors from 1900 to 2000, alongside data on water consumption in Brazil and the Democratic Republic of Congo in 2000.

Overall, global water consumption increased dramatically from 1900 to 2000, with agriculture being the predominant sector. In 1900, agricultural water usage was minimal, but it surged by 2000, surpassing industrial and domestic usage. Industrial water consumption rose steadily but was consistently lower than agriculture's. Conversely, domestic water usage saw a more moderate increase.

In 2000, Brazil exhibited significantly higher water consumption per capita compared to the Democratic Republic of Congo. Brazil's population was 176 million, with 26,500 km² of irrigated land, resulting in a per capita water consumption of 359 m³. In contrast, the Democratic Republic of Congo, with a population of 5.2 million and 100 km² of irrigated land, had a much lower per capita water consumption of 8 m³. This stark difference highlights the substantial disparity in water usage between the two countries.



Question: The diagram below shows how the Australian Bureau of Meteorology collects up-to-theminute information on the weather in order to produce reliable forecasts. Write a report for a university lecturer describing the information shown below. You should write at least 150 words.

VLsI-2B

The diagram illustrates the methodology employed by the Australian Bureau of Meteorology to generate accurate weather forecasts. It begins with the collection of data from various sources, including satellites, radar, and digital buoys. These data points are then processed through analysis and forecasting models, which utilize satellite photos and radar screens to generate synoptic charts. These charts provide detailed information on weather patterns and conditions across different regions.

The collected data undergoes further analysis to predict weather changes, ensuring that the forecast is timely and relevant. Once the analysis is complete, the prepared forecast is broadcasted through various channels, including television newsreaders, radio, and recorded announcements. This comprehensive approach ensures that the public receives timely and accurate weather updates, contributing to better preparedness and informed decision-making.



The diagram illustrates the methodology employed by the Australian Bureau of Meteorology to generate accurate weather forecasts. The process begins with the collection of current weather data from various sources, including satellites, radar systems, and drifting buoys. This data is then processed and analyzed to create a comprehensive understanding of the weather patterns.

Initially, satellite imagery provides a wide view of cloud formations and temperature differences across the globe. This information is crucial for predicting large-scale weather systems. Simultaneously, radar data is gathered from multiple locations to track local weather conditions, such as rainstorms or snowstorms, offering detailed insights into specific regions.

To enhance the accuracy of these data sources, synoptic charts are utilized. These charts depict pressure systems and wind patterns, allowing meteorologists to identify high and low pressure areas, which are key factors in determining weather movements.

The processed data is then used to create detailed weather forecasts. This involves analyzing the collected information to predict changes in temperature, pressure, and wind speeds. The resulting forecast is prepared for dissemination through various channels, including television, radio, and recorded telephone announcements.

Finally, the prepared forecast is broadcasted to the public, ensuring timely and reliable weather updates. This multi-faceted approach allows the Bureau to provide accurate and timely weather information, enabling individuals and businesses to make informed decisions based on the forecasted conditions.

Question: What is this diagram telling?



This diagram illustrates a cloud architecture designed for a healthcare facility using AWS services. It details the setup of a web application hosted on AWS, along with its dependencies and security measures.

Components and Their Roles:

1. **Amazon S3**: Used for storage and backup of medical records.

2. **Amazon Route53**: Manages DNS services for domain names associated with the cloud infrastructure.

3. **AWS WAF**:提供防护,防止DDoS攻击。

4. **VPC (Virtual Private Cloud)**:

- **Availability Zone 2**: Two Availability Zones are configured for high availability.

- **NAT gateway**: Converts traffic between public and private networks.

- **App subnet**: Contains instances of the web application.

- **DB subnet**: Manages databases for patient records.

5. ******Autoscaling Group******: Manages load balancing and resource optimization.

6. **Amazon RDS (Rapid Deployment Service)**:

- **Primary**: The primary database for patient records.

- **Secondary**: A secondary instance for redundancy.

7. **CloudFront**: Manages content delivery to ensure fast load times for users.

8. **CloudTrail**: track and manage AWS events for security and monitoring.

Security and Protection:

- **AWS Shield**: Ensures secure internet access by providing DDoS protection.

- **CloudTrail**: Logs AWS activities, enhancing security and monitoring.

Conclusion:

This architecture provides a robust framework for a healthcare facility to manage its cloud services efficiently. It includes secure components like AWS WAF, RDS, and CloudTrail, as well as auto-scaling and load balancing to ensure reliable and scalable operations.





The diagram illustrates a cloud architecture using AWS services, designed for secure and reliable hosting of a healthcare application. Here's a detailed explanation:

1. **User Interaction **: Users access the application through a URL, which is directed to an Amazon Route53 DNS service. 2. **Application Security **: The DNS service is protected by an AWS WAF (Web Application Firewall), which filters and

protects against web exploits and malicious traffic. 3. **Application Deployment**: The application is hosted on

an auto-scaling group within an EC2 (Elastic Compute Cloud) instance, ensuring high availability and load management.

4. **Database Management**: The database is stored on Amazon RDS (Database instance), providing reliable and secure data management.

5. ******Network Security******: The network is segmented into multiple subnets within two Availability Zones, offering fault tolerance and network resilience.

6. ******Application Access******: The application is accessible through NAT (Network Address Translation) gateways, which facilitate secure internet connections.

7. **Cloud Storage**: Amazon S3 is used for secure storage and backup of application data.

This architecture combines multiple services to ensure the application is secure, reliable, and scalable.



To interpret the blood sugar test results accurately, follow these steps:

1. **Look at the Date and Time**: The test was conducted on a specific date and time. This helps in comparing the results with the expected values for that day.

2. ******Examine the Blood Sugar Level******: The meter displays the blood sugar level. A normal range for blood sugar is typically between 70 and 110 mg/dL. If the level is outside this range, it may indicate a potential health concern.

3. **Consider the Time of Day**: The blood sugar level can change throughout the day. For example, it may be higher in the morning and lower in the evening. Monitoring at different times can provide more comprehensive data.

4. **Look at the Time of Measurement**: The meter may have a feature to indicate the time of measurement. This can help in understanding if the blood sugar level was taken at a peak or a trough.

5. **Compare with a Blood Sugar Tracker**: If you have a blood sugar tracker, compare the readings with it to see if there are consistent patterns or fluctuations.

6. ****** Consult with a Doctor******: If the blood sugar level is high, it's important to consult with a healthcare professional for personalized guidance.

By following these steps, you can better understand and interpret the blood sugar test results.





The blood glucose meter displays a reading of 97 mg/dL, which is considered a normal blood sugar level. This indicates that the individual's glucose level is within the typical range for someone without diabetes. The meter is easy to use, providing quick results in just 6 seconds, and requires a minimal amount of blood. It also features four reminder alarms to help manage blood sugar levels effectively. The large display ensures easy reading of the test results.



VLsI-2B

The image depicts a robotics simulation environment, likely from a robotics simulation software like RViz. This software is used for visualizing and debugging robotic systems. In the image, you can see a 3D grid representing a simulated environment, with a camera positioned to capture the scene. The grid is colored in red, green, and blue, likely indicating different sensor readings or data points. The top right corner of the image displays parameters for the simulation, such as the robot's orientation and the distance from the camera. This setup is crucial for developers and engineers to test and refine their robotic systems before physical deployment.

VLsI-7B

The image depicts a screenshot of a robotics simulation software interface, likely from the Robot Operating System (ROS) environment. The interface is designed for visualizing and debugging robotic systems. The left side of the screen shows a list of components or sensors, such as "camBotRobot_ee" and "odometry," indicating the various parts of a robotic setup. The central area displays a 3D grid with axes, which is a common way to represent robotic positioning and movement. On the right, there are options for adjusting the view and parameters, such as "Orbit (rviz_default)" and "Zero," which are likely tools for navigating and observing the simulated robotic environment. This setup is crucial for developers and engineers to test and refine robotic algorithms and movements.





The image depicts a computer screen with a booking system for a "Paradise Corridor Timed Entry Reservation." The user is attempting to book a ticket for the "Paradise Corridor" but is unable to proceed due to a "Too many people are requesting this tour at this time" error. The message suggests refreshing the page and trying again, indicating that the system is experiencing high demand for the available time slot.



The image shows a computer screen with an error message encountered while attempting to make a reservation for the Paradise Corridor at Mount Rainier National Park. The message indicates that the reservation could not be completed due to high demand, with too many people trying to book the tour simultaneously. It suggests refreshing the page and trying to submit the reservation again. This situation suggests that the website is experiencing heavy traffic, likely due to popular times or events, which is causing booking issues.