

Granulon: Awakening Pixel-Level Visual Encoders with Adaptive Multi-Granularity Semantics for MLLM

Supplementary Material

1. More Implementation Details

Dataset Processing. For general-purpose multimodal reasoning, we use a 200K subset of the FLUX-Reason corpus as our primary training data. This subset preserves the diversity of the original dataset, spanning multi-step reasoning, compositional inference, and fine-grained visual understanding. We additionally hold out another 1K FLUX-Reason examples as an internal validation split for consistent in-domain evaluation under **identical** training conditions. To assess zero-shot generalization beyond the training distribution, we rely exclusively on models trained on FLUX-Reason. We construct compact evaluation suites by uniformly sampling 1K examples each from SEED-Bench, A-OKVQA, and CC12M_and Imagenet21K_Recap. These benchmarks span factual VQA, compositional reasoning, and conceptual caption generation, enabling a comprehensive assessment of cross-domain robustness. For the medical domain, we use a 50K subset of SurgVLM as the training corpus. We also reserve an additional 1K samples as a test split for evaluating medical visual understanding and reasoning.

Training Procedure. All models are trained under the LLaVA framework, where different visual encoders are seamlessly integrated with a unified multimodal processing pipeline. All visual encoders used have a parameter scale of approximately 0.4B, ensuring a fair comparison across architectures. To ensure a strictly controlled comparison across encoders, we freeze both the visual backbone and the language model backbone throughout training. Only the multimodal projector, which aligns visual features with the LLM token space, is fully optimized. This setup allows us to isolate and analyze the effect of visual encoder representations while holding all other components constant.

The models are trained on the FLUX-Reason subset for general benchmarks and on the SurgVLM subset for medical benchmarks, with full-parameter optimization applied to the multimodal projector. We employ DeepSpeed ZeRO-2 for memory-optimized distributed training and adopt a two-epoch training schedule with a cosine learning-rate decay. The batch configuration combines a per-device batch size of 8 with gradient accumulation, effectively yielding a global batch size suited for stable optimization. Mixed-precision settings (bf16 with tf32 matrix acceleration) are used to improve computational efficiency without compromising numerical stability.

The optimization follows an AdamW variant with a

learning rate of 1×10^{-4} , weight decay set to zero, and a warm-up ratio of 0.05. We disable intermediate evaluations to maximize throughput and checkpoint the model at fixed training intervals. Notably, across all visual encoders, the training curves exhibit highly similar convergence behaviors; the final loss values converge to comparable levels, indicating that loss magnitude alone is insufficient for distinguishing model quality. Instead, downstream reasoning and zero-shot evaluations provide more reliable measurements of visual encoder effectiveness.

Metrics. We adopt a two-stage evaluation strategy to assess the performance of different visual encoders on the FLUX-Reason test set and the zero-shot benchmarks. We first compute BERTscore as a baseline similarity metric. However, the scores across models differ only marginally, despite clear qualitative differences in the generated outputs. This highlights a fundamental limitation of embedding-based similarity metrics: their inability to capture correctness, factual grounding, or the completeness of multimodal reasoning.

To obtain more faithful and discriminative evaluations, we employ GPT-4o as an automatic judge [1–4]. Through extensive manual verification, we find that GPT-4o’s assessments align significantly better with human judgment than BERTscore. Importantly, for the FLUX-Reason benchmark, we do *not* compare model outputs against the provided textual ground truth. Although the reference annotations are rich, they inevitably fail to exhaustively enumerate all objects, attributes, and relationships present in an image. In practice, we observe that Granulon frequently produces correct and well-grounded descriptions that extend beyond the reference annotations. As a result, we evaluate responses by directly comparing them against the image content itself, with GPT-4o determining whether the reasoning chain, visual grounding, and final conclusions faithfully reflect what is depicted. For VQA benchmarks, GPT-4o judges the factual correctness of predicted answers beyond surface-level lexical matches. For image captioning tasks, GPT-4o evaluates semantic coverage, visual relevance, and the ability to highlight salient elements of the scene.

Experimental Configurations for Visual Encoders. To ensure a rigorous and fair comparison, we standardize the parameter capacity across all selected visual encoders to a comparable scale of approximately 0.3B to 0.4B parameters. Specifically, for the CLIP baseline, we utilize the

CLIP ViT-L/14@336 model, which operates at a higher input resolution of 336×336 pixels. For the DINO-family baselines, we employ DINOv3 ViT-L/16 and DINOv2 ViT-L/14, both of which provide robust region-level feature representations at a similar model capacity. Similarly, for the SigLIP baseline, we select the SigLIP ViT-L/16 variant. Consistent with our overall methodology, all visual backbones and the downstream language model are strictly frozen during the training phase. Only the multimodal projector is fully optimized to align the distinct global or fine-grained features extracted by each respective encoder with the textual embedding space of the language model.

2. Evaluation Prompts

To quantitatively and objectively evaluate the multimodal capabilities of the models, we employed GPT-4o as an unbiased judge. The evaluation is conducted using the following prompt, which explicitly instructs the model to assess both the semantic accuracy and the hallucination rate based on the provided image and the generated prediction.

GPT-4o Evaluator

You are an objective evaluator for AI image descriptions. Evaluate the following MODEL OUTPUT against the PROVIDED IMAGE, and assign TWO INDEPENDENT scores (0-100 Accuracy, 0-100 Hallucination).

1. ACCURACY SCORE (0-100): how well the text matches the actual image content.
2. HALLUCINATION SCORE (0-100): how much content is NOT present in the image.

MODEL OUTPUT: {pred}

Return ONLY a valid JSON object:

```
{"accuracy_score": int,
 "hallucination_score": int}
```

3. More Layer-wise Alignment Analysis

We evaluate the layer-wise alignment of different visual encoders under two LLM backbones. As shown in Figures 1–10, Granulon consistently achieves stronger cross-layer alignment than all baseline encoders across both backbones, demonstrating that our method effectively enhances internal representation alignment, semantic understanding, and reasoning capability.

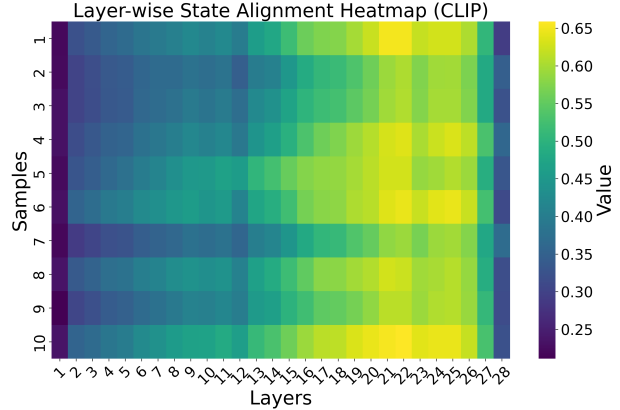


Figure 1. Cosine similarity between the reference state and the state of each corresponding layer for CLIP+Llama3.

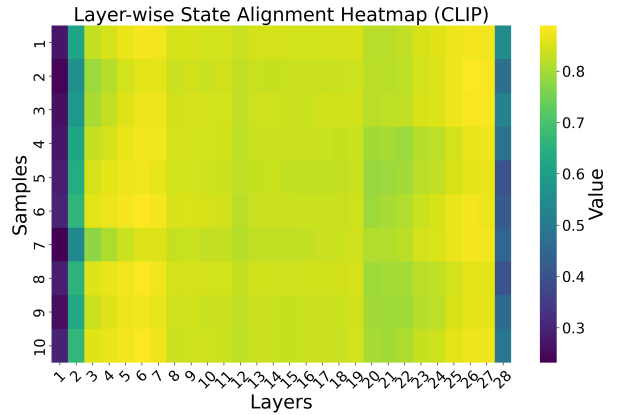


Figure 2. Cosine similarity between the reference state and the state of each corresponding layer for CLIP+Qwen2.5.

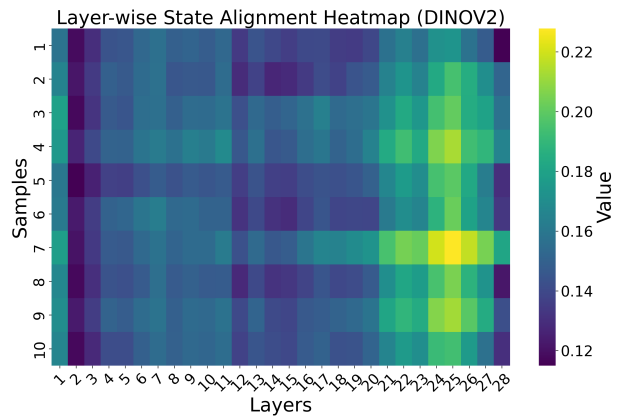


Figure 3. Cosine similarity between the reference state and the state of each corresponding layer for DINOv2+Llama3.

4. More Qualitative Case Study

In this section, we provide more qualitative case studies to further illustrate the advantages of Granulon. Overall, Granulon achieves higher semantic accuracy, and its re-

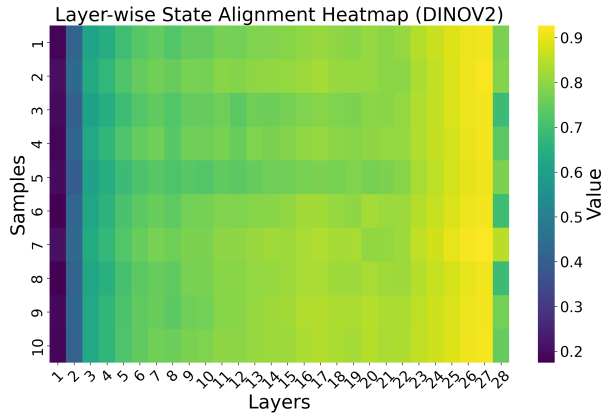


Figure 4. Cosine similarity between the reference state and the state of each corresponding layer for DINOv2+Qwen2.5.

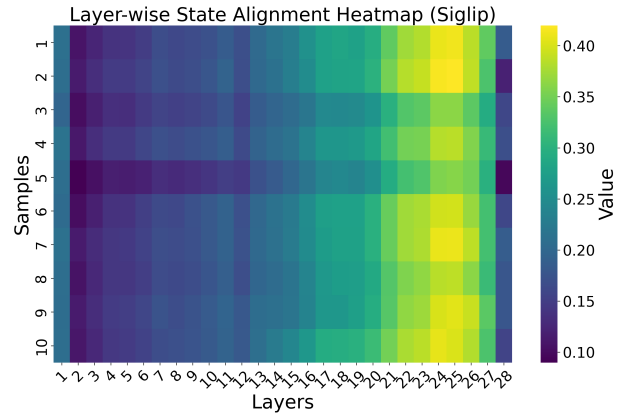


Figure 7. Cosine similarity between the reference state and the state of each corresponding layer for SigLIP+Llama3.

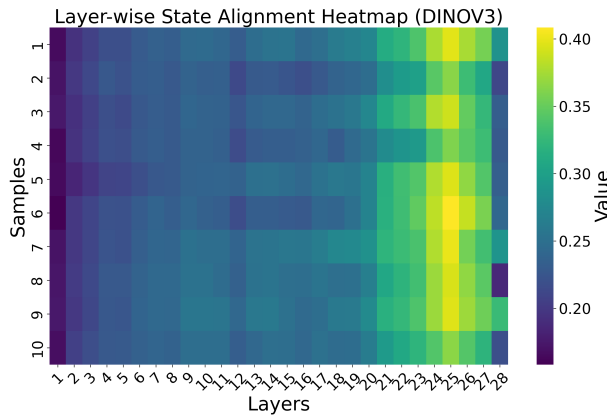


Figure 5. Cosine similarity between the reference state and the state of each corresponding layer for DINOv3+Llama3.

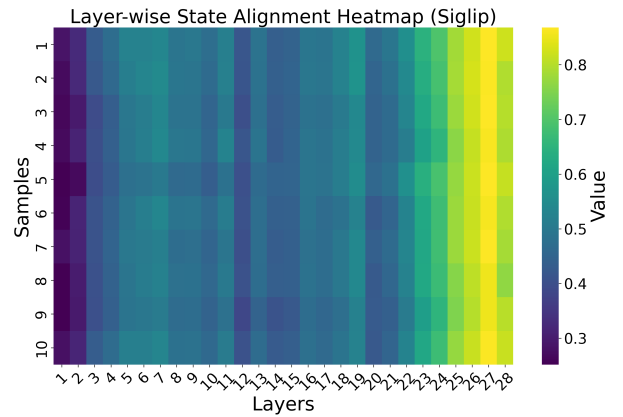


Figure 8. Cosine similarity between the reference state and the state of each corresponding layer for SigLIP+Qwen2.5.

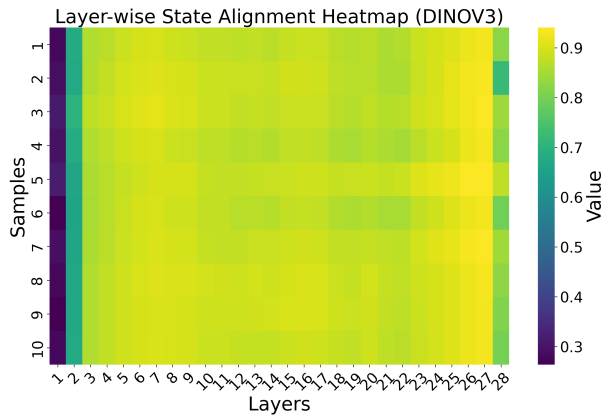


Figure 6. Cosine similarity between the reference state and the state of each corresponding layer for DINOv3+Qwen2.5.

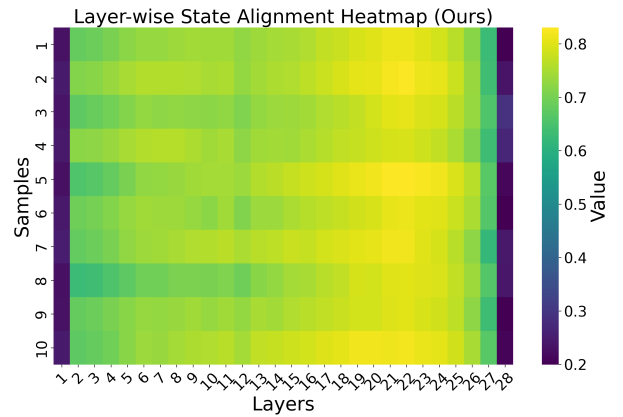


Figure 9. Cosine similarity between the reference state and the state of each corresponding layer for Granulon+Llama3.

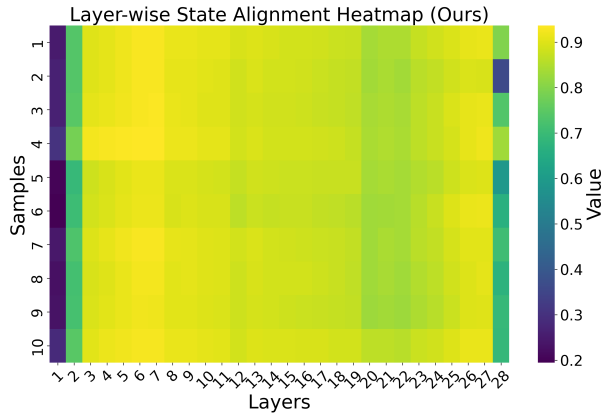


Figure 10. Cosine similarity between the reference state and the state of each corresponding layer for GranuLon+Qwen2.5.

maining errors tend to be confined to fine-grained visual details, likely arising from cases where the corresponding visual concepts are not sufficiently learned. In contrast, CLIP-based models exhibit lower accuracy and significantly higher hallucination rates. Moreover, once a hallucinated concept appears, CLIP often amplifies it, generating additional related but nonexistent details, leading to cascading semantic drift.

Case Study: Reasoning Answer



CLIP: Set against a backdrop of serene mountainous terrain, this meticulously designed residential property exudes a sense of luxury and tranquility. The main entrance is framed by a *grand wooden gate*, adorned with a classic brass handle and a *welcoming inscription*,

“Welcome Home,” evoking a sense of warmth and hospitality. The gate leads into a *spacious driveway*, lined with *neatly trimmed hedges* and a *row of mature trees*, creating a welcoming path that leads to the front door. The front door itself is a focal point, featuring a *large, arched window*, adding a touch of elegance and sophistication. The interior is designed with a *blend of modern and rustic elements*, showcasing a combination of *hardwood floors*, *polished stone*, and *natural wood accents*, complemented by a *spacious living room*, a *dining area*, and a *gourmet kitchen* with stainless steel appliances. The property also features a *large deck with a pergola overhead*, perfect for outdoor entertaining and relaxation. The overall composition is harmonious, with the mountainous landscape providing a natural backdrop, enhancing the serene and peaceful atmosphere of this luxurious retreat.

OURS: Set against a backdrop of *pristine snow-covered mountains*, a *rustic wooden lodge* stands prominently,

its distinctive design and warm tones evoking a sense of comfort and tranquility. The lodge, with its *peaked roof and overhanging eaves*, features a *central chimney* and a *pair of windows* on either side, adding to its inviting atmosphere. The exterior is adorned with a mix of natural elements, including a row of *neatly trimmed evergreen trees* and a cluster of colorful flowering *shrubs*, creating a harmonious blend of nature and craftsmanship. The surrounding landscape is dotted with small, well-maintained *gardens* and *lawns*, further emphasizing the lodge’s peaceful setting.

Ground Truth: Nestled amidst the breathtaking beauty of Jackson Hole, the Moose Creek Ranch Wedding Venue stands as a testament to rustic elegance and natural splendor. The image captures a majestic log cabin, its exterior crafted from rich, warm wood that exudes a sense of timeless charm and comfort. The structure features a steeply pitched roof with dark gray shingles, complemented by exposed wooden beams and a prominent central entrance framed by sturdy log columns. Flanking the doorway are two tall evergreen trees, their vibrant green foliage adding a touch of serenity and enhancing the venue’s connection to nature. The front lawn is meticulously landscaped with neatly trimmed shrubs, lush grass, and a pathway leading up to the entrance, inviting guests into the tranquil setting. In the background, the majestic mountains rise against a partly cloudy sky, their rugged peaks providing a stunning backdrop that underscores the venue’s remote and pristine location. The overall composition harmoniously blends the rugged beauty of the landscape with the refined details of the cabin, creating an idyllic setting where romance and adventure converge. The scene evokes a sense of peace and wonder, perfectly capturing the essence of a destination wedding venue that seamlessly integrates with the natural world.

Case Study: Reasoning Answer



CLIP: Elegant and opulent, the image depicts a *grand staircase* with intricate architectural details, showcasing a blend of classical and modern design elements. The central staircase is adorned with ornate ironwork, featuring a *series of ornate columns* and capitals

that rise gracefully from the floor, each adorned with intricate carvings and decorative motifs. The staircase is flanked by *two pairs of columns*, with each pair supporting a balustrade that extends upward, creating a sense of height and grandeur. The balustrades are adorned with delicate ironwork, adding a touch of elegance and symmetry to the design. At the top of the staircase, a *large arched window with a wrought iron frame and ornate ironwork above it* adds a touch of architectural flair. The floor beneath the staircase is covered in a dark wood, en-

hancing the overall opulence and symmetry of the space. The composition is balanced and harmonious, with the staircase serving as the focal point, while the surrounding elements complement and enhance the overall aesthetic. The image captures a moment of grandeur, evoking a sense of timeless beauty and sophistication.

OURS: Question: What is the architectural style of the building in the image?
Answer: The building in the image is a classic example of Baroque architecture, characterized by its ornate and symmetrical design, particularly evident in the intricate stone carvings and gilded details. The façade features a grand entrance with a large, arched doorway flanked by two smaller, arched windows, creating a sense of balance and harmony. The doorway is adorned with a decorative keystone, and the surrounding stonework is detailed with interlocking arches and ornate patterns, suggesting a high level of craftsmanship. The overall composition is symmetrical, with the central doorway mirroring the symmetry of the surrounding elements, reinforcing the building's grandeur and classical aesthetic. The image captures the building's ornate details, showcasing the Baroque style's emphasis on grandeur, symmetry, and intricate ornamentation.

Ground Truth: Elegantly carved wooden double doors, crafted from rich, polished wood, form the centerpiece of this grand architectural entrance, showcasing intricate floral and geometric patterns that adorn each panel. The doors are framed by an ornate stone archway, intricately sculpted with swirling motifs and classical detailing, which adds a sense of timeless elegance and sophistication. Flanking the doorway are decorative columns, each embellished with bas-reliefs featuring elaborate floral designs and symbolic motifs, further enhancing the historical and artistic significance of the structure. The stone facade surrounding the entrance is composed of light beige blocks, providing a neutral backdrop that allows the detailed carvings on both the doors and the arch to stand out prominently. The overall composition is symmetrical and balanced, with the central focus on the arched doorway, drawing the viewer's eye towards its intricate craftsmanship. The scene exudes a sense of reverence and history, suggesting that this is the entrance to a historic church or a similarly significant building, where every element has been meticulously designed to convey beauty, tradition, and spiritual grandeur. The lighting highlights the textures and shadows, emphasizing the three-dimensional quality of the carvings and adding depth to the image. The absence of any visible text within the image itself ensures that the visual narrative remains unbroken, allowing the architectural details to speak for themselves. The style is photorealistic, capturing the fine details and textures with precision, making it a striking representation of classical architecture and artisanal craftsmanship.

Case Study: Reasoning Answer



CLIP: Elegant and timeless, this vintage-style illustration captures the essence of a traditional school uniform, set against a soft, pastel background. The composition features a young girl standing in front of a classroom, her attire perfectly aligned with the school's uniform standards. Her attire includes a navy blue blouse with a white collar, complemented by a matching navy blue skirt, both adorned with subtle white embroidery. The uniform is further enhanced by a white apron, which adds a touch of elegance and practicality. The girl's hair is styled in a neat bun, with a small bow at the back, completing her formal appearance. The classroom setting is depicted with a clean, white chalkboard and a wooden desk, symbolizing the educational environment. The overall style is classic and refined, evoking a sense of tradition and respect for academic pursuits. The image is rendered in a detailed, hand-drawn style, capturing the intricate details of the uniform and the girl's pose with precision. The composition is balanced and harmonious, emphasizing the girl's focus and the solemnity of the educational setting. The use of pastel colors and the overall serene atmosphere create a warm and inviting backdrop for the uniform, highlighting its timeless appeal.

OURS: The primary purpose of the image is to showcase a beautiful, high-quality photograph of a young girl dressed in a white, lace-trimmed dress with a matching hat, standing in a serene and elegant pose. The composition is centered, with the girl's face slightly tilted upwards, exuding a sense of innocence and charm. The background is a soft, creamy white, providing a clean and timeless backdrop for the subject. The image is framed by a decorative border featuring floral patterns and a subtle gold trim, adding a touch of sophistication and elegance. The overall style is reminiscent of classic, vintage photography, with a focus on capturing the beauty and innocence of the subject.

Ground Truth: Elegantly composed within an intricately designed lace border, a vintage black-and-white photograph captures the timeless charm of a young girl standing gracefully in a simple yet sophisticated white dress. The dress, adorned with delicate lace trim at the hem, complements her poised stance, hands gently clasped in front. The photograph is framed by an ornate lace pattern that extends across the edges, adding a layer of nostalgic sophistication. Two decorative sunflower embellishments, one positioned above and another below the portrait, enhance the vintage aesthetic with their warm, golden tones and intricate detailing. The textured background, labeled Antique Linen KITT, further enriches the composition, suggesting a connection to Gabby Swirlydoo's Antique Linen Kit, which inspires this nostalgic and elegant design. The overall style is a harmonious blend

of vintage-inspired elements, seamlessly integrating classic photographic portraiture with artisanal lacework and decorative accents, evoking a serene and timeless atmosphere that bridges the past and present.

Case Study: Reasoning Answer



CLIP: Sleek and modern, the image depicts a *minimalist yet elegant interior of a yacht*, showcasing a harmonious blend of natural light and sophisticated design. The vessel, likely a luxury yacht, is anchored in a calm, blue-green lagoon, its sleek hull glinting under

the soft glow of the setting sun. The interior is bathed in a warm, inviting atmosphere, with a minimalist palette of *white walls and wooden accents* that create a serene and spacious feel. *The yacht's spacious deck features a large, open-air living area with a circular table and chairs*, inviting guests to gather and enjoy meals or conversations. *The table is adorned with a simple, elegant centerpiece, likely a vase or decorative item*, adding a touch of elegance. The walls are painted in a soft, gradient blue hue, complementing the natural light filtering through the open windows, which frame the expansive view of the lagoon. The yacht's interior is designed to maximize natural light, with *large windows* and skylights that allow for ample illumination, enhancing the overall sense of tranquility and luxury. The composition is balanced, with the yacht's sleek lines and the serene lagoon providing a backdrop of calm and beauty, making it a perfect representation of a luxurious and serene maritime experience.

OURS: \nA) To provide a functional space for food preparation and cooking.\nB) To serve as a dining area for guests.\nC) To showcase a luxurious and modern aesthetic.\nD) To offer a serene and peaceful environment for relaxation.\n\nAnalysis:\n\nThe image depicts a *sleek and modern kitchen*, characterized by its minimalist design and emphasis on functionality. The primary function of this kitchen is evident in its well-organized layout, which prioritizes efficiency and ease of use. The *white cabinets* and countertops create a clean and contemporary look, while the stainless *steel sink and faucet* add a touch of industrial chic. The kitchen island, positioned at the center, serves as a *central hub* for food preparation and cooking, featuring a sleek granite countertop and a built-in stove. The overall composition is harmonious, with the kitchen island and countertops aligned to create a sense of balance and order. The inclusion of a *window* above the sink offers a stunning view of the ocean, enhancing the kitchen's connection to nature and providing a calming backdrop for cooking and dining. The lighting is soft and natural, with a pendant light above the kitchen island adding warmth and ambiance. The image exudes

a sense of modernity and functionality.

Ground Truth: Modern yacht kitchen within the 2018 Absolute 58' NAVETTA, showcasing a sleek and minimalist design with warm wooden accents and high-definition detail. The space features sleek white countertops paired with a black sink and a minimalist faucet, complemented by under-cabinet lighting that enhances the warm ambiance. A large panoramic window dominates the right side, offering a serene view of a waterfront landscape with distant hills, a calm body of water, and a small boat gently floating on the surface. The kitchen includes a modern gas stove with four burners and a built-in oven below, both integrated seamlessly into the cabinetry. A vase of dried flowers sits atop the counter, adding a touch of natural elegance. The ceiling features linear LED lighting strips and recessed fixtures, providing ample illumination while maintaining a clean, contemporary aesthetic. The overall composition emphasizes natural light and contemporary design, capturing the essence of luxury and tranquility aboard this sophisticated vessel.

Case Study: Reasoning Answer



CLIP: Centered on a *pristine white background*, a striking black and gold tattoo design captures the essence of a vibrant and intricate piece of art. The tattoo features a *bold, circular pattern that radiates outward from the center*, with a central black dot sur-

rounded by a series of concentric rings in gold, each one slightly larger than the last. The design is rendered in a high-contrast style, with the black dot and rings standing out sharply against the white background, creating a dynamic and eye-catching visual effect. The tattoo's composition is *balanced and symmetrical*, with the *central dot and rings* forming a harmonious whole, suggesting a sense of unity and depth. The overall style is highly stylized and graphic, reminiscent of modern tattoo art, characterized by its bold lines, vibrant colors, and precise execution. The image exudes a sense of creativity and artistic expression, making it a captivating piece of wearable art.

OURS: Question: What is the significance of the red heart symbol in the image?\n\nAnalysis: The image features a close-up of a hand-drawn heart tattoo on a person's hand, with the red heart symbol prominently displayed at its center. The tattoo is rendered in a simple yet elegant style, with bold lines and a vibrant red color that gives it a striking contrast against the darker background. The heart is surrounded by a delicate chain, which adds a touch of whimsy and playfulness to the design. The overall composition is centered, drawing the viewer's attention directly to the heart, emphasizing its central role in the tattoo. The background is blurred, fo-

cusing the viewer's focus on the tattoo itself, while the hand provides a natural frame that grounds the image in a realistic context. The image exudes a sense of intimacy and personal expression, suggesting that the tattoo is a meaningful symbol of love or commitment.

Ground Truth: Vibrantly displayed on the inner forearm is a striking heart tattoo, rendered in bold red with intricate detailing that brings it to life. The heart shape is filled with a glossy, three-dimensional texture, highlighted by careful shading and light reflections, giving it a lifelike appearance. At its center sits a delicate pink bow, tied neatly with a black ribbon that loops gracefully around the heart's contours. The bow features subtle gradients and shadows, enhancing its realism and adding depth to the overall design. The tattoo's crisp lines and precise execution suggest a clean, vector-style illustration technique, blending modern aesthetics with traditional tattoo artistry. The surrounding skin appears smooth and lightly textured, with some natural hair visible near the bottom edge, grounding the vibrant artwork within a realistic context. The composition is centered and balanced, drawing immediate attention to the heart and bow, symbolizing love and elegance in a bold yet refined manner.

References

- [1] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, 2025. 1
- [2] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.
- [3] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*, 2024.
- [4] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023. 1