

# X-Fusion: Introducing New Modality to Frozen Large Language Models (Supplementary Material)

In the supplementary material, we present implementation details (Section C) and additional experiment results and analysis (Section D). We also discuss the societal impact of our method (Section E) and limitations (Section F). For sections and figures, 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. We hope this document complements the main paper.

## A. Pretrained Representations for Vision Feature Regularization

Inspired by [1], we explore whether aligning vision features in our X-Fusion with a pretrained encoder (*e.g.*, CLIP [2]) can improve X-Fusion’s generation and understanding. Following [1], we align the vision features from the 8th layer of X-Fusion with the penultimate features extracted from a pretrained CLIP model. Specifically, we take the vision feature  $\mathbf{H}^{\text{img}}$  from the dual tower and project it using a trainable linear projection  $\mathbf{W}$  to match the feature dimension of CLIP. The alignment loss is then computed as the cosine distance between the projected feature and the CLIP feature, encouraging our intermediate vision features to closely resemble those of CLIP. The regularization loss is defined as  $\mathcal{L}_{\text{align}} = 1 - \cos(\mathbf{W}\mathbf{H}^{\text{img}}, \mathbf{H}^{\text{CLIP}})$ , where  $\mathbf{H}^{\text{CLIP}}$  is the vision feature extracted from the pretrained CLIP encoder, and  $\cos(\cdot, \cdot)$  represents the cosine similarity.

To evaluate effectiveness and scalability, we tested this approach on our dual-tower model with three different base model sizes: 1B, 3B, and 8B. We use  $\lambda_{\text{AR}} = 0.5$  in this ablation study. As Fig. B shows, alignment accelerates training and improves performance but with two key findings: (1) Its impact diminishes with model size, even slightly degrading 8B model performance at 100k iterations. (2) Regardless of size, alignment loss imposes a performance ceiling, likely set by the external encoder’s representational power. These findings highlight that aligning with external representations is particularly beneficial for smaller models, but the benefit may be less for larger models. Future work could explore whether using a more powerful external encoder (*e.g.*, a larger or improved vision backbone) might further push this performance boundary.

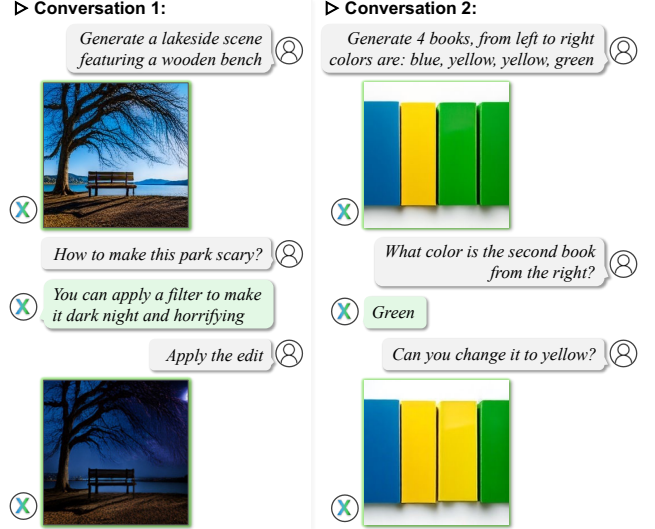



Figure A. **Interactive Generation.** Our X-Fusion model can follow user instructions to understand, generate, or edit images.

 Aligning with pre-trained vision features (*i.e.*, CLIP) improves performance for smaller models but has significantly less impact on larger ones.

## B. Extension of X-Fusion

### B.1. Fine-tune X-Fusion

Our X-Fusion model has been pre-trained on T2I and I2T tasks and has achieved strong cross-modal performance. Can X-Fusion extend its capabilities to other downstream vision-and-language tasks? We fine-tuned our model on four tasks simultaneously—including image editing, localization, outpainting, and Visual Question Answering (VQA)—using internal datasets for 20k training steps. In Figure C, we demonstrate that our unified X-Fusion model can handle multiple tasks without creating task-specific models or weights. We evaluated image-editing performance on publicly available datasets, including PIE-Bench [3] and SmartEdit [4]. Notably, X-Fusion showcased strong instruction-based editing capability when challenged to select between multiple objects—such as “smaller

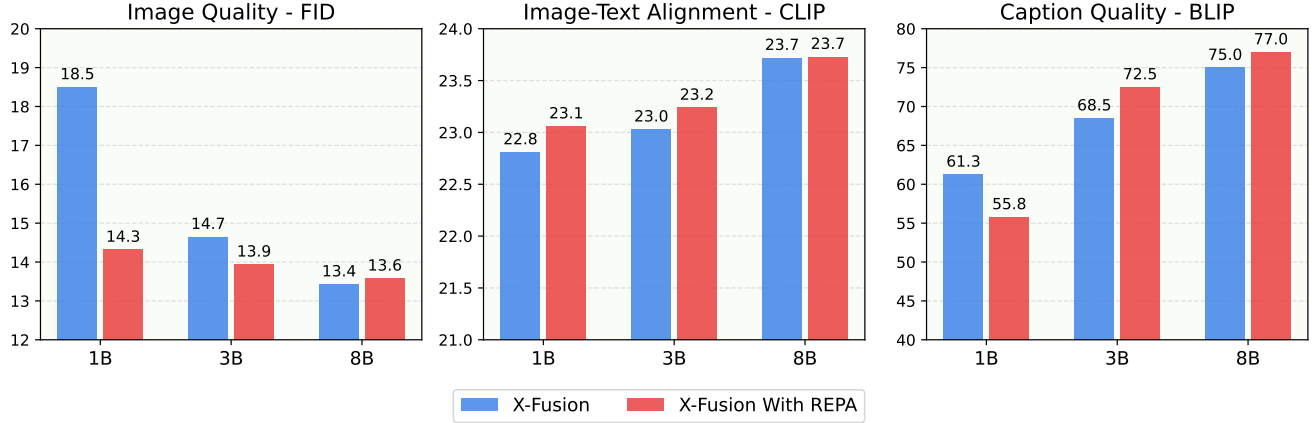


Figure B. **Performance comparison of models of different sizes (1B, 3B, and 8B) with and without additional feature alignment loss.** The effectiveness of alignment diminishes as model size increases.

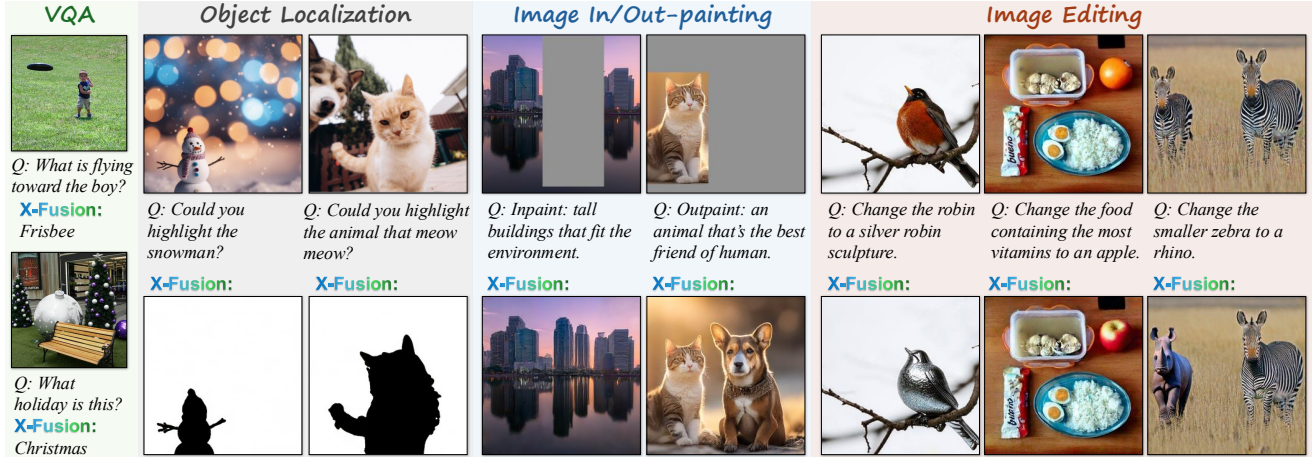


Figure C. **Qualitative results** of fine-tuned X-Fusion model on downstream tasks including: visual question answering (VQA), image editing, localization and outpainting tasks.

zebra” and “food containing the most vitamins.” Furthermore, driven by X-Fusion’s generalized capability across vision-and-language tasks, we unlock new interactive vision–language applications by enabling a single model to both generate, understand, and edit images with natural-language instructions from users, as illustrated in Figure A.

## C. Implementation Details

In this section, we present the implementation details, including the model, experiments, and the choice of evaluation metrics.

### C.1. Model Details

Our X-Fusion uses the pre-trained VAE model with the compression ratio of 4 from Stable Diffusion [5] and follows the flow matching training from Stable Diffusion 3. During inference, we use the classifier-free guidance scale

of 5.5 as we empirically find that it provides optimal visual results.

### C.2. Caption Quality Evaluation Metric

In our main paper, we use pairs of images and long captions (generated from InternVL [6]) as training data for both text-to-image (T2I) and image-to-text (I2T) tasks. The use of long captions encourages the model to generate more detailed images, while also forcing the model captures fine-grained semantic details for understanding tasks. For generation evaluation, we use standard metrics like FID [7] and CLIP scores [2]. However, we discovered that the CIDEr score [8], commonly used to evaluate caption quality in recent papers such as Chameleon and Transfusion, is not suitable for long captions. Therefore, it is not an ideal metric for monitoring our ablation results.

The CIDEr score, which relies on n-gram overlap, tends to give lower scores for long captions. This is because

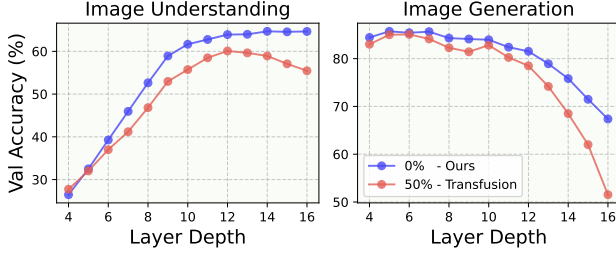


Figure D. **Linear Probe Results.** We use the trained model as a feature extractor and train an additional linear layer for image classification on ImageNet [10]. Models trained with our training strategy constantly obtain higher feature quality.

COCO ground truth captions are short, and even accurate long captions may have minimal n-gram overlap, leading to a lower score. Our evaluation shows that the CIDEr score between human-written COCO short captions and InternVL long captions is only  $1e-5$ , an insignificantly low value. We also evaluate other metrics, such as BertScore [9] and CLIP score [2]. However, neither is well-suited for assessing long captions. Figure E presents qualitative examples where the reference caption is a human-written COCO ground truth caption. Among the generated captions, Sample 1 is a long and informative caption, while Sample 2 is short and of low quality. As shown in the figure, both BertScore and CLIP score assign relatively high scores to these two vastly different captions, making it difficult to gauge the quality of training progress. In contrast, the BLIP score effectively differentiates between the two captions, making it a more suitable evaluation metric.

### C.3. Linear Probing Experiments

Following [1], we analyze the model’s behavior by conducting layer-wise feature representation experiments through linear probing. We report the Top-1 accuracy on the ImageNet dataset [10]. As shown in Fig. D, using clean images results in better feature representation for understanding tasks compared to noisy images (Transfusion setting). Similarly, the generative features are also superior, leading to improved generation results.

For the implementations details, we followed previous literature [1, 11] for the details of the linear probing experiment settings. We use the trained Dual Tower model as the feature extractor and train an additional classification head along with a parameter-free normalization layer. Specifically, we extract the image representation from the vision tower instead of the language tower. We used a batch size of 16,384 with the Adam optimizer without weight decay and set the learning rate to  $6.4 \times 10^{-3}$ . We report the top-1 accuracy on the validation dataset.

As the objective of the linear probing experiments is to examine the quality of the learned visual features, we consider two settings for extracting the visual representation

when the model is used for visual understanding tasks and visual generation tasks. For the understanding mode, we provide clear images without noise and set the corresponding timestep to  $t = 0$ . For the generation mode, since X-Fusion is partially trained with text-to-image generation tasks, we add a text prompt preceding the noisy image and set the corresponding timestep to  $t = 20$ .

## D. Additional Experiments Results

In the main paper, we systematically explored the architectural design choices, data types, and other training factors to identify an effective approach for training a unified model with a frozen language model. Our experiments primarily focused on 1B parameter models. In this supplementary material, we extend our study by training the model with the LLaMA-3.1-8B architecture using our final training recipe. Specifically, we employ a Dual Tower design, with a 2:1 data ratio for generation and understanding tasks. Additionally, we ensure that image-to-text samples are created with cleaned images and without feature regularization from an external encoder. The model is trained with 0.8M tokens per batch for 200k iterations. Table A compares our model’s performance with other state-of-the-art unified models. Our model achieves comparable captioning and image generation quality while maintaining its original language capabilities. We also provide additional qualitative results for image generation and image understanding in Figure F and G, respectively. As shown in Figure F, our X-Fusion achieved great visual quality and image-text alignment. Also, Figure G illustrates the strong performance of X-Fusion for visual understanding.

## E. Social Impact

Our work presents an efficient approach to integrating new modalities into frozen LLMs, enabling both image understanding and generation. By systematically studying architectural and data-centric design choices, we provide insights that enhance the scalability and efficiency of multimodal learning, reducing computational costs and making such models more accessible. This has potential applications in assistive AI, creative content generation, and vision-language understanding, benefiting areas like accessibility and education.

However, multimodal AI also raises ethical concerns, including bias in training data, misinterpretation of images, and potential misuse of generated content. To ensure responsible deployment, future work should focus on dataset fairness, robustness evaluations, and ethical safeguards. Our research contributes to the development of scalable and responsible multimodal AI, promoting more adaptable and efficient vision-language models.


Image	Reference	Caption	BertScore	CLIP	BLIP
	1. This is the view of a kitchen stove, kitchen sink, counter, and cabinets. 2. A wooden kitchen center island with a rug in front of it. 3. The sink is on the island of a large kitchen. 4. A kitchen with a stove a sink and a counter 5 . A kitchen with a sink, stove, flower vase and wine rack.	<b>Sample 1:</b> This image showcases a modern kitchen with a neutral color palette. The space features light wood cabinetry, a stainless steel sink, and an array of hanging pendant lights. A vase with flowers adds a touch of nature, and a patterned rug anchors the flooring. The room is well-lit, with natural light streaming through windows and the ceiling fans providing circulation.	0.86	30.9	91.2
		<b>Sample 2:</b> This image is a photo of kitchen.	0.89	30.9	10.4

Figure E. **Comparison of Different Evaluation Metrics.** The BLIP score effectively differentiates between the two captions with varying levels of detail, whereas the other metrics do not.

Method	Base LLM	Language MMLU ↑	Image Understanding COCO BLIP ↑	Image Generation COCO FID ↓
<b>Language Generation Only</b>				
Llama2 7B		45.3	-	-
LLaMA-3.1 8B		66.7	-	-
LLaMA-3.2-Vision 11B		66.7	82.3	-
InternVL2.0-26B		-	81.1	-
<b>Visual Generation Only</b>				
Stable Diffusion		-	-	9.6
<b>Unified Models</b>				
Emu-3		35.3	79.6	12.8
Janus	DeepSeek1.3B	-	70.8	8.5
Chameleon-7B		52.1	54.1	26.7
Show-O	Phi-1.5B	-	-	9.2
Transfusion		-	-	6.7
LLaVAFusion	LLaVA-Next8B	-	-	8.2
MetaMorph	LLaMA-3.1 8B	-	-	11.8
X-Fusion (Ours)	LLaMA-3.1 8B	66.7	80.0	11.5

Table A. **Quantitative Comparison with State-of-the-Art Models.** X-Fusion achieves competitive performance compared to other leading unified models. For the image understanding task, we prompt open-sourced models to generate detailed captions.

## F. Limitation

X-Fusion is not without limitations. First, like other vision-language models, it is prone to hallucinations, occasionally generating inaccurate or misleading outputs (Fig. H). This may be due to the model operating in the latent space of LDM [12], which might not accurately capture fine details. A potential solution is to use a more robust latent VAE encoder. Second, although our model has shown promising results, there is still room for improvement in image quality, such as by training with higher-resolution images. Third, X-Fusion doubles the number of model parameters, thereby reducing training efficiency compared to Transfusion [13].

Nevertheless, this paper serves as a study that explores and analyzes various factors, including architecture and training strategies for unified MLLM training, offering valuable insights for future research.





A photo of a tranquil night scene with a yellow crescent moon and serene village.



A photo of a large gray elephant underwater on the ocean floor with vibrant marine life.



A croissant-shaped spaceship traveling through a galaxy of bakery planets.



A photo of royal raccoon portraits in a grand castle.



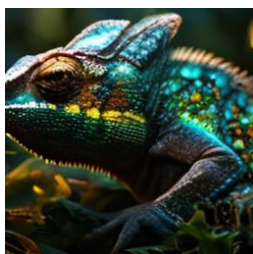
A photo of a robot artist painting in a modern art studio.



A colossal ancient tree with a door carved into its trunk.



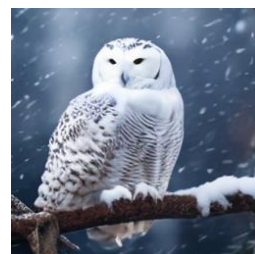
A photo of a black and white dog observing Picasso-style art in a museum.



A mystical chameleon that shifts its colors to reflect ancient runes.



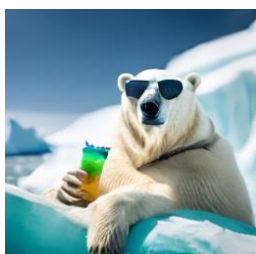
A pumpkin patch with glowing pumpkin lanterns.



A snowy owl perched on a branch under falling snowflakes.



A photo of an astronaut on a camel in an otherworldly desert.



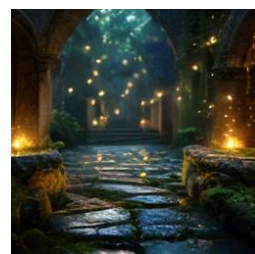
A polar bear wearing sunglasses, sipping a cold drink on an iceberg.



A museum that only exists in dreams, vanishing upon waking.



A smiling cat with a monocle and top hat sipping tea.



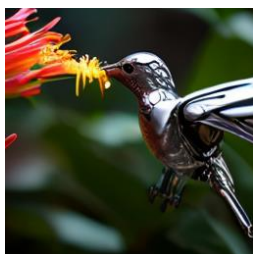
A mystical garden where enchanted fireflies illuminate ancient stone pathways.



A giant mushroom city inhabited by tiny fairies and gnomes.



A photo of a red dog lying in a cozy living room, gazing at a vibrant potted plant.



A robotic hummingbird drinking nectar from a mechanical flower.



A photo of a clear glass fish tank in a tranquil underwater scene with vibrant fishes.



A lone telephone booth in the middle of a desert, ringing persistently.

Figure F. **Additional qualitative results of X-Fusion model.** Our generation samples demonstrated good visual quality and achieved great image-text alignment.

▷ Input Image:



Row of brown elephants is seen walking in single file along a road, possibly in a national park, with each elephant having a rider. The setting is natural, with greenery in the background, and the elephants appear calm.

▷ Input Image:



In the image, a man is leading a buffalo in a grassy field, holding a colorful umbrella to shield them both. The buffalo seems calm and is close to the man, with other buffaloes in the background, indicating a herd. In the distance, there are people, suggesting a public event or fair. The scene is lively, with other individuals and activities in the vicinity.

▷ Input Image:



Three elderly men are seated on park benches in an outdoor setting. The man on the left wears a light-colored shirt and dark pants, the middle man is dressed in a white shirt and dark pants, and the man on the right is in a black shirt and dark pants. They all have white hair, and the setting includes a building with arched windows in the background.

▷ Input Image:



A grey and white cat stands on a grassy area near a curb, gazing intently to the side. Behind the cat, there are two black cars parked next to each other. The scene suggests a quiet street setting, with the cat possibly alerted to something out of frame.

▷ Input Image:



In the image, a group of individuals is engaged in winter sports atop a snowy slope. A prominent figure in a blue jumpsuit and yellow helmet is skiing downhill, while other snowboarders and skiers stand or move along the slope in the background. The scene is one of outdoor activity and camaraderie, with the individuals dressed in colorful winter gear, typical for skiing and snowboarding.

▷ Input Image:



Woman with red hair is holding a vibrant red umbrella with a white pattern, which is open above her head. She appears to be walking outdoors, possibly on a rainy day, with a blurred background that suggests an urban environment.

▷ Input Image:



Two women are seen from behind, each carrying a surfboard, as they walk along a sandy beach. The ocean waves are crashing in the background, and other surfers are visible, some actively riding the waves. The sky is overcast, suggesting a cool or humid day.

▷ Input Image:



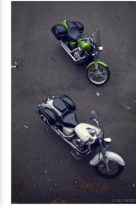
In the image, a black and white dog stands in the foreground, facing away from the camera, with its tail slightly raised. It is positioned in a grassy field that stretches towards a calm body of water, likely the sea, visible in the distance. A flock of white sheep is scattered across the grass, grazing peacefully. The sky is clear, suggesting a calm, sunny day.

▷ Input Image:



Image displays a neatly arranged bathroom featuring a bathtub with a wooden cabinet next to it. Above the bathtub, there is a shower head. To the right, there is a sink with a wooden cabinet below it, and the sink appears to be positioned next to a toilet. The tiles are light beige, and there is a window providing natural light to the room.

▷ Input Image:



Two motorbikes are parked on a dark asphalt road, captured from a high-angle perspective. The top bike is green with black seats, while the bottom bike is primarily white with black accents. Both bikes feature luggage racks at the back, and the one on the top appears to be equipped with additional protective gear.

▷ Input Image:



In a serene park setting, a wooden bench with metal supports casts long, artistic shadows on the pavement. The bench is situated on a grassy area, surrounded by lush greenery. In the background, cars are parked near a curb, and the scene is bathed in the soft glow of sunlight, suggesting a peaceful morning.

▷ Input Image:



Woman with a cheerful expression stands in front of the grand, Gothic-style Duomo di Milano in Italy. She is holding a yellow umbrella, providing a pop of color against the grey, overcast sky. The Duomo's intricate details and tall spires are visible in the background, creating a striking contrast between the historical architecture and the woman's modern attire.

Figure G. **Additional qualitative results of X-Fusion model.** Our generation captions demonstrated good text quality and achieved great image-text alignment.



▷ Input Image:



*A gray tabby cat with striking yellow eyes is the focal point, lying down on a white surface with a beige bed in the background. To the left of the cat, there's a small, brown, and white **Chihuahua** sleeping in a pet bed. Both animals exude a calm demeanor, with the tabby cat's gaze directed towards the viewer. The scene is set in a room with a workbench in the background, creating a cozy and comfortable atmosphere for the pets.*

*[the dog is a Shiba Inu]*

▷ Input Image:



*An adult giraffe stands in the center of the image, its long legs firmly planted amidst a bustling urban setting. The giraffe is positioned on a road flanked by parked cars, with a **red SUV** visible on the far side. The sky above is overcast, casting a muted light over the scene. The surrounding environment suggests a juxtaposition of wildlife and city life, with the giraffe taking center stage amidst the chaos of cars and buildings.*

*[there is no visible "red SUV"]*

Figure H. **Limitations.** Similar to other large multimodal models, our model is prone to hallucinating in its generated concepts.

## References

- [1] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *ArXiv*, abs/2410.06940, 2024. [1](#), [3](#)
- [2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. [1](#), [2](#), [3](#)
- [3] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. [1](#)
- [4] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024. [1](#)
- [5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. [2](#)
- [6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. [2](#)
- [7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. [2](#)
- [8] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. [2](#)
- [9] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. [3](#)
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. [3](#)
- [11] Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. *ArXiv*, abs/2401.14404, 2024. [3](#)
- [12] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. [4](#)
- [13] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. 2024. [4](#)