Transferability Between Understanding and Generation in Unified Multimodal Models

Jiwon Kang[1*], Heeji Yoon[1*], Jaewoo Jung[1], Jaewon Min[1], Minkyeong Jeon[1], Biyeon Hwang[2], Sangwon Jung[3†], and Seungryong Kim[1†]

[1] KAIST AI, South Korea [2] Blynx, South Korea [3] Trillionlabs, South Korea

Abstract. Unified Multimodal Models (UMMs) integrate image understanding and generation within a single architecture, yet how the two tasks interact remains understudied. We investigate **transferability** in UMMs: whether training a capability on one task improves the same capability on the other without explicit supervision. Through controlled experiments, we empirically find that transferability depends on architecture—models with fully shared transformer backbone and a unified visual encoder exhibit consistent cross-task transfer, while loosely coupled designs show little or none. Leveraging this transferability, we propose a practical training strategy. The most straightforward way to improve a target generative capability (*e.g.*, counting) is to fine-tune generation directly, but this can degrade visual quality due to distribution shift. Instead, we train the corresponding understanding task and let it *transfer* into generation, which improves capability-specific generative performance while minimizing distribution shift. We validate this across three capabilities—counting, spatial relation, and text recognition/generation—showing that cross-task transferability can be systematically exploited in UMMs.

Keywords: Unified Multimodal Models · Cross-Task Transferability · Image Understanding · Image Generation

1 Introduction

Propelled by the success of Large Language Models (LLMs) [3,33,62,79], Vision-Language Models (VLMs) [5, 16, 20, 37, 54, 101] have achieved significant breakthroughs in image understanding. Concurrently, diffusion-based generative models [10,65,67] have driven tremendous advancements in image generation. Despite their profound progress, these two domains have mostly evolved in isolation. This historical separation has recently motivated the development of Unified Multimodal Models (UMMs) [13, 15, 18, 21, 93, 94], which integrate both tasks—image understanding and generation—within a single architecture. While recent UMMs demonstrate that such unification can achieve performance comparable to specialized experts for each task, it remains fundamentally unclear how the two tasks interact once they are trained within a single model.

* Equal contribution.
† Co-corresponding author.

Several recent studies [24, 52, 90, 95] have investigated interactions between image understanding and generation, arguing that understanding can improve generation and vice versa. However, existing analyses typically examine these interactions through aggregate performance changes on general benchmarks [30, 34, 51, 102] (*e.g.*, adding generation training improves understanding benchmark scores). While such results provide useful high-level observations, relying on broad benchmark improvements makes it difficult to determine whether the observed gains stem from genuine cross-task knowledge transfer or merely from indirect factors such as additional data or regularization effects.

We address this gap by studying a more direct and actionable signal of cross-task interaction: *transferability*. Instead of inferring interaction from aggregate benchmark gains, we examine whether training a given capability (*e.g.*, counting) in one task (understanding or generation) reliably improves the same capability in the other task, under controlled interventions. Through a series of experiments, we find that such cross-task transferability emerges in current UMMs, and this study leads to two key findings: (i) the existence and strength of transferability vary across models, empirically revealing a dependency on architectural design; (ii) when transfer is present, it can be exploited as a practical mechanism to improve targeted generative capabilities by training on the corresponding understanding task.

Specifically, we begin with a controlled analysis on architecture using *counting* as a probe task. Across several representative open-source UMMs [13, 15, 21, 93], we empirically observe that transferability depends on architectural design. In particular, models that employ a fully shared transformer backbone along with a unified visual encoder [93] exhibit consistently the strongest transfer between tasks, whereas architectures that leverage a separate backbone or an image encoder [13, 15] show little or no transfer. These results suggest that the degree of representation sharing may play an important role in enabling cross-task capability transfer.

Using the architecture with the strongest transfer, our next step is to investigate whether transferability can be utilized as a practical mechanism for enhancing generative capabilities. While directly fine-tuning with a generation objective may be the most straightforward approach to acquire a specific capability, it carries the risk of inducing a distribution shift [12, 47, 59] that may degrade overall visual quality. Instead, we find that training on the corresponding understanding task can effectively introduce the capability into generation through transfer, minimizing the distribution shift. Extensive experiments across multiple visual capabilities, including counting, spatial relation, and text recognition/generation, confirm that cross-task transfer can consistently improve generation performance on specific capabilities while preserving general multimodal understanding ability and generative quality.

Our contributions are summarized as follows:

- We introduce **transferability** as a tool for analyzing cross-task interactions in Unified Multimodal Models, and show that capabilities learned in understanding or generation can be transferred to the other task.

- Through controlled experiments, we empirically demonstrate that transferability depends on architectural design, emerging most clearly in models with fully shared transformer backbones with a unified visual encoder.
- We show that transferability can be leveraged as a practical strategy for improving targeted generative capabilities while avoiding degradation of original multimodal understanding ability or generative quality, and validate this approach across counting, spatial relation, and text recognition/generation tasks.

2 Related Work

2.1 Image Understanding and Generation in UMMs

The success of large language models [3, 33, 62, 79] has catalyzed the extension of language modeling into the multimodal domain. Vision-Language Models (VLMs) [5, 16, 20, 37, 54] enable strong image understanding by pairing a visual encoder with an LLM. More recently, Unified Multimodal Models (UMMs) have pushed this paradigm further by incorporating image generation alongside understanding within a single model [13, 21, 76, 88, 93].

Visual representations. A key design axis in UMMs is the choice of visual representation for encoding input images. One line of work adopts language-supervised encoders such as CLIP [66] and SigLIP [80, 103], whose representations are aligned with text and thus preserve high-level semantics amenable to reasoning. A complementary line leverages autoencoders [6] which are typically employed for image generation. These autoencoders, which prioritize pixel-level reconstruction fidelity, are usually trained with variational objectives [43] for continuous representation [10, 25, 45, 65, 67] or a vector quantization objective [81] for discrete representation [26, 81]. Analyses in Janus [15, 88] and related works [77] show that language-supervised representations substantially outperform compression-oriented tokenizers for understanding tasks, motivating several UMMs to employ separate encoders for each task.

Generation objectives. Image generation in UMMs follows two main paradigms. *Autoregressive* models [15, 18, 44, 76, 88, 90] employ discrete image tokenizers and generate image tokens sequentially, directly extending the next-token prediction framework of LLMs. *Diffusion*-based models instead learn a denoising objective [2, 39, 53, 56, 73, 74] and can be further divided into discrete-token diffusion models [70, 93, 100] which apply discrete diffusion [4, 58, 68] over quantized latent representations, and continuous diffusion models [13, 24, 32, 75, 92, 110] which operate on continuous latent representations.

Architectural taxonomy of unified multimodal models. Although finer-grained categorizations are possible [109], we group existing unified models into three dominant architecture families for our analysis. **(1) Single transformer** designs route all modalities through a shared transformer backbone, maximizing parameter sharing [15, 18, 70, 76, 85, 88, 93, 100]. These models either use unified image encoders [70, 76, 85, 93, 100] or separate ones [15, 88] for understanding/generation. **(2) Mixture-of-Transformers (MoT)** designs maintain modality-specific transformer branches but couple them through joint attention layers, allowing selective parameter sharing while retaining specialized capacity [21,49,84,87]. **(3) Decoupled transformer** designs employ separate transformers for language and vision, where the LLM provides conditioning signals (*e.g.*, hidden states) that are subsequently fed into a dedicated diffusion module for generation [13, 32, 63, 75, 78].

2.2 Understanding Interactions Between Understanding and Generation in UMMs

A central question in designing Unified Multimodal Models is whether multimodal understanding and generation genuinely benefit each other within a shared architecture. Several works report potential synergy, arguing that high-level semantic representations learned for understanding can improve the semantic fidelity of generation, while fine-grained visual representations acquired through generation may enhance visual grounding and compositional reasoning [44, 52, 89, 90, 93, 97, 106]. On the other hand, research such as UniFork [49] empirically demonstrates that ideal representations for these two tasks throughout the network layers are distinct, suggesting that simple parameter sharing can lead to a representational compromise. This perspective aligns with a broader trend of decoupled architectures that utilize task-specific branches in higher layers or separated visual paths to mitigate potential conflicts [15, 21, 88].

While these studies provide valuable insights into the potential benefits and trade-offs of joint training, most analyses examine interaction through aggregate performance changes on general benchmarks. For example, prior work often evaluates whether increasing generation-oriented training improves overall understanding benchmarks (*e.g.*, MMMU [102] or POPE [51]), or conversely whether additional understanding supervision improves generation benchmarks such as GenEval. Although such evaluations provide useful high-level signals, they make it difficult to isolate how specific capabilities learned in one modality influence the other. In contrast, our work investigates cross-task interaction at the capability level through the lens of transferability. Rather than inferring interaction from aggregate benchmark improvements, we directly examine whether learning a capability in one modality improves the same capability in the other modality.

3 Does knowledge transfer emerge in UMMs?

In this section, we investigate cross-task transfer in current open-source UMMs: (i) whether learning a capability in one task (understanding or generation) improves the same capability in the other, and (ii) whether this transfer depends on architectural design. To this end, we conduct an analysis using counting as a probe capability, which provides a clean and quantifiable concept that both understanding and generation should encode.

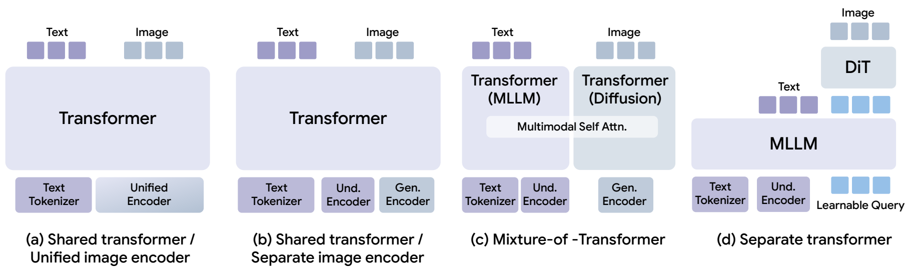


Fig. 1: Architecture of chosen unified multimodal models for cross-task transferability analysis on counting. (a) Shared Transformer with a unified image encoder. (b) Shared Transformer with separate image encoders for each task (understanding/generation). (c) Mixture-of-Transformer which uses separate transformers for understanding and generation but connects each other via joint(multimodal) self-attention. (d) Separate Transformer where the MLLM transformer provides conditioning signals that are fed into a diffusion transformer (DiT).

Train → Test	**(a) Lumina-DiMOO**		**(b) Janus-Pro**		**(c) BAGEL**		**(d) BLIP3-o**	
	Acc. (%) ↑	MAD ↓	Acc. (%) ↑	MAD ↓	Acc. (%) ↑	MAD ↓	Acc. (%) ↑	MAD ↓
Generation Evaluation								
Baseline	48.0	1.11	38.0	1.46	42.0	1.16	31.0	2.08
und → *gen*	57.0 **(+9.0)**	0.83 **(-0.28)**	32.0 (-6.0)	1.55 (+0.09)	47.0 (+5.0)	1.11 (-0.05)	37.0 (+6.0)	1.80 (-0.28)
Understanding Evaluation								
Baseline	22.0	2.48	37.0	2.70	63.0	0.65	63.0	0.54
gen → *und*	30.0 **(+8.0)**	1.88 **(-0.60)**	38.0 (+1.0)	3.34 (+0.64)	64.0 (+1.0)	0.62 (-0.03)	62.0 (-1.0)	0.56 (+0.02)

Table 1: Probing cross-task transferability on counting across different UMM architectures. We conduct two separate fine-tuning experiments and compare them against the baseline. (i) *und* → *gen*: training only on understanding and evaluating on generation. (ii) *gen* → *und*: training only on generation and evaluating on understanding. Acc. denotes exact-match accuracy (%) and MAD denotes mean absolute deviation from the ground-truth count. Shared transformer models with a unified visual tokenizer (Lumina-DiMOO) show the strongest transferability in both directions.

3.1 Experimental setup

Considering the diverse architectural choices introduced in Section 2.1, we select four representative open-source models: Lumina-DiMOO [93] for **(a) Shared Transformer with unified image encoder**, Janus-Pro [15] for **(b) Shared Transformer with separate image encoder**, BAGEL [21] for **(c) Mixture of Transformers**, and BLIP3-o [13] for **(d) Separate Transformer**. The architectural overview of each selected model is illustrated in Figure 1. We select 7B–8B parameter versions of each model.

For each model, we conduct two separate training runs (understanding and generation) using LoRA-based supervised fine-tuning. We construct training datasets by filtering PixMo-Points and PixMo-Count [20] to a counting range of 0–20. For understanding, the model learns counting through visual question an-

Fig. 2: Qualitative results of *und* → *gen* transferability across models. For each prompt, we compare the baseline model with the model trained only on counting understanding data (*und* → *gen*) and evaluated on generation. All images are generated using the same random seed for fair comparison.

swering, where it answers how many objects appear in an image. For generation, the model learns counting through text-to-image generation, where it is supervised to produce images containing objects whose count matches the prompt. To get the generation prompts, we use Qwen3-VL [5] to caption the same images and append object count information to form the generation prompts.

For evaluation, we use two separate protocols. Understanding is evaluated on the test split of PixMo-Count [20] in a VQA format, where the model is asked to count target objects in an image and responds with a number. For generation, we follow the counting protocol from GenEval [34] and construct prompts in the form `"A photo of [NUMBER] [OBJECT]s"` with counts ranging from 2 to 10. Generated images are evaluated using an object detector [17] to verify whether the number of detected objects matches the specified count. We report both accuracy and mean absolute deviation (MAD) to capture exact correctness and proximity to the ground truth.

3.2 Transferability across architectures

Transferability varies across architectures. Table 1 summarizes the results. The most consistent finding is that knowledge transfer from understanding to generation, which we term *understanding-to-generation transfer* (*und* → *gen*), is observed in all models except Janus-Pro [15]. BAGEL [21], Lumina-DiMOO [93], and BLIP3-o [13] all show improved generation accuracy and lower MAD when trained with the understanding objective, indicating that counting capability acquired through understanding can be transferred to generation. In contrast, the inverse direction ($gen \rightarrow und$) is more selective: only BAGEL and Lumina-DiMOO show gains in counting understanding from generation training, making transfer bi-directional in these two models. Among all models, Lumina-DiMOO shows the strongest effect in both directions, underscoring that transferability is not uniform across UMM architectures. Qualitative examples illustrating this behavior are shown in Figure 2.

Which architecture results in the strongest transferability? From an architectural perspective, we empirically observe that models which employ a fully shared transformer backbone for both understanding and generation, along with a unified vision encoder (Figure 1-(a)), such as Lumina-DiMOO [93], exhibit the strongest transferability. This result is consistent with recent empirical findings that adopting a unified encoder for image understanding and generation produces synergistic effects in UMMs [89,90]: because a shared transformer backbone with a unified encoder lets both tasks operate over a shared representation, knowledge acquired through one objective naturally propagates to the other, facilitating cross-task transfer. By contrast, architectures that maintain separate visual pathways for each task offer fewer opportunities for such propagation, which explains the weaker or absent transfer observed in models like Janus-Pro.

4 Using Transferability to Improve Generative Capabilities

In this section, we investigate how the transferability that emerged in UMMs can be practically exploited. In practice, generative models often exhibit weaknesses in specific capabilities such as spatial relation [14,29,36,55,64,86,91] or object counting [8,11,31,50]. Generation objectives [2,39,53,56,73,74] fundamentally learn to approximate their training data distribution, so a straightforward way to inject a specific skill (*e.g.*, counting) is fine-tuning directly on a skill-focused dataset with generation objectives. However, fine-tuning pulls the model away from its original image generation distribution, which can cause a distribution shift that may degrade generative quality [12,47,59].

This issue is even more pronounced for recent models whose generation distribution is carefully aligned with human preferences in a final post-training stage, using SFT over curated high-quality datasets [19,65] or RL-based methods [9,27,46,48,83,98]. However, direct fine-tuning readily destroys this alignment. Workarounds would be either curating a dataset that is simultaneously high-quality and skill-rich, or repeating the costly post-training alignment after fine-tuning, neither of which is practical.

We propose a practical alternative that leverages transferability. Rather than optimizing the generation objective directly, we train the target capability through

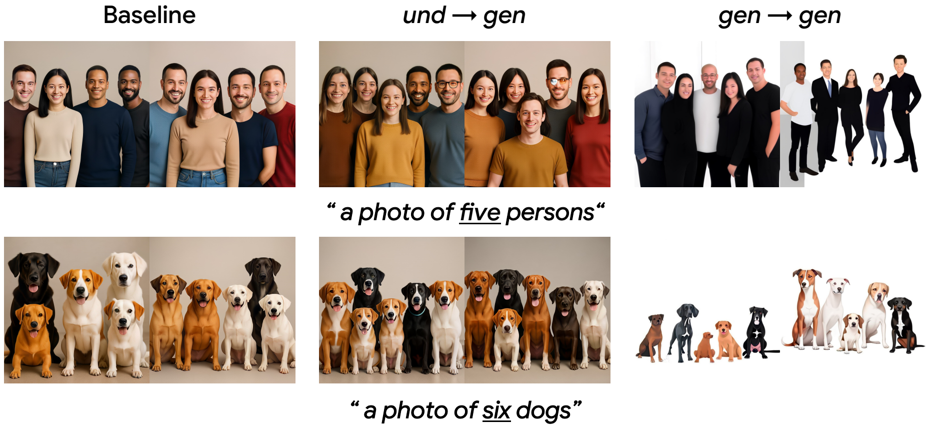


Fig. 3: Qualitative comparison of understanding-to-generation transfer (*und* → *gen*) versus direct generation training (*gen* → *gen*) on counting. While direct generation training (*gen* → *gen*) significantly degrades image quality due to distribution shift, *und* → *gen* preserves generation quality while achieving improved counting accuracy.

Train → Test	Acc. (%) ↑	MAD ↓	IS [69] ↑	FID [38] ↓
Baseline	48.0	1.11	16.86	-
und → *gen*	**57.0** (+9.0)	**0.83** (-0.28)	**17.55**	**31.47**
gen → *gen*	57.0 (+9.0)	0.91 (-0.20)	15.29	52.51

Table 2: Quantitative comparison of understanding-to-generation transfer (*und* → *gen*) versus direct generation training (*gen* → *gen*) on counting. Using transferability (*und* → *gen*) attains task accuracy on par with direct training (*gen* → *gen*) while retaining image quality (IS [69], FID [38]). In contrast, *gen* → *gen* degrades both metrics.

the understanding and let it transfer to generation. This approach enables targeted improvements in generative capabilities while maintaining the overall generation quality of the model. To study this mechanism, we analyze understanding-to-generation transfer across three representative tasks, *i.e.*, counting, spatial relation, and text recognition/generation. We conduct the following experiments on Lumina-DiMOO [93], as its single-transformer design with a unified image encoder empirically exhibited the strongest cross-task transfer among the UMMs studied in Section 3. Result on another model [100] sharing the same architecture is reported in the Appendix.

4.1 Counting

We first revisit the counting generation from this perspective. Table 1 previously showed that counting exhibits strong transferability from understanding to generation. To examine whether this transfer can be practically beneficial, we compare the model trained with the understanding objective (*und* → *gen*) with a model directly trained with the generation objective (*gen* → *gen*) in Table 2.

Surprisingly, the model trained through understanding achieves a lower mean absolute deviation (MAD) in counting generation. In other words, introducing counting capability through the understanding leads to more accurate counting behavior in generation than directly optimizing the generation objective.

The qualitative results in Figure 3 further highlight this difference. When the model is trained with the generation objective ($gen \rightarrow gen$), the narrow distribution of the dataset causes the model to drift from its original training distribution, resulting in noticeable degradation of overall image quality. In contrast, the model trained via understanding ($und \rightarrow gen$) preserves the overall image quality. To quantify this, we additionally measure distribution shift using Fréchet Inception Distance (FID) [38] and image quality using Inception Score (IS) [69] in Table 2. Specifically, FID is computed between the baseline generated images and those from the trained models. Training directly with the generation objective ($gen \rightarrow gen$) yields a higher FID than $und \rightarrow gen$ and degrades IS relative to the baseline, confirming both a larger distribution shift and loss of generative quality. In contrast, the understanding-based approach ($und \rightarrow gen$) maintains generation quality with lower distribution shift, while improving counting accuracy, demonstrating that transfer through understanding provides a more stable way to introduce new capabilities into generation.

4.2 Spatial relation

We next consider a structurally different capability, spatial relation, where the model must encode the relative layout of objects. Unlike counting, this task requires capturing compositional relationships between multiple entities. This allows us to examine whether the transferability observed in counting extends to more complex relational concepts.

Experimental setting. We observe that pretrained models already handle simple spatial relations such as left, right, top, and bottom reasonably well. We therefore focus on a more challenging setting: diagonal relations between two objects (*e.g.*, top-left, top-right, bottom-right, bottom-left). Since existing spatial reasoning datasets [23,28,40,72,96,99,107] are difficult to use directly for our purpose—owing to their limited scale, mismatched spatial complexity, or inconsistent annotation formats—we construct a synthetic dataset of 200K samples from ImageNet [22]. For each sample, as shown in Figure 4 we select two class–image pairs and place them on a canvas according to a randomly sampled diagonal direction. Each sample is paired with a prompt of the form `"[class A] is located in the [direction] of [class B]"`.

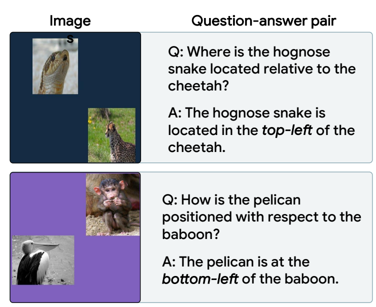


Fig. 4: Example of synthetic dataset for spatial relation.

For evaluation, we use the same T2I prompt format as the training data and follow GenEval [34] protocol. We apply a detection model [17] to localize

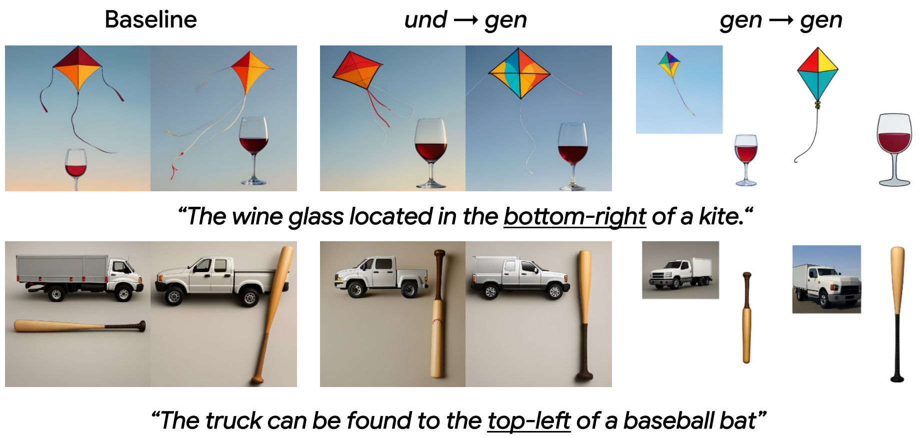


Fig. 5: Qualitative comparison between *und* → *gen* and *gen* → *gen* on spatial relation tasks. While generation-only training (*gen* → *gen*) significantly degrades image quality reflecting traits of synthetic training data, *und* → *gen* preserves generation quality while achieving improved spatial relation accuracy.

Train → Test	**Acc.** ↑	**IS** [69] ↑	**FID** [38] ↓
Baseline	67.0	16.86	-
und → *gen*	74.0 (+7.0)	**17.50**	**30.95**
gen → *gen*	**80.0** (+13.0)	17.41	32.28

Table 3: Quantitative comparison of understanding-to-generation transfer (*und* → *gen*) versus direct generation training (*gen* → *gen*) on spatial relation. Both *und* → *gen* and *gen* → *gen* improve spatial accuracy over the baseline, with *und* → *gen* attaining a substantial gain (+7.0%) from transferred understanding supervision alone, approaching direct generation training (+13.0%). *und* → *gen* also stays closer to the baseline image distribution (lower FID)

bounding boxes of the two objects and compute the center points of the detected boxes and verify whether their relative position matches the specified direction. More details of the dataset construction and evaluation pipeline can be found in the Appendix.

Results. Table 3 confirms that understanding-to-generation transferability also holds for spatial relation. The model trained through understanding (*und* → *gen*) achieves higher spatial accuracy than the baseline, showing that relational knowledge transfers from understanding to generation. Training directly on generation data (*gen* → *gen*) yields slightly higher accuracy but at the cost of a higher FID. Qualitative results in Figure 5 illustrate this trade-off: *gen* → *gen* drifts toward the training distribution, producing images that resemble cropped objects pasted onto a canvas rather than natural scenes, whereas *und* → *gen* improves spatial correctness while preserving realistic appearance.

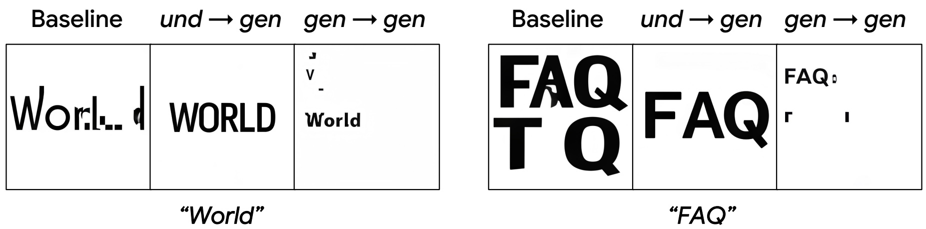


Fig. 7: Qualitative comparison between *und* → *gen* and *gen* → *gen* on text generation task. *und* → *gen* generates clear and accurate text, preserving rendering fidelity.

Train → Test	WER ↓	CER ↓	Edit Dist. ↓	BLEU ↑	METEOR ↑	F1 ↑
Baseline	0.654	0.362	36.812	0.277	0.403	0.446
und → *gen*	0.641 (-0.013)	0.351 (-0.011)	35.636 (-1.176)	0.274 (-0.003)	0.410 (+0.007)	0.452 (+0.006)
gen → *gen*	0.641 (-0.013)	0.367 (+0.005)	37.312 (+0.500)	0.354 (+0.077)	0.533 (+0.130)	0.525 (+0.079)

Table 4: Quantitative comparison of understanding-to-generation transfer (*und* → *gen*) versus direct generation training (*gen* → *gen*) on text generation. We apply an OCR model [1] to the generated images for text extraction and compute WER, CER, edit distance, BLEU, METEOR, F1 between the extracted text and the original prompt. *und* → *gen* transfer yields improvements in most metrics.

4.3 Text recognition and generation

We next examine text-contained image generation, where the model must render legible characters with accurate strokes, spacing, and glyph structure. Unlike counting or spatial relation, this requires the model to capture highly localized and fine-grained visual details. We investigate whether such fine-grained capabilities can be introduced through understanding and transferred into generation.

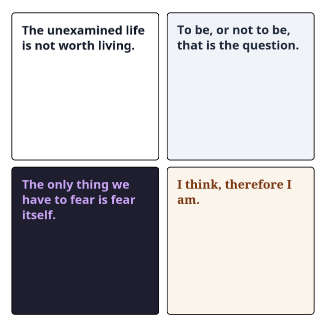


Fig. 6: Synthetic dataset for text recognition and generation.

Experimental setting. We find that existing OCR datasets tend to focus on either too small, sign-level text in natural scenes [71,82,108] or dense, full-page documents [61]. Both may be too hard for models based on discrete image tokenizers, as quantizing such images into discrete tokens discards much of the fine-grained visual details. Thus, we construct a synthetic dataset of 200K image-text pairs by rendering Markdown documents into images (Figure 6). Our rendered dataset provides diverse text content at a moderate difficulty level. For evaluation, we apply an OCR model [1] to read text from the generated image and compute five metrics against the original prompt: word error rate (WER) and character error rate (CER) measure exact correctness at the word and character level.

Edit Distance captures the minimum number of operations needed to align the prediction with the ground truth. METEOR [7] and F1 assess partial matches by accounting for token overlap and ordering. Details of the dataset construction and evaluation pipeline can be found in the Appendix.

Results. Table 4 summarizes the results. The model trained through understanding (*und* $\rightarrow$ *gen*) shows improvements over the baseline on most metrics, including WER, CER, Edit Distance, METEOR and F1, confirming that understanding-to-generation transfer also exists for text generation. The qualitative examples in Figure 7 further verify this: the understanding-trained model produces more legible text compared to the baseline. While direct training with generation objective (*gen* $\rightarrow$ *gen*) yields larger gains in text accuracy, this comes at the cost of generation quality. These results together indicate that the transfer-based approach remains a safer alternative that avoids degrading generative quality when boosting specific visual capabilities.

5 Discussion

While our experiments demonstrate that understanding-to-generation transfer can effectively improve generative capabilities, several natural questions remain. In this section, we further analyze the properties and implications of transferability in UMMs. We first investigate whether the reverse direction, *i.e.*, generation-to-understanding transfer, can also provide meaningful improvements. We then examine whether jointly training understanding and generation tasks could serve as an alternative to transfer-based training. Finally, we analyze whether our understanding training negatively affects general multimodal understanding ability.

5.1 Can generation-to-understanding transfer also be useful?

While the previous section demonstrated that *und* $\rightarrow$ *gen* transfer can be exploited to improve specific generative capabilities, we observed that the strength of transferability is not uniform across capabilities. To quantify the strength, we assume direct (*e.g.*, generation) training on the same dataset as the upper bound for transfer (*e.g.*, *und* $\rightarrow$ *gen*). We then measure transfer strength as the ratio of the improvement from transfer to that from direct training, with the results presented in Table 6. We found that for text generation, the transfer strength of *und* $\rightarrow$ *gen* is overall lower than that of counting and spatial relation. This raises the question of what factors determine how well a capability transfers.

We hypothesize that the strength of *und* $\rightarrow$ *gen* transfer depends on the nature of the visual knowledge each capability requires. Counting and spatial relation rely on relatively high-level, structural concepts, such as numerical quantity and relative object layout. Consequently, transferring from understanding to generation for these tasks does not strictly require pixel-level precision. The text capability, by contrast, intrinsically demands highly fine-grained visual details, such as stroke geometry, character spacing, and glyph composition. Such

Train $\rightarrow$ Test	Counting		Spatial Rel.	Text Recognition/Generation					
	Acc. (%) $\uparrow$	**MAD $\downarrow$**	**Acc. (%) $\uparrow$**	**WER $\downarrow$**	**CER $\downarrow$**	**Edit Dist. $\downarrow$**	**BLEU $\uparrow$**	**METEOR $\uparrow$**	**F1 $\uparrow$**
Baseline	22.0	2.48	61.0	0.129	0.073	7.464	0.835	0.899	0.885
gen $\rightarrow$ *und*	30.0 (+8.0)	1.88 (-0.60)	67.0 (+6.0)	0.118 (-0.011)	0.061 (-0.012)	6.232 (-1.232)	0.845 (+0.010)	0.907 (+0.008)	0.892 (+0.007)
und $\rightarrow$ *und*	57.0 (+35.0)	0.89 (-1.59)	89.0 (+28.0)	0.091 (-0.038)	0.042 (-0.031)	4.276 (-3.188)	0.875 (+0.040)	0.928 (+0.029)	0.912 (+0.027)

Table 5: Generation-to-understanding transfer (*gen* $\rightarrow$ *und*) versus direct understanding training (*und* $\rightarrow$ *und*) on counting, spatial relation, and text recognition for Lumina-DiMOO. Across all three capabilities, *gen* $\rightarrow$ *und* transfer consistently improves over the baseline.

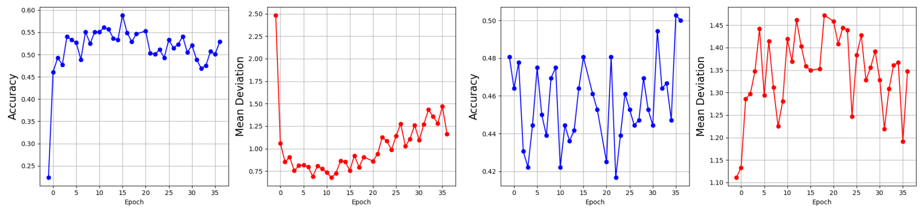
Transfer Direction	Counting Acc.	Counting MAD	Spatial Rel. Acc.	Text Recog./Gen. WER	Text Recog./Gen. METEOR	Text Recog./Gen. F1
und $\rightarrow$ *gen*	**100%**	**140%**	**53.8%**	**100%**	5.4%	7.6%
gen $\rightarrow$ *und*	22.9%	37.7%	21.4%	28.9%	**27.6%**	**25.9%**

Table 6: Transfer strength comparison across tasks and directions. Assuming direct training as an upper bound that transfer can achieve, we quantify transfer strength as the ratio of the improvement obtained by transfer to that obtained by direct training. For counting and spatial-relation, *und* $\rightarrow$ *gen* transfer is stronger than *gen* $\rightarrow$ *und*, whereas for text recognition/generation the opposite holds, indicating that the dominant transfer direction may be capability-dependent.

fine-grained precision is difficult to acquire purely through the understanding objective [35, 41, 42, 104, 105], which often relies on partial or contextual cues to perform recognition without fully encoding the exact visual details needed for rendering. This gap between the coarse-grained representations learned through the understanding objective [60, 90] and the fine-grained precision inherently required by the text capability explains the weaker *und* $\rightarrow$ *gen* transfer.

This reasoning leads to a natural follow-up question: if text generation requires pixel-level precision, and text recognition fundamentally demands the same level of detailed visual discrimination, does the *reverse transfer direction (gen $\rightarrow$ und) work more effectively* for such capabilities? Since the generation objective inherently provides more detailed supervision than the understanding objective, we expect that learning to generate text forces the model to capture the fine-grained representations highly beneficial for text recognition.

To investigate this, we extend our evaluation to the remaining capabilities—text and spatial relation—by reporting the *gen* $\rightarrow$ *und* transfer performance on Lumina-DiMOO alongside direct training (*und* $\rightarrow$ *und*) for comparison (Table 5), and calculate their respective transfer strengths (Table 6). Consistent with our hypothesis, Table 6 demonstrates that while *und* $\rightarrow$ *gen* transfer remains stronger for counting and spatial relation, the reverse direction (*gen* $\rightarrow$ *und*) more robustly improves performance for text capabilities compared to *und* $\rightarrow$ *gen*. These findings confirm that the dominant direction and strength of transferability are capability-dependent and governed by the granularity of supervision required.



(a) Quantitative results of $und + gen \to und$

(b) Quantitative results of $und + gen \to gen$

Fig. 8: Trade-off in joint *und*+*gen* training. (a) Counting understanding accuracy and mean absolute deviation (MAD) over training epochs. (b) Counting generation accuracy and MAD over training epochs. Blue lines denote accuracy and red lines denote MAD.

5.2 Does joint training lead to stronger transferability?

Throughout this work, we examined transferability in UMMs and showed that training a capability through one task can improve it in the other task. This naturally raises the question of whether training both tasks simultaneously leads to improved performance for both.

To investigate this, we trained Lumina-DiMOO on understanding and generation jointly ($und + gen$) for counting. Results shown in Figure 8 indicate that the two objectives do not improve simultaneously during training. Instead, their performance exhibits an oscillatory pattern. Understanding accuracy increases in the first half of training while generation accuracy declines, and the trend reverses in the latter half. This behavior suggests that the presence of transferability does not necessarily translate into consistent gains when the two are optimized simultaneously.

One possible explanation is that understanding and generation require different alignment patterns within the network, which may lead to conflicting optimization signals during joint training, as suggested by [49]. At the same time, our experimental setting may limit how clearly synergy can be observed. We conduct our experiments as post-training on a pretrained model whose capabilities are already largely established, which may lead to different training dynamics from those that arise during large-scale pre-training. In addition, the joint training uses the same concept-focused dataset constructed for the individual experiments in Section 3, where each image-text pair is tightly aligned around *counting*. Such paired data distributions are uncommon in typical multimodal pre-training corpora and may introduce an unfamiliar training configuration when the two objectives are optimized together. Distinguishing between these possibilities is difficult within the scope of post-training experiments and therefore remains an open question.

5.3 Does understanding training harm general multimodal ability?

A potential concern of capability injection through understanding training is that it may negatively affect the model's overall multimodal ability. Since our

Train (*und*)	**POPE** [51]	**MMBench** [57]	**MMMU** [102]	**MME-P** [30]
Baseline	86.7	90.8	57.7	1534
Counting	86.4	90.9	62.0	1492
Spatial Relation	87.3	90.8	61.3	1554
Text recognition	86.7	90.5	61.3	1538

Table 7: Performance on general multimodal benchmarks after understanding-based training. Despite introducing task-specific capabilities through understanding training for $und \rightarrow gen$ transfer, the model's general multimodal understanding remains largely preserved across all benchmarks.

approach introduces additional supervision targeting specific capabilities, one might expect such training to overfit to the target task or degrade performance on general multimodal benchmarks.

To examine this, we evaluate the models on several widely used multimodal understanding benchmarks, including POPE [51], MMBench [57], MMMU [102], and MME [30]. The results are summarized in Table 7. Across all benchmarks, we observe that understanding-based training does not degrade general multimodal performance. The scores on POPE and MMBench remain nearly unchanged compared to the baseline model, while performance on MMMU even improves slightly after capability-focused training. These results suggest that injecting capabilities through understanding tasks introduces only minimal changes to the model parameters while preserving the model's general multimodal knowledge.

6 Conclusion

In this work, we studied cross-task transferability between understanding and generation in UMMs. Through controlled experiments, we empirically showed that capabilities transfer between the two tasks and that the strength of this transfer depends on architectural design. We further exploited this as a practical strategy: introducing a target capability through understanding, rather than fine-tuning generation directly, improves capability-specific performance while preserving overall generative quality. We hope our findings provide new insights into the interaction between understanding and generation and open promising directions for post-training in unified multimodal models.

Acknowledgements

This research was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279, RS-2025-II212068, RS-2023-00227592, RS-2025-02214479, RS-2024-00457882, RS-2025-25441838, RS-2025-02217259, RS-2026-25519202) and the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2024-00333068), and National Research Foundation of Korea (RS-2024-00346597).

References

1. AI, Z.: Glm-ocr. `https://github.com/zai-org/GLM-OCR` (2026), accessed: 2026-07-01
2. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. arXiv preprint arXiv:2209.15571 (2022)
3. Anthropic: Introducing the next generation of claude (2024), `https://www.anthropic.com/news/claude-3-family`, accessed: 2026-07-01
4. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. Advances in neural information processing systems **34**, 17981–17993 (2021)
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
6. Ballard, D.H.: Modular learning in neural networks. In: Proceedings of the sixth National conference on Artificial intelligence-Volume 1. pp. 279–284 (1987)
7. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
8. Binyamin, L., Tewel, Y., Segev, H., Hirsch, E., Rassin, R., Chechik, G.: Make it count: Text-to-image generation with an accurate number of objects. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13242–13251 (2025)
9. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. In: International Conference on Learning Representations. vol. 2024, pp. 4965–4987 (2024)
10. Black Forest Labs: FLUX.2: Analyzing and enhancing the latent space of FLUX – representation comparison (2025), `https://bfl.ai/research/representation-comparison`, accessed: 2026-07-01
11. Boo, H., Kim, H., Lee, M., Lee, S., Lee, J., Choi, J.H., Cho, H.: Countsteer: Steering attention for object counting in diffusion models. arXiv preprint arXiv:2511.11253 (2025)
12. Chae, J., Kim, J., Choi, J., Kim, K., Hwang, S.: Apt: Adaptive personalized training for diffusion models with limited data. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28619–28628 (2025)
13. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025. URL https://arxiv.org/abs/2505.09568
14. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5343–5353 (2024)
15. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
16. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 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. pp. 24185–24198 (2024)

17. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
18. Cui, Y., Chen, H., Deng, H., Huang, X., Li, X., Liu, J., Liu, Y., Luo, Z., Wang, J., Wang, W., et al.: Emu3. 5: Native multimodal models are world learners. arXiv preprint arXiv:2510.26583 (2025)
19. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
20. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 91–104 (2025)
21. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
22. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
23. Deng, N., Gu, L., Ye, S., He, Y., Chen, Z., Li, S., Wang, H., Wei, X., Yang, T., Dou, M., et al.: Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. arXiv preprint arXiv:2506.18385 (2025)
24. Dong, R., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., Kong, X., et al.: Dreamllm: Synergistic multimodal comprehension and creation. In: International Conference on Learning Representations. vol. 2024, pp. 6666–6702 (2024)
25. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
26. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
27. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. Advances in Neural Information Processing Systems **36**, 79858–79885 (2023)
28. Fan, Y., He, X., Yang, D., Zheng, K., Kuo, C.C., Zheng, Y., Narayanaraju, S.J., Guan, X., Wang, X.E.: Grit: Teaching mllms to think with images (2025), `https://arxiv.org/abs/2505.15879`, accessed: 2026-07-01
29. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. arXiv preprint arXiv:2212.05032 (2022)
30. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
31. Fu, S., Zhou, J., Chen, Q., Jing, H., Nguyen, H.A., Liu, X., Zeng, Z., Ma, L., Zhang, Q., Wu, Q.: Counting hallucinations in diffusion models. arXiv preprint arXiv:2510.13080 (2025)

32. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024)
33. Gemini Team Google: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
34. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
35. Ghosh, D., Zhang, Y., Schmidt, L.: Understanding the fine-grained knowledge capabilities of vision-language models. arXiv preprint arXiv:2602.17871 (2026)
36. Gokhale, T., Palangi, H., Nushi, B., Vineet, V., Horvitz, E., Kamar, E., Baral, C., Yang, Y.: Benchmarking spatial relationships in text-to-image generation. arXiv preprint arXiv:2212.10015 (2022)
37. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
38. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
39. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
40. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6700–6709 (2019)
41. Jung, W., Go, J., Jeon, M., Yoon, S., Kim, J.: Visual funnel: Resolving contextual blindness in multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8962–8971 (2026)
42. Khayatkhoei, M., Chhikara, P., Ilievski, F., et al.: Mllms know where to look: Training-free perception of small visual details with multimodal llms. In: International Conference on Learning Representations. vol. 2025, pp. 68194–68213 (2025)
43. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
44. Kou, S., Jin, J., Liu, Z., Liu, C., Ma, Y., Jia, J., Chen, Q., Jiang, P., Deng, Z.: Orthus: Autoregressive interleaved image-text generation with modality-specific heads. In: International Conference on Machine Learning. pp. 31619–31633. PMLR (2025)
45. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
46. Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023)
47. Lee, K., Kwak, S., Sohn, K., Shin, J.: Direct consistency optimization for robust customization of text-to-image diffusion models. Advances in neural information processing systems **37**, 103269–103304 (2024)
48. Li, S., Kallidromitis, K., Gokul, A., Kato, Y., Kozuka, K.: Aligning diffusion models by optimizing human utility. Advances in Neural Information Processing Systems **37**, 24897–24925 (2024)

49. Li, T., Lu, Q., Zhao, L., Li, H., Zhu, X., Qiao, Y., Zhang, J., Shao, W.: Unifork: Exploring modality alignment for unified multimodal understanding and generation. arXiv preprint arXiv:2506.17202 (2025)
50. Li, Y., Wan, P., Han, L., Wang, Y., Nie, L., Zhang, M.: Countdiffusion: Text-to-image synthesis with training-free counting-guidance diffusion. arXiv preprint arXiv:2505.04347 (2025)
51. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
52. Li, Z., Li, H., Shi, Y., Farimani, A.B., Kluger, Y., Yang, L., Wang, P.: Dual diffusion for unified image generation and understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2779–2790 (2025)
53. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
54. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
55. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European conference on computer vision. pp. 423–439. Springer (2022)
56. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
57. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
58. Lou, A., Meng, C., Ermon, S.: Discrete diffusion modeling by estimating the ratios of the data distribution. arXiv preprint arXiv:2310.16834 (2023)
59. Lv, H., Xiao, J., Li, L.: Pick-and-draw: Training-free semantic guidance for text-to-image personalization. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10535–10543 (2024)
60. Marsili, D., Mehta, A., Lin, R.Y., Gkioxari, G.: Same or not? enhancing visual perception in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17303–17315 (2026)
61. NVIDIA: Llama-nemotron-vlm-dataset-v1. `https://huggingface.co/datasets/nvidia/Llama-Nemotron-VLM-Dataset-v1` (2025), accessed: 2026-07-01
62. OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
63. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., et al.: Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256 (2025)
64. Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7932–7942 (2024)
65. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations. vol. 2024, pp. 1862–1874 (2024)
66. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)

67. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
68. Sahoo, S.S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J.T., Rush, A., Kuleshov, V.: Simple and effective masked diffusion language models. Advances in Neural Information Processing Systems **37**, 130136–130184 (2024)
69. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
70. Shi, Q., Bai, J., Zhao, Z., Chai, W., Yu, K., Wu, J., Tong, Y., Li, X., Li, X., Yan, S.: Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. arXiv preprint arXiv:2505.23606 (2025)
71. Singh, A., Pang, G., Toh, M., Huang, J., Galuba, W., Hassner, T.: Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8802–8812 (2021)
72. Song, C.H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., Birchfield, S.: Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15768–15780 (2025)
73. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
74. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019)
75. Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., Wang, X.: Emu: Generative pretraining in multimodality. In: International conference on learning representations. vol. 2024, pp. 12352–12380 (2024)
76. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
77. Tong, P., Brown, E., Wu, P., Woo, S., Iyer, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)
78. Tong, S., Fan, D., Li, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17001–17012 (2025)
79. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
80. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
81. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
82. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140 (2016)

83. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
84. Wang, X., Zhang, Z., Zhang, H., Lin, Z., Zhou, Y., Liu, Q., Zhang, S., Li, Y., Liu, S., Zheng, H., et al.: Hbridge: H-shape bridging of heterogeneous experts for unified multimodal understanding and generation. arXiv preprint arXiv:2511.20520 (2025)
85. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
86. Wang, Z., Hu, X., Wang, Y., Xiong, F., Zhang, M., Chu, X.: Everything in its place: Benchmarking spatial intelligence of text-to-image models. In: The Fourteenth International Conference on Learning Representations (2026), `https://openreview.net/forum?id=ddFN3lWpIr`
87. Wang, Z., Chen, Z., Gou, C., Li, F., Deng, C., Zhu, D., Li, K., Yu, W., Tu, H., Fan, H., et al.: Lightfusion: A light-weighted, double fusion framework for unified multimodal understanding and generation. arXiv preprint arXiv:2510.22946 (2025)
88. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12966–12977 (2025)
89. Wu, J., Jiang, Y., Ma, C., Liu, Y., Zhao, H., Yuan, Z., Bai, S., Bai, X.: Liquid: Language models are scalable and unified multi-modal generators. International Journal of Computer Vision **134**(1), 39 (2026)
90. Wu, S., Zhang, W., Xu, L., Jin, S., Wu, Z., Tao, Q., Liu, W., Li, W., Loy, C.C.: Harmonizing visual representations for unified multimodal understanding and generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17739–17750 (2025)
91. Xiao, J., Lv, H., Li, L., Wang, S., Huang, Q.: R&b: Region and boundary aware zero-shot grounded text-to-image generation. In: International Conference on Learning Representations. vol. 2024, pp. 26088–26116 (2024)
92. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: International Conference on Learning Representations. vol. 2025, pp. 28240–28264 (2025)
93. Xin, Y., Qin, Q., Luo, S., Zhu, K., Yan, J., Tai, Y., Lei, J., Cao, Y., Wang, K., Wang, Y., et al.: Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. arXiv preprint arXiv:2510.06308 (2025)
94. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al.: Qwen3-omni technical report. arXiv preprint arXiv:2509.17765 (2025)
95. Xu, Z., Chen, J., Lin, Z., Pan, X., Huang, L., Zhou, T., Khabsa, M., Wang, Q., Jin, D., Yasunaga, M., et al.: Pisces: An auto-regressive foundation model for image understanding and generation. arXiv preprint arXiv:2506.10395 (2025)
96. Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M.S., Kendre, S., Zhang, J., Qin, C., Zhang, S., Chen, C.C., Yu, N., Tan, J., Awalgaonkar, T.M., Heinecke, S., Wang, H., Choi,

Y., Schmidt, L., Chen, Z., Savarese, S., Niebles, J.C., Xiong, C., Xu, R.: xgen-mm(blip-3): A family of open large multimodal models. arXiv preprint (August 2024)

97. Yang, J., Yin, D., Zhou, Y., Rao, F., Zhai, W., Cao, Y., Zha, Z.J.: Mmar: Towards lossless multi-modal auto-regressive probabilistic modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7974–7985 (2025)
98. Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Shen, W., Zhu, X., Li, X.: Using human feedback to fine-tune diffusion models without any reward model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8941–8951 (2024)
99. Yang, K., Russakovsky, O., Deng, J.: Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In: International Conference on Computer Vision (ICCV) (2019)
100. Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: Mmada: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809 (2025)
101. Yoon, H., Jung, J., Kim, J., Choi, H., Shin, H., Lim, S., An, H., Kim, C., Han, J., Kim, D., et al.: Visual representation alignment for multimodal large language models. arXiv preprint arXiv:2509.07979 (2025)
102. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024)
103. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
104. Zhang, J., Hu, J., Khayatkhoei, M., Ilievski, F., Sun, M.: Exploring perceptual limitation of multimodal large language models. arXiv preprint arXiv:2402.07384 (2024)
105. Zhang, J., Khayatkhoei, M., Chhikara, P., Ilievski, F.: Visual cropping improves zero-shot question answering of multimodal large language models. In: R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models (2023)
106. Zhang, J., Li, T., Li, L., Yang, Z., Cheng, Y.: Are unified vision-language models necessary: Generalization across understanding and generation. arXiv preprint arXiv:2505.23043 (2025)
107. Zhang, P., Li, C., Qiao, L., Cheng, Z., Pu, S., Niu, Y., Wu, F.: Vsr: a unified framework for document layout analysis combining vision, semantics and relations. In: International conference on document analysis and recognition. pp. 115–130. Springer (2021)
108. Zhang, Y., Gueguen, L., Zharkov, I., Zhang, P., Seifert, K., Kadlec, B.: Uber-text: A large-scale dataset for optical character recognition from street-level imagery. In: SUNw: Scene Understanding Workshop - CVPR 2017. Hawaii, U.S.A. (2017)
109. Zhao, S., Zhang, X., Guo, J., Hu, J., Duan, L., Fu, M., Chng, Y.X., Wang, G.H., Chen, Q.G., Xu, Z., et al.: Unified multimodal understanding and generation models: Advances, challenges, and opportunities. arXiv preprint arXiv:2505.02567 (2025)
110. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)