GESTALT: LARGE MULTIMODAL INTERPLAY MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose **Gestalt**, a new paradigm of large multimodal model built around multimodal interplay. Despite rapid advances, large multimodal models are reaching a bottleneck: existing approaches focus primarily on accommodating additional modalities while overlooking the distinct characteristics of each modality and the relations among them. Motivated by the multistage property of human multisensory perception, we propose a multimodal interplay pyramid that organizes multimodal modeling as a progression from modality-specific processing, through cross-modal alignment, to deeper multimodal integration. Guided by this pyramid, Gestalt adopts a unified discrete diffusion framework and an interplay-partitioned architecture, with learnable interplay tokens mediating cross-modal exchange and integration. The pyramid also structures its data organization and progressive training strategy. Strong performance across image generation, multimodal understanding, and text-only evaluation shows that Gestalt significantly improves cross-modal integration while preserving modality-specific information, effectively harnessing the strengths of diffusion-based multimodal models and offering a promising path toward unified multimodal intelligence. Code, data, and model will be publicly released.

1 INTRODUCTION

Modern large multimodal models have evolved from architectural compatibility to data compatibility. Early systems, such as LLaVA (Liu et al., 2023) and Qwen2.5-VL (Bai et al., 2025), achieve architectural compatibility by connecting modality-specific encoders to a pretrained language model through learned projectors. More recent native models, such as Kimi K3 (Kimi Team et al., 2026) and Qwen 3.8, pursue data compatibility by jointly pretraining on large-scale mixed-modal corpora within a shared model. Yet progress along this path is reaching a bottleneck. We argue that multimodal intelligence requires more than accommodating multiple modalities: *more is different*, and merely bringing modalities together cannot adequately address the emergent properties and distinctive challenges inherent to multimodality.

In multimodal settings, information may be available in one modality, shared across modalities, or emerge only from their combination. Such information relationships define multimodal interplay, a feature unique to multimodal learning. Human multisensory perception offers a useful analogy. Sensory information is first encoded within individual pathways and then progressively integrated across multiple stages into a coherent representation (Stein & Meredith, 1993; Gazzaniga et al., 2011; Tang et al., 2016). Inspired by this multistage organization, we propose a multimodal interplay pyramid that arranges multimodal learning into successive levels (Figure 1a), with each level building on the preceding one to enable increasingly complex interplay.

The pyramid handles these information relationships in a bottom-up progression. The *lower stage* preserves unique information available from only one modality. Building on this foundation, the *intermediate stage* aligns redundant information shared across modalities. At the top, the *upper stage* integrates information across modalities to derive synergistic information unavailable from either modality alone. Figure 1b illustrates this progression within a single image-text pair. The text contributes the size–collar relation, whereas the image supplies the visible collar color. Their shared description of a Golden Retriever on grass establishes cross-modal correspondence, and combining the two modality-specific cues identifies the depicted dog as the smaller one.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **(a)** The multimodal interplay pyramid progresses from modality-specific modeling to cross-modal alignment and synergy. **(b)** Combining the textual size-collar relation with the visible collar identifies the depicted dog as the smaller one. **(c)** Gestalt supports multimodal understanding, image generation, and language tasks within a unified model.

Based on this principle, we propose **Gestalt**[1], a large multimodal model that explicitly accounts for different forms of information interplay across modalities. Gestalt implements this idea through its architecture, data organization, and training strategy. Architecturally, Gestalt adopts a discrete diffusion framework that represents images and text in a unified token space and trains them with the same objective, enabling both multimodal understanding and generation. To implement the interplay pyramid, the Transformer is divided into two successive zones, with learnable interplay tokens mediating cross-modal information flow throughout. In *Interplay Zone I*, image and text tokens are processed separately by modality-specific feed-forward networks, and direct attention between them is restricted; cross-modal exchange is instead mediated by the interplay tokens. In *Interplay Zone II*, this restriction is removed, allowing image, text, and interplay tokens to interact through full multimodal self-attention and shared feed-forward processing. The two zones thus instantiate the pyramid as a progression from modality-specific processing, through controlled exchange, to full multimodal integration.

Beyond architecture, the interplay pyramid also guides data organization and the training strategy. The modality-unique foundation of its lower stage is already established by the pretrained backbone. Multimodal training therefore begins with cross-modal alignment: Gestalt conducts joint multimodal pretraining and conditional continual pretraining on 78M aligned image-caption pairs to learn a joint vision-language distribution and bidirectional cross-modal dependencies. It then performs interplay-oriented supervised fine-tuning on 13.7M instruction examples organized into redundant, text-unique, visual-unique, and synergistic categories. Across three curriculum phases, all four categories are retained while the sampling emphasis progressively shifts from consolidating cross-modal alignment and modality-specific capability, and finally to deeper multimodal synergy.

A unified multimodal model is expected to do more than support generation and understanding. It should also build coherent cross-modal representations. This calls for a broad set of evaluations. On benchmarks that emphasize complex, fine-grained generation, Gestalt outperforms both generation-only and unified baselines, reaching 73.30 on UniGenBench and 74.57/78.86 on the short/long settings of TIIF-Bench. Gestalt also demonstrates strong multimodal understanding, particularly on vision-centric benchmarks, scoring 78.9 and 86.1 on the two CV-Bench tracks and outperforming the representative autoregressive-based model BAGEL on both. Representation analyses further show that Gestalt forms an integrated vision-language space while largely preserving pretrained linguistic geometry. Despite multimodal adaptation, it maintains strong language capability and delivers leading overall text-only performance among the evaluated diffusion-based unified models. Finally, results on visual-specific and synergy-oriented benchmarks, together with the organization of interplay-token representations, suggest that Gestalt adapts information exchange to different interplay demands. As illustrated in Figure 1c, Gestalt establishes leading performance among the evaluated diffusion-based unified models, and it effectively harnesses the potential of diffusion-based multimodal models, offering a promising path toward unified multimodal intelligence.

[1]The name is inspired by Gestalt psychology (Köhler, 1967): the whole is greater than the sum of its parts.

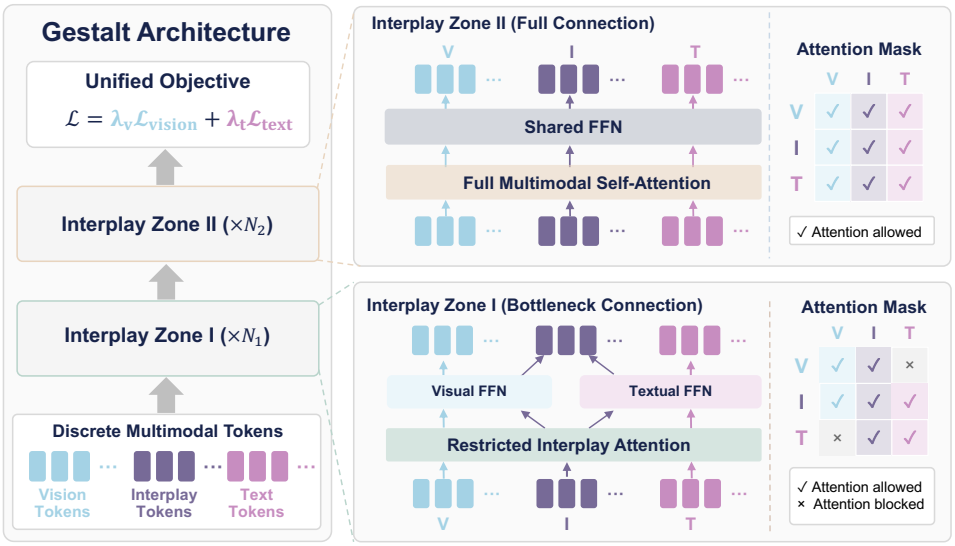

Figure 2: **Overview of the Gestalt architecture.** Gestalt processes discrete vision (orange), interplay (green), and text (blue) tokens through Interplay Zone I (N_1 bottleneck interplay layers) and Interplay Zone II (N_2 full multimodal interplay layers) under a unified objective. Zone I uses restricted interplay attention, with cross-modal exchange mediated by interplay tokens, together with modality-specific FFNs; Zone II uses full multimodal self-attention and a shared FFN.

2 METHODOLOGY

Guided by the interplay pyramid, Gestalt organizes multimodal learning as a progression from modality-specific processing to cross-modal alignment and multimodal synergy. Within a unified discrete diffusion framework, this principle shapes its interplay-centric architecture, interplay-oriented data organization, and progressive training strategy.

2.1 UNIFIED INTERPLAY-CENTRIC ARCHITECTURE

As illustrated in Figure 2, Gestalt represents vision and text within a shared token space, enabling multimodal understanding and generation in a unified framework. To explicitly organize information exchange within this shared space, we introduce learnable interplay tokens as an interface between the two modalities. The architecture further organizes cross-modal interplay into two hierarchical zones, progressively transitioning from bottleneck interplay to full multimodal interplay. Together, these components realize a preserve-then-integrate pathway for unified multimodal modeling.

Discrete Multimodal Representation. To support multimodal interplay within a single model, Gestalt maps visual and textual inputs to a unified codebook and embedding space. This fully discrete formulation allows both modalities to be optimized under the same objective. The underlying masked diffusion language model uses bidirectional attention, allowing visual and textual tokens to condition on each other without the fixed direction imposed by autoregressive models. A fixed VQ-VAE encoder (Shi et al., 2025) converts each image into discrete tokens using a codebook of 16,384 entries. These visual tokens are then combined with textual tokens in a shared sequence.

Interplay Tokens. A shared discrete space enables joint modeling but does not explicitly organize cross-modal information exchange. We therefore introduce dedicated learnable interplay tokens as a shared interface between the visual and textual streams, whose function is progressively expanded from mediated exchange in Zone I to full multimodal integration in Zone II.

Zone I: Bottleneck Interplay. Unique information depends on the distinct characteristics of each modality, yet these characteristics can be diluted by fully shared processing. Zone I therefore adopts a bottleneck interplay design that combines modality-specific modeling with controlled cross-modal exchange. Visual and textual tokens pass through separate feed-forward networks, while restricted attention prevents direct interaction between them. Instead, cross-modal information is exchanged through the interplay tokens. Through this mediated exchange, Zone I preserves modality-specific representations while gradually establishing cross-modal correspondence.

Zone II: Full Multimodal Interplay. After modality-specific representations and basic cross-modal correspondence have been established, Zone II replaces the bottleneck with full multimodal self-attention and a shared feed-forward network. Visual, textual, and interplay tokens can now attend to one another directly, enabling information exchange across the entire sequence. The interplay tokens continue to aggregate task-relevant information from both modalities. Together, full multimodal attention and shared processing support the modeling of synergistic information.

2.2 Training Data Organization

Gestalt organizes training data progressively across pretraining, continual pretraining (CPT), and supervised fine-tuning (SFT). For both pretraining and CPT, we use 78M aligned image–caption pairs to learn basic cross-modal correspondence and acquire multimodal knowledge. For SFT, we organize instruction data according to the interplay pyramid. Each example is categorized as redundant, text-unique, visual-unique, or synergistic. The resulting corpus contains 13.7M examples, including 2.8M redundant, 4.2M text-unique, 4.0M visual-unique, and 2.7M synergistic examples. This organization enables controlled learning of different forms of multimodal interplay during fine-tuning. A detailed description of our data organization strategy is provided in Appendix A.

2.3 Progressive Training Strategy

Training follows the interplay pyramid through joint multimodal pretraining, conditional continual pretraining, and interplay-oriented supervised fine-tuning under a unified masked-diffusion framework. Because Gestalt is initialized from a pretrained backbone, its modality-specific foundation is already established before multimodal training. The multimodal stages therefore begin by aligning cross-modal redundant information, then strengthen the modeling of visual-unique information, and finally advance to cross-modal synergy.

Unified Masked Diffusion Objective. With vision and text represented as discrete tokens, Gestalt jointly models both modalities with a unified masked-diffusion objective. For a multimodal sequence $\mathbf{x} = [\mathbf{x}^v, \mathbf{x}^t]$, we randomly mask a set of positions and reconstruct the corresponding tokens:

$$\mathcal{L}_{\text{mask}} = -\mathbb{E}_m \sum_{i \in m} \log p_\theta(x_i \mid \mathbf{x}_{\setminus m}), \tag{1}$$

where m denotes the set of masked positions. Unlike autoregressive modeling, this enables bidirectional conditioning across modalities without a fixed generation order.

Progressive Multimodal Pretraining. Multimodal pretraining begins with joint masked prediction. Visual and textual tokens are masked independently and reconstructed jointly:

$$p(\mathbf{x}^v_{m_v}, \mathbf{x}^t_{m_t} \mid \mathbf{x}^v_{\setminus m_v}, \mathbf{x}^t_{\setminus m_t}). \tag{2}$$

This objective encourages the model to capture both intra-modal structure and cross-modal dependencies in either direction. This stage establishes a basic joint distribution over the two modalities.

Continual pretraining then adopts conditional masked prediction: one modality remains visible while the other is masked and predicted. Learning bidirectional vision-to-text and text-to-vision dependencies strengthens cross-modal conditional modeling and provides a natural bridge between joint multimodal modeling and downstream generation conditioned on either modality.

Interplay-Oriented Supervised Fine-tuning. Building on the cross-modal dependencies learned during pretraining and CPT, supervised fine-tuning focuses on teaching the model how to exploit different types of multimodal information for downstream tasks. To this end, we adopt a three-phase curriculum under conditional masked prediction. All four interplay categories are retained throughout training, while their sampling ratios are progressively adjusted to place increasing emphasis on visual-unique and synergistic information, as illustrated in Figure 3. This three-phase curriculum gradually shifts the learning focus from consolidating cross-modal alignment and language capability, to strengthening visual-specific capability, and finally to promoting deeper multimodal integration.

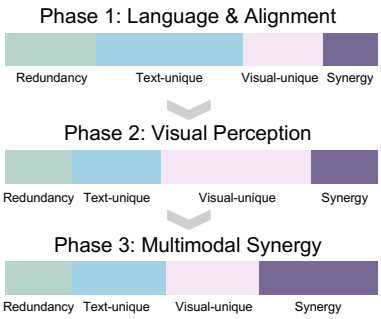

Figure 3: Training recipe for interplay-oriented data organization during SFT.

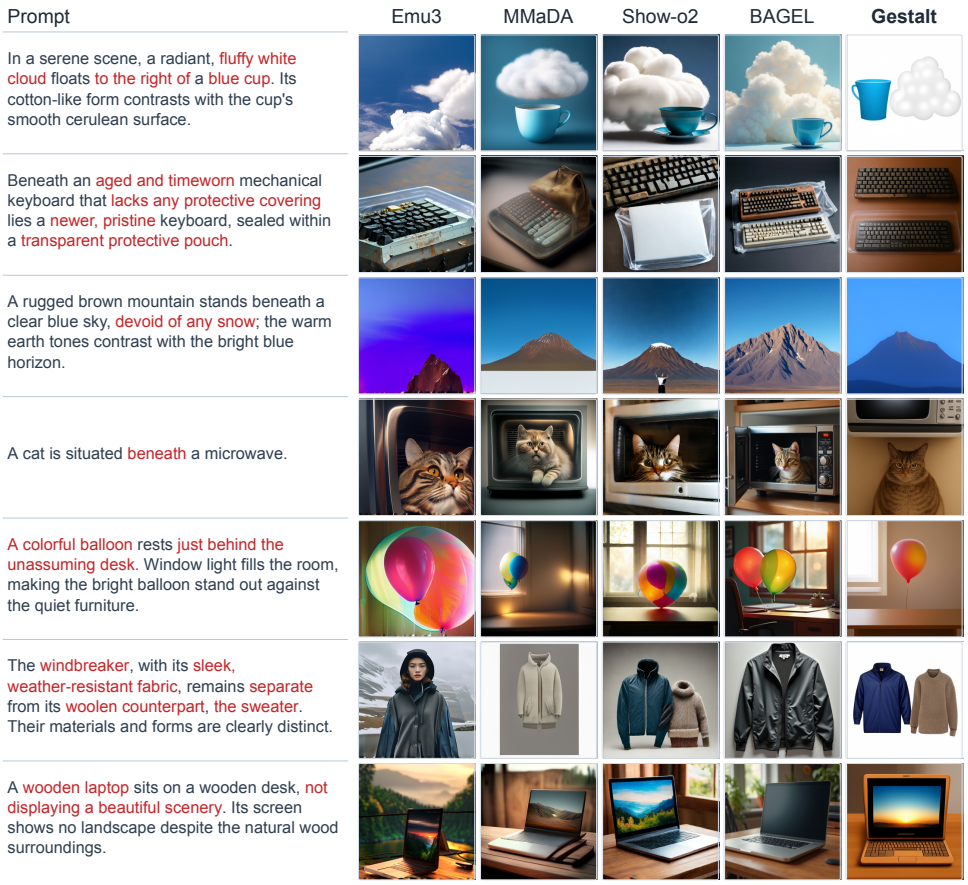

Figure 4: **Qualitative comparison of fine-grained text-to-image generation.** Red text highlights key semantic constraints in the prompts, such as spatial relations, attributes, negation, and properties.

3 Experiment

3.1 Training and Evaluation Overview

Training and Model Configuration. Gestalt is initialized from LLaDA-8B-Instruct (Nie et al., 2026). Images are encoded into discrete tokens using IBQ (Shi et al., 2025), with its visual codebook of 16,384 added to the language vocabulary, while interleaved MRoPE preserves their two-dimensional spatial positions. We use 32 interplay tokens and designate the first four Transformer layers as Interplay Zone I. The newly introduced visual FFN in each of these layers is initialized from the corresponding text FFN in LLaDA. Training proceeds in three stages: joint pretraining on 70M image-caption pairs from LLaVA-OneVision-1.5 (An et al., 2025), conditional continual pretraining on 8M image-text pairs, and supervised fine-tuning on 13.7M interplay-organized instruction examples from LLaVA-OneVision-1.5, MAmmoTH-VL (Guo et al., 2025), FineVision (Wiedmann et al., 2025), and InternVL (Wang et al., 2025a). The final stage additionally includes 2.5M text-to-image examples collected from text-to-image-2M (Zou et al., 2026), ShareGPT-4o-Image (Chen et al., 2025b), and BLIP3o-60k (Chen et al., 2025a). All the parameters are trainable at all stages.

Evaluation Overview. Our evaluation proceeds from what Gestalt achieves to how its multimodal representations and interactions are organized. We begin by establishing its unified capabilities across fine-grained text-to-image generation and multimodal understanding. We then look beneath these outcomes to characterize the joint vision-language space and the retention of pretrained linguistic structure and text-only capability. Finally, we focus on cross-modal interplay, examining performance under modality-unique and synergistic information settings together with the information encoded by the interplay tokens. This progression links benchmark performance with representation geometry and internal information exchange, yielding a unified empirical account of Gestalt. Further analyses in Appendix B examine interplay-oriented curriculum learning, adaptation to remote-sensing tasks, and qualitative results on fine-grained image generation.

Table 1: Evaluations on UniGenBench English Long for fine-grained text-to-image generation. The best and the second-best results are highlighted in **bold** and underline, respectively.

Model	Overall	Style	World	Attr.	Action	Rel.	Comp.	Grammar	Layout	Logic	Text
Gen. Only											
DALL-E-3	70.82	95.08	**92.71**	84.98	68.36	77.90	73.88	68.19	71.76	57.11	18.26
SD-3.5-Large	64.35	88.12	88.15	78.78	59.63	67.62	62.21	65.23	71.19	44.90	17.66
OmniGen2	71.39	94.35	84.83	83.03	66.57	73.06	70.49	76.40	80.63	56.55	**27.99**
Unified											
Emu3	50.95	89.36	76.16	66.81	43.80	51.70	46.00	50.25	56.67	27.43	1.36
Show-o2	70.33	93.11	88.44	**86.35**	69.02	77.37	76.45	70.30	80.63	59.71	1.90
Janus-Pro	71.11	94.02	88.15	81.81	69.14	77.96	76.53	74.62	82.14	62.62	4.08
MMaDA	40.10	75.83	52.75	49.90	32.42	39.06	38.37	50.00	43.02	19.42	0.27
BAGEL	71.26	92.44	89.31	84.21	67.62	75.70	74.71	74.75	81.90	59.71	12.23
Lumina-DiMOO	71.81	86.88	88.58	83.71	**69.66**	73.33	74.93	74.49	84.84	58.01	23.64
Gestalt	**73.30**↑1.49	**95.85**↑0.77	81.65	82.51	68.57	**78.12**↑0.16	**81.20**↑4.67	**82.23**↑5.83	**85.79**↑0.95	**73.28**↑10.66	3.80

Table 2: Evaluations on TIIF-Bench for text-to-image instruction following under short and long prompts. The best and the second-best results are highlighted in **bold** and underline, respectively.

Model	Overall		Basic		Advanced		Designer	
	Short	Long	Short	Long	Short	Long	Short	Long
Gen. Only								
PixArt-Sigma	62.00	58.12	70.66	75.25	57.65	49.50	62.11	52.41
FLUX.1 Pro	67.32	69.89	79.08	78.91	61.10	65.37	71.80	68.80
MidJourney V7	68.74	65.69	77.41	76.00	64.66	60.53	68.83	63.61
SD 3.5 Large	71.15	66.96	78.34	79.56	67.67	61.18	64.43	66.39
Unified								
Emu3	43.38	39.44	49.88	42.08	37.09	33.52	53.73	60.45
MMaDA	52.32	52.90	65.78	66.71	50.32	50.76	60.45	55.60
Show-o2	67.38	68.76	81.16	84.06	69.25	72.99	75.37	75.75
Janus-Pro	66.50	65.02	79.33	78.25	59.71	58.82	65.84	60.25
BAGEL	71.50	71.70	81.79	80.05	70.24	72.19	68.28	67.91
Lumina-DiMOO	71.27	68.53	75.50	78.29	70.49	68.33	69.78	70.90
Gestalt	**74.57**↑3.07	**78.86**↑7.16	**83.79**↑2.00	**86.06**↑2.00	**72.60**↑2.11	**77.69**↑4.70	**84.70**↑9.33	**88.81**↑13.06

3.2 Unified Multimodal Capabilities

3.2.1 Fine-Grained Text-to-Image Generation

Gestalt demonstrates strong fine-grained text-to-image generation under complex semantic constraints, including multiple objects, attributes, spatial relations, and compositional requirements, as shown in Figure 4. We compare Gestalt with two groups of baselines: generation-only models specialized for image synthesis, such as DALL-E 3 (Betker et al., 2023) and FLUX.1 Pro (Labs, 2024), and unified multimodal models that support both generation and understanding, such as BAGEL (Deng et al., 2025) and Janus-Pro (Chen et al., 2025c).

Fine-grained semantic alignment. We evaluate fine-grained text-to-image alignment on UniGenBench (Wang et al., 2025b), which covers attributes, actions, spatial relations, composition, and logical reasoning. As shown in Table 1, Gestalt achieves the best overall score of 73.30, surpassing the highest-scoring baselines in the generation-only and unified groups, OmniGen2 (71.39) and Lumina-DiMOO (71.81), respectively. Gestalt further ranks first in style, relation, composition, grammar, layout, and logic. These dimensions span local semantic details and global compositional structure, requiring the model to preserve relations among concepts and jointly realize multiple constraints. Together, the results reflect precise alignment between textual semantics and visual structure beyond coarse concept matching.

Complex instruction following. We further evaluate complex text-to-image instruction following on TIIF-Bench (Wei et al., 2025). As Table 2, Gestalt achieves the best overall scores under both short and long instructions, reaching 74.57 and 78.86, respectively, ahead of all generation-only and unified baselines. It also ranks first across all three evaluation categories in both settings. Its advantage grows under longer instructions, which introduce more entities and interdependent constraints, demonstrating stronger retention and joint realization of detailed semantic requirements. Together, UniGenBench and TIIF-Bench show that Gestalt unifies understanding and generation without sacrificing fine-grained text-to-image performance, outperforming even specialized generation models.

Table 3: Evaluations on multimodal understanding benchmarks. The symbol † denotes the results from the official paper, while the rest of the results are evaluated using the official checkpoint and inference scripts. The **best** and the second-best results in diffusion-based models are highlighted.

Models	General					Vision-Centric		
	MME-P	GQA	MMStar-P	POPE	RWQA	MMVP	CVB^{2d}	CVB^{3d}
AR-Based								
BAGEL	1687.0†	66.4	70.9	88.2	67.6	69.3†	77.7	84.2
Diffusion-based								
MMaDA	1410.7†	**61.3**†	43.0	86.1†	48.2	17.3	55.3	54.8
Lumina-DiMOO	1534.2†	43.3	-	**87.4**†	35.9	34.0	54.3	52.0
LaViDA-o	1431.0	54.1	55.9	–	56.6	47.3	73.4	70.8
LLaDA-o	1412.0†	58.0	55.6	87.2	**66.4**	46.7	78.0	75.9
Omni-diffusion	1216.7†	–	–	76.6†	–	–	–	–
Gestalt	**1600.2**↑66.0	60.2	**60.8**	86.5	59.1	**48.7**↑1.4	**78.9**↑0.9	**86.1**↑10.2



Figure 5: Visualization of image-text hidden representation across diffusion and AR architecture.

3.2.2 Multimodal Understanding

Gestalt is evaluated on eight multimodal understanding benchmarks, grouped into (1) general VQA benchmarks: MME-P (Fu et al., 2026), GQA (Hudson & Manning, 2019), MMStar-Perception (Chen et al., 2024), POPE (Li et al., 2023), and RealworldQA (xAI, 2024); and (2) vision-centric benchmarks: MMVP (Tong et al., 2024b) and CV-Bench (Tong et al., 2024a). We compare Gestalt with diffusion-based unified multimodal models, including MMaDA (Yang et al., 2026), Lumina-DiMOO (Xin et al., 2025), Lavida-o (Li et al., 2026b), LLaDA-o (You et al., 2026), and Omni-diffusion (Li et al., 2026a), as well as BAGEL (Deng et al., 2025), a representative autoregressive unified model. As shown in Table 3, Gestalt achieves the best results among diffusion-based models on five benchmarks and ranks among the top two on seven of the eight benchmarks. This performance establishes Gestalt as a strong diffusion-based unified model across both general multimodal understanding and vision-centric perception. Gestalt shows particularly strong performance on vision-centric understanding. It achieves the best results among diffusion-based models on MMVP and both CV-Bench tracks. Notably, Gestalt also surpasses BAGEL, a representative autoregressive unified model, on both CV-Bench tracks: Gestalt scores 78.9 and 86.1, whereas the AR-based BAGEL obtains 77.7 and 84.2, respectively. These gains on vision-centric benchmarks align with Gestalt's explicit emphasis on modality-unique information and highlight the potential of interplay-oriented modeling to strengthen visual understanding within a unified diffusion model.

3.3 Cross-modal Representation Analysis

Beyond benchmark performance, we examine the representations underlying Gestalt's multimodal capabilities, including the organization of visual and textual representations and the preservation of pretrained language representations and text-only capability.

Unified Vision-Language Representation Space. We examine how visual and textual token representations are organized in the representation space at layer 24 using image–caption pairs from the DOCCI dataset (Onoe et al., 2024). We compare Gestalt with Lumina-DiMOO (Xin et al., 2025), a diffusion-based unified model, and Qwen3.8-9B-Distill[2], an autoregressive native multimodal model, pretrained from scratch with early-fused multimodal data. As shown in Figure 5, Lumina-DiMOO exhibits a clear separation between the two modalities, while Qwen3.8 shows par-

[2]As no official 9B checkpoint of Qwen3.8 is publicly available, we use the distilled variant Qwen3.8-9B-Distill: `https://huggingface.co/empero-ai/Qwen3.8-9B-Distill`.

Table 4: Comparison of distribution shift (Norm and Distance) and representation preservation (Cosine Similarity and CKA) of LLaDA-based multimodal models in the language embedding space.

Type	Model	Norm Mean	Norm Std	Mean Distance	Cosine Sim.	CKA
Base	LLaDA	7.82	1.11	–	–	–
Projector Alignment	LLaDA-V	7.82	1.11	0.043	0.99998	0.99997
	LaViDa	7.83	1.11	0.160	0.99976	0.99964
	LLaDA-O	8.02	1.16	1.520	0.98136	0.97424
Unified Token Space	Lumina-DiMOO	0.85	0.49	7.426	0.58795	0.43894
	Gestalt	5.29	0.76	2.590	0.99677	0.99494

Table 5: Evaluation of language capabilities based on six text-only benchmarks. The **best** and the second-best results among diffusion-based unified models are highlighted.

Model	MMLU	TruthfulQA	WinoGrande	HellaSwag	ARC-E	ARC-C
AR-Based						
Show-o2	71.70	46.94	74.03	76.67	84.43	58.53
Janus-Pro	49.90	41.72	67.17	68.41	65.74	40.70
BAGEL	28.02	40.51	50.75	28.59	27.53	23.63
Diffusion-based						
MMaDA	40.14	43.81	54.85	45.81	46.72	28.67
Lumina-DiMOO	29.75	43.34	51.62	39.11	44.44	26.45
LLaDA-o	25.25	**52.30**	51.22	30.72	36.66	24.66
Gestalt	**49.50**↑9.36	47.55	**60.62**↑5.77	**53.44**↑7.63	**62.29**↑15.57	**43.69**↑15.02

tial cross-modal overlap, with a substantial portion of textual tokens remaining separated from the visual representations. In contrast, Gestalt exhibits a more interleaved distribution of visual and textual tokens, suggesting a more integrated vision-language representation space than both diffusion-based and autoregressive baselines.

Representation Preservation under Multimodal Adaptation. We compare how two multimodal adaptation paradigms reshape the language representation space. Projector-alignment models align visual features to the language space while largely preserving textual representations, whereas unified token-space models jointly represent vision and language within a shared discrete token space. We report representation changes using the mean and standard deviation of embedding norms, and measure preservation relative to pretrained LLaDA using embedding distance, cosine similarity, and CKA. As shown in Table 4, projector-alignment models introduce only minor changes to the original language space, while unified models exhibit more substantial distribution shifts. Notably, Lumina-DiMOO reduces cosine similarity and CKA with pretrained LLaDA to 0.5880 and 0.4389, respectively, whereas Gestalt maintains substantially higher values of 0.9968 and 0.9949 despite a noticeable shift in embedding scale. These results suggest that Gestalt build a new space for unified multimodal modeling while largely preserving its original representation structure.

Language Capability. We next assess representation-level preservation through text-only capability. We compare Gestalt with diffusion-based unified models MMaDA, Lumina-DiMOO, and LLaDA-o, and AR-based unified models Show-o2, Janus-Pro, and BAGEL (Yang et al., 2026; Xin et al., 2025; You et al., 2026; Xie et al., 2026; Chen et al., 2025c; Deng et al., 2025). Evaluation is 5-shot on MMLU, WinoGrande, and TruthfulQA and 0-shot on HellaSwag, ARC-Easy, and ARC-Challenge, with TruthfulQA reported using MC2. As shown in Table 5, Gestalt leads diffusion-based unified models on five of six benchmarks, with particularly large gains on ARC-E and ARC-C. Notably, it is also competitive with AR-based unified models, outperforming BAGEL across all six benchmarks, closely matching Janus-Pro on MMLU, and surpassing it on TruthfulQA and ARC-C. These results confirm that Gestalt retains strong language capability after multimodal adaptation.

3.4 Understanding the Role of Cross-Modal Interplay

Multimodal interplay is central to Gestalt, preserving modality-unique information while deriving synergy through cross-modal integration. This section evaluates these capabilities and examines how interplay tokens encode their distinct information demands.

Evaluation of Multimodal Interplay. We evaluate two representative forms of multimodal interplay. MIBench-Visual subset (MIB-V) (Miao et al., 2026) and CoreCognition Sensorimotor subset (CoreCog-SM) (Li et al., 2024) are designed for evaluating basic visual-specific capabilities, while MM-IMDb (Arevalo et al., 2017) and SRBench (Stogiannidis et al., 2025) focus on cross-modal synergy through multimodal integration and reasoning. As shown in Table 6, Gestalt achieves strong performance on both visual-oriented and synergy-oriented benchmarks, including the best results among diffusion-based models on CoreCognition, MM-IMDb, and SRBench. These results demonstrate that Gestalt can effectively handle different forms of multimodal interplay, maintaining strong visual capability while benefiting from cross-modal synergy.

Table 6: Evaluation of visual-specific and synergistic multimodal interplay.

Model	Visual		Synergy	
	MIB-V	CoreCog-SM	MM-IMDb	SRBench
AR-Based				
BAGEL	65.96	65.00	60.60	51.89
Diffusion-based				
MMaDA	44.23	43.20	30.39	36.72
Lumina-DiMOO	56.43	42.20	31.22	45.50
LLaDA-o	**60.31**	<u>53.30</u>	39.64	<u>50.89</u>
LaViDa-O	58.07	50.60	<u>51.54</u>	40.89
Gestalt	<u>59.89</u>	**55.20**↑1.90	**67.07**↑15.53	**53.17**↑2.28

Interplay-Aware Representations. We further analyze how interplay tokens respond to samples with different forms of multimodal interplay. Specifically, we visualize their hidden representations at layer 16 on MIBench, which categorizes samples into visual uniqueness, textual uniqueness, and cross-modal synergy according to the information required for prediction. As shown in Figure 6, the interplay-token representations exhibit clear structures associated with different interplay types. Samples dominated by visual or textual information occupy largely separated regions, whereas synergy samples form a distinct distribution that partially bridges the two. These patterns suggest that interplay tokens adapt to different multimodal information requirements, reflecting whether prediction depends primarily on visual information, textual information, or information that emerges from their combination.

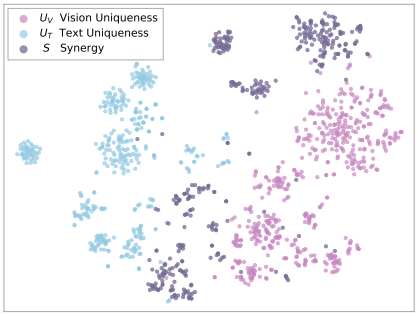

Figure 6: Visualization of interplay token distribution across samples with different multimodal interplay.

4 Related Work

Large multimodal models have evolved from architectural compatibility to data compatibility. Early systems connect modality-specific vision encoders to pretrained AR-based language models through learned projectors (Liu et al., 2023; Bai et al., 2025), whereas native unified multimodal models jointly pretrain visual and textual data in a shared model to support both understanding and generation (Zhao et al., 2025). These unified models use AR next-token prediction (Team, 2024; Wang et al., 2024; Chen et al., 2025c), discrete diffusion (Nie et al., 2026; You et al., 2026; Yu et al., 2025; Li et al., 2026b; Yang et al., 2026; Xin et al., 2025), or hybrid AR-diffusion and flow-based formulations (Zhou et al., 2025; Xie et al., 2025; 2026). Despite this progress, existing work largely focuses on bringing visual and textual data into a shared model, without explicitly organizing the information relationships and emergent properties that arise across modalities.

5 Discussion

Moving beyond simply accommodating additional modalities, Gestalt frames unified multimodal modeling around a multistage interplay pyramid. Our findings indicate that unification need not require sharing all computation from the earliest layer; modality-specific processing and controlled exchange before full integration may better balance specialization and cross-modal synergy. In existing studies of mixed-modal scaling laws, multimodal synergy is treated as a key component (Aghajanyan et al., 2023). This connection suggests that multimodal scaling may depend not only on model and data size, but also on how information exchange is organized across stages. The interplay pyramid provides a structured setting for examining this possibility and may therefore offer a useful basis for future research on scaling laws in large multimodal models.

AI DISCLOSURE STATEMENT

In this work, we used generative AI tools to assist with environment setup, code checking and debugging, improving the clarity and readability of the manuscript, and literature summarization. All AI-assisted outputs were reviewed, verified, and revised by the authors. Experimental designs and conclusions were determined by the authors, code suggestions were inspected and tested before use, and claims concerning related work were checked against the original literature. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the Gestalt architecture, multimodal sequence construction, interplay routing mechanism, and unified masked-diffusion objective in the main paper. The training procedure, including the three-stage training recipe, data organization strategy, optimization settings, and major hyperparameters, is described in the experimental setup. Code, data, and model will be publicly released.

REFERENCES

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pp. 265–279. PMLR, 2023.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. Technical report, OpenAI, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025c.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2026.

Michael S. Gazzaniga, Richard B. Ivry, and George R. Mangun. *Cognitive Neuroscience: The Biology of the Mind*. W. W. Norton, 2011.

Jiawei Guo, Tianyu Zheng, Yizhi Li, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13869–13920, 2025.

Drew A. Hudson and Christopher D. Manning. 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.

Kimi Team et al. Kimi K3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Wolfgang Köhler. Gestalt psychology. *Psychologische forschung*, 31(1):XVIII–XXX, 1967.

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024.

Lijiang Li, Zuwei Long, Yunhang Shen, Heting Gao, Haoyu Cao, Xing Sun, Caifeng Shan, Ran He, and Chaoyou Fu. Omni-diffusion: Unified multimodal understanding and generation with masked discrete diffusion. *arXiv preprint arXiv:2603.06577*, 2026a.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. Lavida: A large diffusion language model for multimodal understanding. *Advances in Neural Information Processing Systems*, 38: 105101–105134, 2026b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 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.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, et al. Core knowledge deficits in multimodal language models. *arXiv preprint arXiv:2410.10855*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. Skysensegpt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024.

Yu Miao, Zequn Yang, Yake Wei, Ziheng Chen, Haotian Ni, Haodong Duan, Kai Chen, and Di Hu. Mibench: Evaluating lmms on multimodal interaction. *arXiv preprint arXiv:2603.13427*, 2026.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, pp. 291–309. Springer, 2024.

Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Scalable image tokenization with index backpropagation quantization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16037–16046. IEEE, 2025.

Barry E Stein and M Alex Meredith. *The merging of the senses*. MIT press, 1993.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*, 2025.

Xiaoyu Tang, Jinglong Wu, and Yong Shen. The interactions of multisensory integration with endogenous and exogenous attention. *Neuroscience & Biobehavioral Reviews*, 61:208–224, 2016.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Yibin Wang, Zhimin Li, Yuhang Zang, Jiazi Bu, Yujie Zhou, Yi Xin, Junjun He, Chunyu Wang, Qinglin Lu, Cheng Jin, et al. Unigenbench++: A unified semantic evaluation benchmark for text-to-image generation. *arXiv preprint arXiv:2510.18701*, 2025b.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. Finevision: Open data is all you need. *arXiv preprint arXiv:2510.17269*, 2025.

xAI. Grok-1.5 vision preview. `https://x.ai/news/grok-1.5v`, April 2024. Accessed: 2026-05-20.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *Advances in Neural Information Processing Systems*, 38:47490–47518, 2026.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *Advances in Neural Information Processing Systems*, 38:138867–138907, 2026.

Zebin You, Shen Nie, Xiaolu Zhang, JUN ZHOU, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10093–10105, 2026.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. Dimple: Discrete diffusion multimodal large language model with parallel decoding. *arXiv preprint arXiv:2505.16990*, 2025.

Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025.

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. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025.

Kai Zou, Zhiwei Zhao, Bin Liu, and Nenghai Yu. Advancing aesthetic image generation via composition transfer. *International Journal of Computer Vision*, 134:252, 2026. doi: 10.1007/s11263-026-02862-8. URL https://doi.org/10.1007/s11263-026-02862-8.

A Data Organization Strategy

A.1 Data curation.

We apply task-specific filtering to construct high-quality training data at different stages. For multimodal pretraining, we select high-quality image–text pairs from LLaVA-OneVision-1.5-Midtraining data An et al. (2025) and use Qwen3-VL-32B-Instruct to verify image–text consistency, removing samples with weak or mismatched correspondence. For supervised fine-tuning, we further perform multiple rounds of quality control. In particular, since the VQ-VAE tokenizer compresses images into discrete visual tokens, fine-grained visual details may be lost during tokenization, making some originally answerable samples no longer recoverable from the tokenized image. We therefore first apply rule-based filtering and then use Qwen3-VL-32B-Instruct to verify that the visual evidence required by the target response remains available after tokenization.

Table 7: Composition of the 13.7M instruction-tuning examples, including multimodal understanding and text-only data, together with their primary interplay types, including redundant information (R), visual-unique information (U_v), text-unique information (U_t), and synergistic information (S)

Primary Task	Secondary Capability	Samples	Percentage	Interplay
Multimodal Understanding (MMU)	**MMU Total**	**9.743M**	**100.0%**	R, U_v, U_t, S
	Image Captioning & Scene Understanding	3.086M	31.7%	R, U_v
	Knowledge-intensive Visual Understanding	2.105M	21.6%	U_t, S
	General VQA & Instruction Following	1.630M	16.7%	U_v, S
	Object, Attribute, State & Counting	1.558M	16.0%	U_v
	Scene Text & OCR	0.437M	4.5%	U_v
	Visual Grounding & Spatial Understanding	0.366M	3.8%	U_v, S
	Hallucination Suppression & Hard Negatives	0.324M	3.3%	U_v
	Relational & Compositional Reasoning	0.236M	2.4%	S
Text-only	Text Understanding & Knowledge	**3.957M**	**100.0%**	U_t
Total Instruction Data		**13.700M**	–	R, U_v, U_t, S

A.2 Interplay-aware data organization.

After data curation, we organize the 13.7M instruction examples according to four forms of task-relevant multimodal interplay: redundant information (R), visual-unique information (U_v), text-unique information (U_t), and synergistic information (S). Rather than exhaustively annotating every training example, we perform the analysis at the sub-task level. For each sub-dataset, we randomly sample 100 examples and combine task-specific heuristics with model-assisted verification to characterize its dominant information structure. Specifically, we examine whether the two modalities provide aligned or overlapping evidence, whether prediction mainly depends on the visual or textual modality, or whether both modalities are jointly required to derive information unavailable from either modality alone. Each sampled example is scored along these criteria and assigned to its dominant interplay type. The aggregated assignments are then used to characterize the corresponding sub-task and organize its training examples. As summarized in Table 7, this procedure provides a unified interplay-based organization across diverse multimodal understanding tasks, ranging from captioning and visual perception to knowledge-intensive reasoning and compositional inference.

The resulting instruction corpus contains 13.7M examples, including 2.8M redundant, 4.0M visual-unique, 4.2M text-unique, and 2.7M synergistic examples. We use these categories to construct the three-stage curriculum described in Section 2.2, progressively shifting the training emphasis from preserving pretrained capabilities, to strengthening visual perception, and finally to promoting cross-modal synergy.

Table 8: Comparison between uniform data mixing and interplay-oriented curriculum learning using 10% of the SFT data.

Metric	Phase 1			Phase 2			Phase 3		
	Uniform	Curriculum	Diff.	Uniform	Curriculum	Diff.	Uniform	Curriculum	Diff.
MMStar-P	49.52	45.08	-4.44	49.42	49.72	+0.30	51.67	53.87	+2.20
MME-P	917.1	1030.9	+113.8	1135.8	1178.5	+42.7	1249.5	1211.6	-37.9
MMVP	26.67	28.00	+1.33	32.67	26.00	-6.67	29.33	31.33	+2.00
CVBench-2D	68.64	68.08	-0.56	73.02	72.11	-0.90	73.99	74.62	+0.63

Table 9: Evaluation on remote-sensing benchmarks. **RS Data** denotes the number of remote-sensing samples used for fine-tuning.

Model	RS Data	FGRS	FGRC
BAGEL	–	42.51	36.65
GeoChat	–	53.47	21.90
SkySenseGPT	1.415M	**79.76**	**55.50**
Lumina-DiMOO*	50K	77.77	–
Gestalt	50K	<u>79.34</u>	<u>38.20</u>

* FGRC evaluation did not complete successfully.

A.3 Image generation data.

In addition to the above understanding-oriented instruction data, we collect 2.46M image-generation examples. Since these examples primarily maintain and improve conditional image-generation capability rather than determine the interplay curriculum, we distribute them approximately uniformly across the three stages.

B Additional Results

B.1 Effect of Interplay-Oriented Curriculum Learning.

To investigate the effect of our interplay-oriented curriculum, we conduct a controlled study starting from the same CPT checkpoint and using the same 10% subset of SFT data. We compare two training strategies: uniformly mixing all training samples throughout SFT and progressively organizing them according to the interplay-centric curriculum. We evaluate both general multimodal capability and visual perception across different stages of training.

As shown in Table 8, the interplay-oriented curriculum induces a clear shift in the capabilities learned across stages. In the early stage, where training places greater emphasis on redundant and text-unique information, curriculum learning yields substantially stronger broad visual perception, improving MME-Perception by 113.8 points over uniform mixing, while fine-grained visual perception remains weaker. As training progressively shifts toward modality-unique and synergistic information, the model becomes increasingly better at exploiting detailed and task-relevant visual cues. In particular, the curriculum initially shows a clear advantage on broad perceptual benchmarks such as MME-Perception, while its performance on MMStar gradually improves across stages and eventually surpasses uniform mixing.

These results indicate that the ordering of different interplay types affects which capabilities are emphasized during training. Earlier redundancy- and text-oriented stages favor broad perceptual capability, whereas later modality-unique and synergy-oriented stages increasingly strengthen task-sensitive visual reasoning and fine-grained perception. Overall, the curriculum does not simply provide a uniform gain across benchmarks, but progressively reshapes the model toward different aspects of multimodal capability as the information composition of the training data changes.

B.2 Adaptation to Remote-Sensing Tasks.

To examine whether Gestalt can efficiently adapt to a substantially different visual domain, we introduce only 50K remote-sensing VQA samples from FIT Luo et al. (2024) into the training mixture and evaluate the resulting model on FGRS and FGRC. The two benchmarks emphasize complementary aspects of remote-sensing understanding: FGRS evaluates fine-grained recognition and understanding of remote-sensing imagery, whereas FGRC places greater emphasis on reasoning about relations among objects and regions. To provide a controlled comparison in terms of domain-specific data exposure, we additionally adapt the official Lumina-DiMOO checkpoint using the same 50K remote-sensing samples, denoted as Lumina-DiMOO*.

As shown in Table 9, Gestalt achieves 79.34 on FGRS, substantially outperforming the general-purpose BAGEL baseline and GeoChat, and approaching the 79.76 performance of SkySenseGPT despite being exposed to only 50K remote-sensing samples, compared with 1.415M for SkySenseGPT. Under the same 50K domain-specific data exposure, Gestalt also surpasses Lumina-DiMOO* on FGRS. On the more relation-oriented FGRC benchmark, Gestalt further improves over BAGEL and GeoChat, although a gap remains compared with the heavily specialized SkySenseGPT. These results show that Gestalt can effectively absorb domain-specific visual knowledge from limited additional data while retaining strong general multimodal capabilities, demonstrating its adaptability and data efficiency when extending to new domains.

B.3 Qualitative Analysis of Fine-Grained Generation.

Figure 7 presents qualitative examples on UniGenBench, covering diverse visual styles, multi-object compositions, attribute constraints, and spatial relations. Across these elaborate and challenging prompts, Gestalt demonstrates strong instruction-following ability while maintaining coherent scene composition and visual quality.

In particular, Gestalt can faithfully organize multiple semantic constraints within a single image. For example, the generated scenes preserve clear object identities and relative layouts in complex compositions, such as the polar bear and penguins on a snowy field, the camel caravan against the pyramids, the couple and cruise ship along the Seine, and multiple flower arrangements with distinct containers. These examples require more than generating the requested objects independently: the model must correctly bind attributes to their corresponding entities and arrange them according to the specified scene structure. Gestalt also handles fine-grained object properties, as illustrated by examples involving similar bottles with different internal contents, flowers placed in different types of vases, and a puppy positioned beside an empty metallic bowl.

Meanwhile, Gestalt exhibits strong controllability over visual style and global appearance. The generated results span substantially different styles, including watercolor, Van Gogh-inspired oil painting, Impressionism, surreal digital art, product photography, cinematic landscapes, and Japanese anime, while preserving the semantic content of the corresponding prompts. Notably, stylistic constraints affect the entire visual composition rather than appearing as superficial local textures: lighting, color palette, brushwork, perspective, and scene atmosphere remain mutually consistent within each image. The watercolor night scene, warm cinematic desert landscapes, and Impressionist canal scene provide particularly clear examples of such global stylistic coherence.

Overall, these qualitative results highlight Gestalt's ability to jointly model *what* should appear, *where* different elements should be placed, and *how* the resulting scene should be rendered. This is especially important for fine-grained text-to-image generation, where success depends not only on visual realism but also on preserving compositional and attribute-level constraints from the input instruction. Together with the quantitative results on UniGenBench, the examples suggest that Gestalt can effectively translate detailed textual specifications into coherent visual structures across a wide range of generation scenarios.

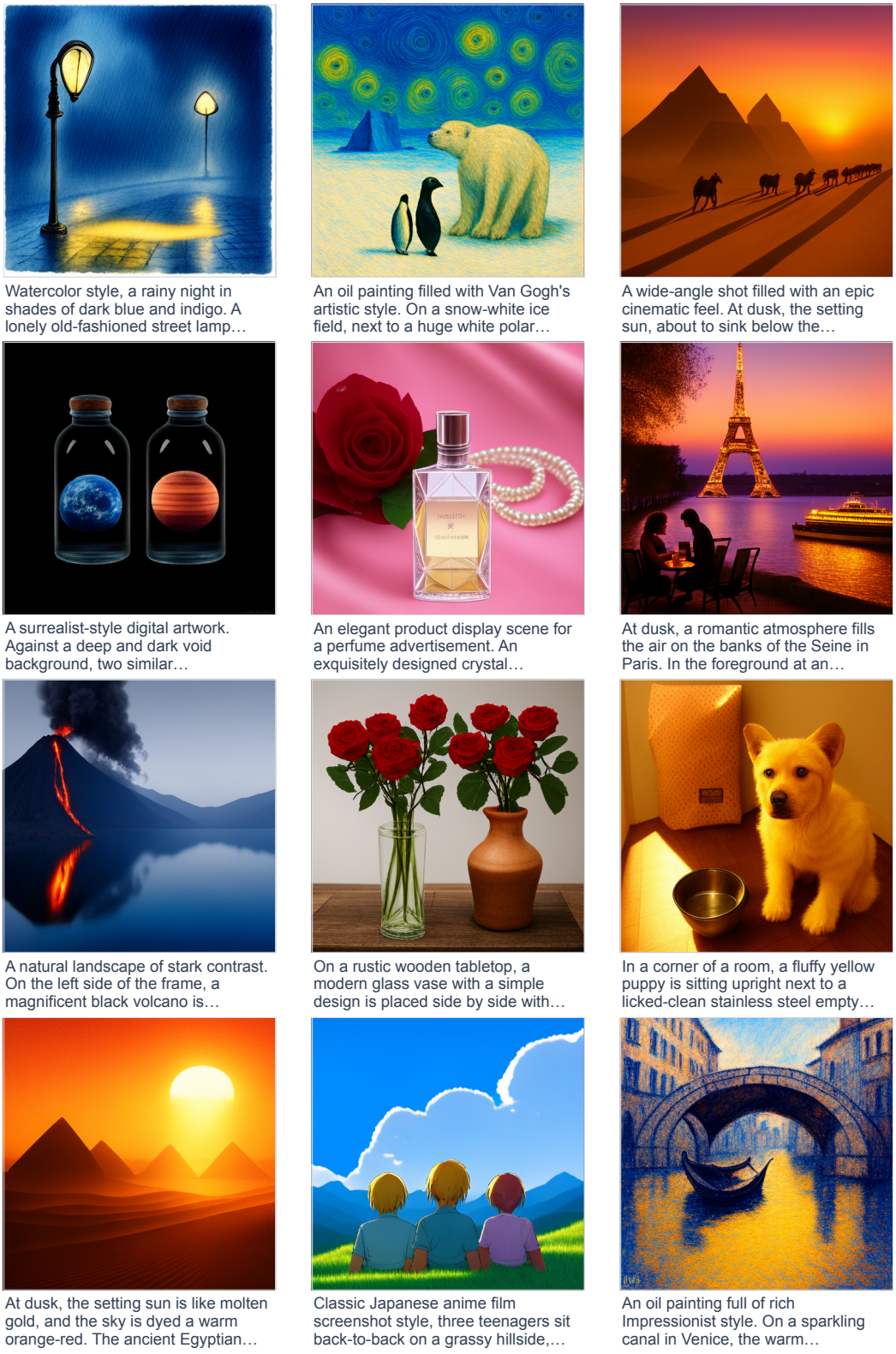

Figure 7: Qualitative results on UniGenBench. Gestalt demonstrates strong fine-grained generation capabilities on complex prompts involving multiple objects, attributes, and spatial relations.