# A Survey on Bridging VLMs and Synthetic Data

**Mohammad Ghiasvand Mohammadkhani**[1] **Saeedeh Momtazi**[1] **Hamid Beigy**[2]
[1]Amirkabir University of Technology [2]Sharif University of Technology
{mohammad.ghiasvand, momtazi}@aut.ac.ir beigy@sharif.edu

## Abstract

Vision-language models (VLMs) have significantly advanced multimodal AI by learning joint representations of visual and textual data. However, their progress is hindered by challenges in acquiring high-quality, aligned datasets, including issues of cost, privacy, and scarcity. On the other hand, synthetic data, created through the use of generative AI—which can even include VLMs—offers a scalable and cost-effective solution to these challenges. This paper presents the first comprehensive survey on bridging VLMs and synthetic data, exploring both the role of synthetic data in VLMs and the role of VLMs in synthetic data. First, we provide a preliminary overview by briefly explaining the architecture of two basic VLMs and, after studying a large number of previous works, offer an extensive survey of the previously proposed methodologies and potential future directions in this area. The repository for this work is available at https://github.com/mghiasvand1/Awesome-VLM-Synthetic-Data.

## 1 Introduction

In recent years, vision-language models (VLMs) have emerged as a transformative class at the intersection of computer vision and natural language processing. As (Qi et al., 2024b) noted, by learning joint representations from both modalities and bridging the gap between them, remarkable progress has been achieved in various tasks involving visual and textual content. However, their advancement heavily depends on high-quality datasets with well-aligned visual and textual information. Collecting such resources is expensive and time-consuming, and these issues, along with challenges such as data scarcity and privacy concerns, pose a significant bottleneck to further progress.

Synthetic data has emerged as a promising direction, offering automatic approaches for data generation aimed at replacing or augmenting real-world data. With rapid advancements in generative AI models—including large language models (LLMs) and VLMs—using these models in the synthetic data generation process has become a prevalent strategy for many tasks. Thanks to their strong instruction-following capabilities and deep prompt understanding (Liu et al., 2023c), developed during the instruction-tuning phase (Liu et al., 2023b; Peng et al., 2023), generative models can be used in a controlled and scalable way to produce synthetic datasets. Compared to manual data construction or algorithms not based on generative AI, this approach yields higher-quality data while also being more customizable and cost-effective.

The role of synthetic data in LLMs and the role of LLMs in synthetic data have been extensively explored in previous surveys. For example, (Liu et al., 2024d) provides an overview of applications in which synthetic data has been utilized within the scopes of training or evaluation, such as reasoning, alignment, and factuality. Similarly, (Tan et al., 2024) reviews studies in which LLMs themselves functioned as data annotators, examining three key aspects: annotation generation, annotation assessment, and annotation utilization. (Long et al., 2024) surveyed works that focused on LLM-driven synthetic data, including its generation through different approaches, its curation, and even its evaluation through direct or indirect methods. (Wang et al., 2024b) emphasized LLM-oriented data synthesis, covering the role of synthetic data throughout the full lifecycle of LLMs and across their core functionalities. However, unlike the role of synthetic data in LLMs or vice versa—and other topics in VLMs previously explored in surveys, such as hallucination (Liu et al., 2024a), prompt engineering (Gu et al., 2023), alignment and evaluation (Li et al., 2025d)—a deep and focused survey on bridging VLMs and synthetic data has yet to be conducted.

In summary, this study thoroughly explores the VLM-Synthetic Data bridge by considering the role of VLMs in synthetic data and the role of synthetic data in VLMs, surveying 125 papers and proposing potential future directions. To the best of our knowledge, it is the first focused attempt to investigate this area within a survey paper. Section 2 provides an explanation of the architecture of two basic VLMs as a preliminary. Section 3 comprehensively surveys previous related works that meet the condition mentioned in that section, and in Section 4, we present several future directions based on our findings from exploring this area.

## 2 Preliminary

In this section, we briefly describe the architecture of two common and basic VLMs for a better understanding of the following concepts. These two are CLIP (Radford et al., 2021) and LLaVA (Liu et al., 2023b), and as (Ghosh et al., 2024) noted, they were developed to follow the goals of vision-language understanding and text generation with multimodal input, respectively.

### 2.1 CLIP Architecture

The pretraining procedure involves a vision encoder $e$ and a text embedding model $e'$ to encode each image-text pair in the training batch $\{(i_j, t_j)\}_{j=1}^{n}$, with size $n$. Images are encoded into embeddings $e_{i_j} \in \mathbb{R}^{d_e}$ and texts into embeddings $e'_{t_j} \in \mathbb{R}^{d_{e'}}$, where $d_e$ and $d_{e'}$ are the output dimensions of the vision and text encoder models, respectively. Two linear projection matrices, $W_i \in \mathbb{R}^{d_e \times d}$ for images and $W_t \in \mathbb{R}^{d_{e'} \times d}$ for texts, project the encoded representations into a joint embedding space of dimension $d$. Denoting the batch matrices of embeddings as:

$$E_i = [e_{i_1}, \ldots, e_{i_n}]^\top \in \mathbb{R}^{n \times d_e}$$
$$E_t = \left[e'_{t_1}, \ldots, e'_{t_n}\right]^\top \in \mathbb{R}^{n \times d_{e'}}$$

The projections are computed as:

$$Z_i = E_i W_i, \quad Z_t = E_t W_t$$

Next, each row of the $Z_i, Z_t \in \mathbb{R}^{d \times n}$ projections is normalized, creating $\hat{z}_i, \hat{z}_t \in \mathbb{R}^{d \times n}$, respectively. Then, the similarity matrix $S \in \mathbb{R}^{n \times n}$ is computed, where each entry $S[i, t]$ measures the similarity between image $i$ and text $t$ via their dot product, scaled by a temperature parameter $\tau \in \mathbb{R}$:

$$S = \hat{z}_i^\top \hat{z}_t \times \exp(\tau)$$

CLIP then computes a symmetric contrastive loss over the batch, consisting of two cross-entropy losses: one for predicting the correct text given an image, and another for predicting the correct image given a text. Finally, the model optimizes all learnable parameters to minimize this loss.

### 2.2 LLaVA Architecture

The architecture includes a vision encoder $e$ and a decoder-only LLM denoted as $f$, with the word embedding dimension $d_f$. Given a sample image–question–answer triplet $(i, q, a)$, the vision encoder first produces $E_i \in \mathbb{R}^{p_e \times d_e}$, which represents the encoded image, where $p_e$ is the number of patches the encoder divides the image into, and $d_e$ is the encoding dimension. A projection matrix $W \in \mathbb{R}^{d_e \times d_f}$ is then applied to map these encoded features into the LLM's word embedding space:

$$H_i = E_i W \in \mathbb{R}^{p_e \times d_f}$$

Considering $H_q \in \mathbb{R}^{m \times d_f}$ as the embedding representation of the question, where $m$ is the number of tokens in the question, the concatenated representation for the input is formed as follows:

$$X = \begin{bmatrix} H_i \\ H_q \end{bmatrix} \in \mathbb{R}^{(p_e + m) \times d_f}$$

which is then fed into the LLM, and this model autoregressively outputs logits at each step $t$:

$$\ell_t = f(X, a_{<t}) \in \mathbb{R}^{|\mathcal{V}|}$$

where $\mathcal{V}$ is the LLM vocabulary set and $a_{<t}$ denotes the previously generated tokens. These logits are turned into probabilities via the softmax:

$$p(a_t \mid i, q, a_{<t}) = \frac{\exp\left(\ell_t^{(a_t)}\right)}{\sum_{v \in \mathcal{V}} \exp\left(\ell_t^{(v)}\right)}$$

The total conditional likelihood of the grounded answer $a = (a_1, \ldots, a_T)$ is defined as follows:

$$p(a \mid i, q) = \prod_{t=1}^{T} p_\theta(a_t \mid i, q, a_{<t})$$

This probability is used to calculate the negative log-likelihood loss, which is minimized by updating $\theta$, the set of trainable parameters, which was set to $\{W\}$ in the pretraining stage.

## 3 VLMs and Synthetic Data

This section surveys prior works—grouped based on the main task they addressed—in which one of the following criteria is met: (i) *it proposes a synthetic data generation process utilizing VLMs and uses the data for an AI model*, or (ii) *it generates synthetic data without VLM involvement by using other generative AI models, such as LLMs or diffusion models, but later employs it for VLMs*. The first criterion addresses the role of VLMs in synthetic data and that of synthetic data (via VLMs) in VLMs, while the second considers the role of synthetic data (via non-VLM generative AI models) in VLMs—together examining both the role of VLMs in synthetic data and vice versa.

### 3.1 Instruction-Tuning & Alignment

***TextBind*** (Li et al., 2023b) tackles the challenge of creating high-quality multimodal instruction data using accessible image-caption pairs. It selects diverse image samples via k-means on image encodings, then uses a text-only LLM fed with captions to generate multi-turn conversations, followed by post-processing for data quality.

***LLaVA*** (Liu et al., 2023b) bases the use of synthetic data for VLM instruction tuning by replacing images with captions and using bounding boxes, followed by employing a text-only LLM with few-shot examples to generate QA pairs in conversational, descriptive, and complex reasoning formats.

***LLaVAR*** (Zhang et al., 2023) enhances text-rich image comprehension in instruction tuning by selecting relevant images via an image classifier, reducing noise with clustering, extracting text using two OCR tools, and generating captions with a captioning model. A text-only LLM then creates conversational instruction data from the captions and extracted texts.

***LVIS-Instruct4V*** (Wang et al., 2023) resolves inconsistencies in multimodal instruction data generation without image inputs by using GPT-4V with precise prompts to produce conversational data through self-reasoning and detailed image descriptions based on bounding box information.

***LLaVA-Med*** (Li et al., 2023a) improves biomedical image comprehension through instruction tuning by gathering biomedical images and using their captions or related text to prompt GPT-4 for generating conversational question-answering data.

***STIC*** (Deng et al., 2024b) enhances self-generated preference data for image comprehension using only unlabeled images, generating preferred samples via step-by-step engineered description prompts, and dispreferred ones through image corruption or poor prompting.

***VLFeedback*** (Li et al., 2024c) removes the need for human annotation via scaled alignment, using a pool of VLMs to answer diverse instructions and ranking responses by helpfulness, visual fidelity, and ethical consideration to build a preference annotation dataset.

***InBoL*** (Wang et al., 2024e) enables answer refusal for insufficiently informative questions by generating candidate responses for an image-question-answer triplet, using an answer-grounded LLM-as-a-Judge to select correct and incorrect responses. It also modifies questions to be unanswerable and generates an incorrect answer. Data is categorized based on confidence, creating pairs of chosen-rejected, correct-refusal, correct-incorrect, and refusal-incorrect.

***Multimodal Self-Instruct*** (Zhang et al., 2024b) improves models for daily tasks by proposing a visual idea, prompting an LLM to generate related text and code, executing the code to create a synthetic image, then generating QA pairs and enriching answers with strong reasoning.

***ALLaVA*** (Chen et al., 2024c) addresses scaling efficiency by relying on rich synthetic data rather than model size. It introduces compact models trained on instruction data constructed through the steps of image selection, fine-grained captioning, diverse and complex questioning, and detailed answering.

***Agri-LLaVA*** (Wang et al., 2024c) focuses on agricultural domain tuning by aligning features using pest and disease data, and generating knowledge-based synthetic instructions through multi-turn conversations grounded in agricultural images and web-sourced textual knowledge.

***MM-Instruct*** (Liu et al., 2024b) broadens instruction tuning by using an LLM to generate application-based instructions for each image-description pair. Instructions are clustered via k-means, and representative ones are selected. CLIP retrieval selects the best instruction per image, which, along with the description and few-shot examples, is used to generate the final output.

***VIGC*** (Wang et al., 2024a) tackles multimodal data generation, where text-only methods struggle with image content. It employs separate generation and correction models, trained on quadruplets (image-instruction-question-answer) and triplets (image-question-answer), respectively. The generation model produces QA pairs, and the correction model refines answers sentence by sentence.

***C*$^3$*L*** (Ma et al., 2024a) enhances visual-textual content relevance by defining a correspondence score based on the KL-divergence between answer token probabilities with and without the input image. Initially, synthetic instruction data is provided, and high-scoring samples are selected for fine-tuning. To mitigate exposure bias, contrastive training treats the highest-scoring QA pair as positive and others as negative. This two-stage process enables the model to generate instruction tuning data with strong image-text alignment.

***GENIXER*** (Zhao et al., 2024) generates data without relying on GPT models by collecting large-scale instruction data and designing a two-level instruction chat template. The VLM is trained on this data to serve as an independent data generator.

***Vision-Flan*** (Xu et al., 2024) tackles bias and diversity issues in instruction tuning data by first applying instruction tuning with diverse, high-quality expert-written data, followed by additional tuning with a small amount of synthetic data.

***STLLaVA-Med*** (Sun et al., 2024) tackles data scarcity in medical question answering through a two-step strategy: training VLMs to generate and answer questions from input images, then using GPT-4o to create preference data from the generated content for alignment.

***BioMed-VITAL*** (Cui et al., 2024) addresses misalignment between biomedical instruction data and domain expertise by using expert-driven few-shot demonstrations to generate instruction data, then filtering misaligned data with a classifier trained on human and model preference data.

***CaReVL*** (Dai et al., 2025) handles non-aligned answers by generating response candidates for an image-question pair, evaluating them with GPT-4 and a caption-based reward model pool, and using them for instruction tuning as high-confidence data if agreement is sufficient, or for negative sampling as low-confidence data otherwise.

***OmniAlign-V*** (Zhao et al., 2025a) tackles human preference alignment in multimodal models by using unlabeled images, discarding low-content ones, and generating synthetic QA pairs from seed tasks with varied prompt templates. After refinement, it aligns preferences by generating high-temperature responses, selecting the worst via a judge, and pairing it with the gold answer to create negative-positive preference data.

***Unicorn*** (Yu et al., 2025) lowers multimodal data costs by removing real images, generating a seed dataset from captions, expanding it, and creating diverse instruction data using a text-only LLM. Modality representation transfer, based on modality gap theory, enriches the data with corresponding synthetic image embeddings.

***EndoChat*** (Wang et al., 2025) improves VLMs' surgical scene understanding by extracting scene attributes, generating diverse conversational data with various instruction types, and fine-tuning the model on this data.

***AnyPrefer*** (Zhou et al.) mitigates self-rewarding alignment bias by generating candidate answers for an image-question pair using the target model, ranking them with a tool-assisted judge, and iteratively updating prompts if the score gap between top and bottom answers is below a threshold, before adding the pair to a synthetic preference dataset.

***CoSyn*** (Yang et al., 2025) enhances VLMs' text-rich image understanding by generating diverse data content, rendering images through generated code with tools, and prompting LLMs with the code to produce instruction data.

### 3.2 Evaluation

***Auto-Bench*** (Ji et al., 2023) evaluates the alignment of VLMs with human intelligence. Each image in the collection is textually described using captions, object locations, and instance relationships. This information is used by an LLM to generate QA data covering capabilities in reasoning, perception, values, and planning, forming a benchmark dataset after human verification.

***AutoHallusion*** (Wu et al., 2024b) addresses poor generalization in handcrafted hallucination benchmarks via a data generation pipeline. It enriches short image descriptions with an LLM, generates images using a diffusion model, identifies specific objects with a VLM, manipulates images using other models, and constructs existence and spatial questions for VLM evaluation.

***MVP-Bench*** (Li et al., 2024a) explores the perception gap between VLMs and humans through evaluation. It generates ideas about image aspects (background, clothes, facial state) using chain-of-thoughts, applies image manipulations via generative models, and constructs benchmark data with high and low-level visual perception questions, including cross-image and yes/no formats.

***VLBiasBench*** (Wang et al., 2024d) addresses data insufficiency in existing bias evaluation benchmarks. It constructs diverse prompts for diffusion-based image generation using two methods: (1) extracting keywords from a text corpus, enriching them with related terms and applying style control, and (2) leveraging GPT for prompt creation. Subsequently, it generates questions to describe image content, embedding bias-inducing elements to elicit biased responses from models, thereby forming a benchmark for evaluation.

***SMMQG*** (Wu et al., 2024a) addresses the scarcity of multimodal retrieval-augmented generation evaluation data by sampling a seed source, extracting a key entity, retrieving cross-modal content, and prompting an LLM with few-shot examples to generate a QA pair and reference sources. The generated data is kept for retriever and generator evaluation if it meets verification criteria such as answer correctness and the necessity of all references.

***VIVA*** (Hu et al., 2024b) constructs the first benchmark to evaluate VLMs on decision-making grounded in human values. It starts with collecting seed image–description pairs, then employs an LLM to brainstorm new situations. Corresponding images are gathered, and action candidates are generated along with their rationale and underlying human values, which are then used to evaluate VLMs via a judge model.

***MLLM-as-a-Judge*** (Chen et al., 2024b) addresses the divergence between human preferences and model-based multimodal evaluation, whether via scoring or batch ranking. It collects diverse images and instruction templates to create image–instruction pairs via random sampling. Model responses are then gathered in scoring, pairwise comparison, and batch ranking formats, with human annotations used as evaluation references.

***Prometheus-Vision*** (Lee et al., 2024) trains a VLM to act as an evaluator, taking an image, instruction, response, rubric, and reference as input, and outputting feedback and a score. It begins by collecting images and hand-crafted seed data, then uses GPT-4V to generate new rubrics, augmented instructions and references, as well as novel feedback and responses, to train an open-source VLM.

***VL-RewardBench*** (Li et al., 2024b) conducts a meta-evaluation of multimodal reward models using a benchmark built from instruction, hallucination, and reasoning datasets. For instruction and hallucination data with preference annotations, it uses a pool of small VLMs as judges and retains samples where most judges fail, thereby exposing their limitations. For reasoning data without preferences, it generates answer candidates, evaluates them using GPT-4o and the gold answer, and retains the full sample if at least one synthetic response is correct.

***ConMe*** (Huang et al., 2024a) constructs a compositional reasoning benchmark via a multi-stage pipeline. Starting with an image collection, both a strong model and smaller VLMs generate image descriptions. The strong model then formulates challenging reasoning questions with one correct and several adversarial answers. Two-choice QA pairs are posed to the smaller VLMs, and those that all smaller models answer correctly are discarded to ensure the remaining samples induce failure. The smaller VLMs subsequently provide reasoning texts, which the strong model uses to generate final QA pairs for evaluation.

***UNO*** (Wu et al., 2025) adds subject-level visuals to images based on instructions. An LLM generates a scene description for a diffusion model, which returns two images: the first with the main subject, and the second combining it with an added object. A VLM assesses the images in a chain-of-thought format to remove inconsistencies. The final inputs—cropped images of the main subject and the added object, plus the instruction—are paired with the mixed-subject image as the training output.

***LSDBench*** (Qu et al., 2025) introduces a benchmark for evaluating VLMs on long-video comprehension. Its data synthesis involves segment-level video captioning, hierarchical clustering of scenes and actions, LLM-based question generation and refinement using cluster summaries, followed by feeding the question, summary, and video clip to a multimodal model for answer generation, and formatting the QA as multiple choice using an LLM.

***SPO*** (Liang et al., 2025b) introduces two synthetic benchmarks to evaluate VLMs on long-horizon planning. In the first, GPT-4o takes an object list to generate a task proposal and plan, which are reviewed under specific configurations, simulated, and the result is used by the VLM to refine the task-plan pair. In the second, using a different simulator, GPT-4o creates task templates from scene assets; objects are then sampled to instantiate tasks, simulated, retained only if successful, and augmented to enrich task instruction diversity.

***YesBut*** (Liang et al., 2025a) focuses on assessing VLMs' understanding of humor. It constructs a benchmark by collecting image pairs with contradictory humor and synthetically generating foundational and deep image content—including descriptions, contradiction explanations, and underlying symbolism—through a human-AI collaborative process. Each generated text is used as a reference to evaluate the image pairs using a judge model.

***DrawEduMath*** (Baral et al., 2025) addresses the challenge of VLM content understanding in noisy visual contexts. It begins by collecting student response images along with expert-authored textual descriptions. These descriptions are then decomposed into atomic points using an LLM, from which synthetic question–answer pairs are generated to create evaluation data for each image.

***KITAB-Bench*** (Heakl et al., 2025) introduces a benchmark for evaluating VLMs' Arabic OCR performance. The data synthesis process begins with generating sample topics, each expanded into multiple topic names. Raw data is then generated within these topic scopes, followed by code generation to render the raw data into various visual formats.

### 3.3 Multimodal Reasoning

***ComVint*** (Du et al., 2023) highlights the role of visual reasoning training data for general instruction following. Starting with an image-caption-objects triplet, it is converted into an instruction-response format using an LLM with caption and object information. The datapoint is then made more complex with LLM, and if verified by a judge model, the instruction and response are reformulated into the considered format.

***NavGPT-2*** (Zhou et al., 2024a) focuses on instruction-following for navigational reasoning. Training data generation is a single-step procedure where an image and instruction are provided, and the VLM is prompted to follow the instruction based on observations of surrounding components and pathways to determine the next step.

***VPD*** (Hu et al., 2024a) addresses limitations of visual reasoning methods based on programming or tool use, such as cost, latency, and errors. Given an image–question–answer triple, a tool-assisted LLM generates candidate programs; the correct one is selected, and its entire procedure is written in a chain-of-thought format to fine-tune the VLM.

***R-CoT*** (Deng et al., 2024a) addresses the limited accuracy and diversity of geometry reasoning data generation methods. Given image-description pairs, the method patches the description and applies single-step prompting to reason on the image, continuing in a step-by-step reasoning format called chain-of-thought fusion. Each reasoning step is treated as an answer, and a corresponding question is generated to form the training data.

***Math-LLaVA*** (Shi et al., 2024) addresses the low amount of question-answer pairs for each image in instruction data. It first collects image-question-answer triplets and uses a multimodal model to select high-quality, complex examples. Then, synthetic question-answer pairs are generated for each image using zero- or few-shot prompting to ensure the dataset fully captures each image's content.

***ShareGPT-4o-Reasoning*** (Zhang et al., 2024a) tackles the limited effectiveness of minimal rationales in training data by first applying chain-of-thought data distillation to fine-tune the VLM, then using high-temperature sampling to generate diverse answers and form positive-negative answer pairs for further alignment.

***Least-to-Most*** (Cheng et al., 2024) addresses the shortage of training datasets with multi-step vision and language processing and the limitations of proprietary models for reasoning data generation. It proposes a bottom-up pipeline for data synthesis, creating reasoning paths by relating adjacent objects in an image, forming sub-questions answerable via tool use, and progressively generating higher-level sub-questions that combine earlier ones until reaching the main question-answer.

***URSA*** (Luo et al., 2025) addresses the scarcity of chain-of-thought data for multimodal mathematical reasoning. After pretraining on math contexts for improved vision-language understanding, it generates training data by converting other data types (e.g., answer-only) into CoT using a multimodal model, beyond existing chain-of-thought datasets.

***MM-Verify*** (Sun et al., 2025) introduces a strong verification mechanism for multimodal reasoning. It collects a large question pool, uses a simulation-based algorithm with a binary tree for chain-of-thought candidate solutions, and applies a strong model to generate synthetic verifications for each tree path. After extracting answers and verifying solutions, the data is cleaned and used for training.

***SMIR*** (Li et al., 2025a) emphasizes multi-image reasoning data for training or evaluation. It computes multimodal embeddings as a weighted sum of textual and visual embeddings from a large image-caption dataset, then gathers related images via random sampling to ensure coherence and diversity. Finally, multi-turn conversation data is synthetically generated by creating question-answer pairs.

***TACO*** (Ma et al.) tackles multimodal complex question answering with multistep reasoning. Its data generation process involves an image-question-answer triplet, producing a multistep answer through a chain-of-thought and action. The data is added to the dataset after successful verification and parsing.

***Text-Only Training*** (Hu et al., 2025) explores improving VLMs in human-centered reasoning by enhancing their LLM component with synthetic data. For each image, a situation description is generated, and manually crafted seed questions are used as in-context demonstrations to prompt an LLM to create a decision-making question, including a rationale. The generated dataset is then used to train the model as a reasoning data generator.

***SCRAMBLe*** (Mishra et al., 2025) addresses the weakness of VLMs to differentiate similar compositions. Given image-caption pairs, it prompts an LLM with few-shot examples and step-by-step chain-of-thought guidance to generate new meaningful captions by rearranging word order. It then forms positive-negative preference pairs to improve VLM compositional ability through alignment.

***MindGYM*** (Xu et al., 2025) focuses on synthetic self-questioning to enhance reasoning. For data generation, given an image-question pair, it creates several single-hop questions from defined perspectives, then combines them to form challenging multi-hop questions, aiming to improve thinking breadth and depth. This strategy is applied across datasets to fine-tune the model using circilium fine-tuning, which regularly varies the input-output formats of the question-thought-answer triplet to ensure a guided learning pathway.

***MMR*** (Jang et al., 2025) is motivated by the shortage of broad object-level reasoning in existing datasets. For data synthesis, it annotates an image collection with object- and part-level information, sets a system message to customize GPT's role, and prompts it with the image to generate a global caption and QA pairs under specific conditions.

### 3.4 Retrieval-Augmented Tasks

***VCoT*** (Rose et al., 2023) bridges multimodal gaps for data synthesis by replacing images with captions, generating intermediate texts via an LLM, retrieving the best using CLIP embeddings based on initial inputs, and generating image candidates via a diffusion model, again using CLIP to select the best image to form the intermediate pair.

***MegaPairs*** (Zhou et al., 2024b) tackles the shortage of multimodal retrieval training data by synthesizing source–instruction–target triplets, where the source and target are relevant images. It employs multiple similarity models to retrieve related image pairs, uses a VLM to describe their relationships, and then leverages an LLM to generate instructions from these descriptions to form the triplets.

***VisRAG*** (Yu et al., 2024) enhances multimodal retrieval-augmented generation using VLMs with separate retriever and generator modules: the retriever selects top relevant documents per query, and the generator uses them to produce answers. Training is augmented with synthetic query–document pairs, created by prompting a VLM to extract answers from grounded document images and generate corresponding queries.

***InstructCIR*** (Zhong et al., 2024) addresses compositional image retrieval to better follow modification-based instructions. A strong model is prompted with the image caption, a guided prompt, and few-shot demonstrations to brainstorm novel modifications in a chain-of-thought format, resulting in a modified caption. The VLM is then trained using triplet data containing the source caption, instruction, and target caption.

***VISA*** (Ma et al., 2024b) addresses visual source attribution in retrieval-augmented generation. After retrieving relevant documents, the generator outputs both an answer and a bounding box, ensuring the answer is grounded in the specified region. Using a dataset with annotated bounding boxes, a VLM is prompted to generate question-answer pairs from defined regions. The resulting training data uses questions and pure images as inputs, and answers with bounding boxes as outputs.

***SK-VQA*** (Su et al., 2024) explores synthetic training data for multimodal context-augmented generation using a single-step pipeline. A VLM, given an image and guiding points, first generates contextual knowledge, then QA pairs based on the context and image. Samples overly reliant on the image for context or QA are filtered out before training use.

***NegBench*** (Alhamoud et al., 2025) targets the comprehension of negated captions by extracting objects from image captions as positive concepts, then prompting an LLM to generate related negative concepts. It constructs multiple-choice questions by pairing concepts that may or may not be present in the image, and these choices are subsequently paraphrased by the LLM into refined captions.

***ImageRef-VL*** (Yi et al., 2025) addresses contextual image referencing in query answering by generating responses that follow prompts while citing input images by ID. For data construction, an LLM creates an initial textual response, and a VLM generates descriptions for all input images. These are used to produce image captions, which, along with the initial response, guide the VLM to generate an interleaved output with contextual citations. This output, combined with the prompt and images, is used to fine-tune the VLM.

***good4cir*** (Kolouju et al., 2025) improves composed image retrieval by leveraging synthetic data to enrich instructions, replacing short texts. Given source–instruction–target triplets, it uses a VLM to describe objects and attributes in the source image, generate object descriptions for the target, and enrich the instruction by generating the differences between the two sets of descriptions.

### 3.5 Hallucination

***CIEM*** (Hu et al., 2023) uses synthetic data to evaluate and mitigate hallucination. Given an image-caption pair, an LLM generates factual and contrastive QA pairs—answered with yes and no, respectively—for evaluation. For instruction tuning, the answers are enriched with chain-of-thought reasoning, forming the final QA data.

***HallE-Control*** (Zhai et al., 2023) presents a control layer to regulate hallucination in visual detail comprehension during captioning. It employs a synthetic data pipeline using image-object pairs, where an object categorization model distinguishes grounded and omitted object categories. Two captions are generated by an LLM: one with grounded objects (contextual knowledge, control = -1) and another including omitted objects in brackets (contextual plus parametric knowledge, control = 1). Training a VLM integrated with their control layer enables inference-time hallucination control between -1 and 1. Controllability is evaluated by extracting captioned objects via an LLM and matching them with gold objects to measure hallucination.

***OpenCHAIR*** (Ben-Kish et al., 2023) tackles open-vocabulary hallucination evaluation in image captioning. It uses a caption collection and prompts an LLM with few-shot examples to generate stylistically similar yet diverse captions. A diffusion model then generates corresponding images for these captions, forming benchmark image-caption pairs. For evaluation, objects are extracted from the evaluatee-generated caption and passed, along with the gold caption, to the LLM to identify hallucinations—bypassing closed-vocabulary limitations.

***LRV*** (Liu et al., 2023a) targets both hallucination mitigation and instruction following in VLMs by using LLM-generated negative and positive samples, respectively. Positive data is built from various tasks using images with bounding boxes and dense captions to form concise question-answer pairs in both declarative and interrogative formats. Negative data is generated by manipulating questions—adding nonexistent objects, modifying attributes of existing objects, or altering known facts.

***VGA*** (Ziyang et al., 2024) mitigates hallucination in GUI comprehension caused by over-reliance on textual knowledge by generating two synthetic training datasets: one uses an LLM prompted with the preprocessed Android view textual hierarchy and a guided prompt to create QA data, and the other prompts a VLM with interface images and a guided prompt for the same purpose.

***V-DPO*** (Xie et al., 2024) addresses VLM hallucination in weird image comprehension by aligning on synthetic preference data. This data is created by prompting a VLM with an image to generate detailed captions and object positions, then using an LLM to suggest visual element replacements. After applying these via inpainting, preference pairs are constructed for alignment.

***TLDR*** (Fu et al., 2024) proposes an interpretable reward model for multimodal hallucination detection across token-, sentence-, and response-level granularities. It gathers batched visual QA pairs and uses an LLM to generate detailed captions. These captions, along with primitive ones, are perturbed by the LLM using taxonomy-guided edits to create hard negative examples, with binary token-level labels indicating the perturbations. Finally, the image-caption pairs and the corresponding binary token labels are used for training.

***LongHalQA*** (Qiu et al., 2024) introduces an automated pipeline for generating synthetic hallucination data in long-context settings. It starts by collecting rich-content images, then employs GPT-4V to produce long, grounded positive responses. Hallucinations are identified using object detection and model interpretation. If the initial response is positive or negative, GPT-4V is used to generate its complementary version, forming final QA pairs.

### 3.6 Planning & Manipulation

***TaPA*** (Wu et al., 2023) presents a data generation pipeline to enhance VLMs' indoor planning capabilities. Starting from a collection of indoor scenes with multiple objects, the pipeline uses object positions and a carefully crafted prompt with few-shot examples to query an LLM for generating an instruction-plan pair. Each instruction describes a specific event, and each plan is a sequence of executable actions. The resulting dataset covers complex tasks with diverse and feasible plans.

***ECoT*** (Zawalski et al., 2024) introduces a reasoning-before-acting method for robotics planning. It generates synthetic plans by extracting scene descriptions from images, performing object detection and motion analysis, and prompting a multimodal model with this data and the task description to reason step-by-step and produce plans, subtasks, and movements.

***ReLEP*** (Liu et al., 2024e) uses GPT-4V to generate synthetic data for long-horizon embodied planning, enabling fine-tuning of other VLMs. It collects indoor scene images and prompts the multimodal model to create a robot task, which is refined and used—along with modules of skill library, robot setup, and memory—to produce an execution plan.

***Manipulate-Anything*** (Duan et al., 2024) addresses automatic generation of robot manipulation data using GPT-4V. It first decomposes the main task into subtasks, each processed by an action generation and a subtask verification module. The verification module checks subtask completion based on the temporary goal state, while the action generation module classifies the subtask as object-centric or agent-centric and generates either a task-specific grasp pose or executable action code. This modular design allows flexible handling of diverse manipulation scenarios by tailoring actions to each subtask, forming the demonstrations.

***CogCoM*** (Qi et al., 2024a) trains a VLM to answer questions through an intrinsic chain of manipulations by imagining a robot within the image scenario. Its training data is generated via a pipeline that begins with an image-question pair. An LLM produces a step-by-step path, where each step includes a new question, image view, and manipulation chain. This chain is executed using various tools to obtain intra-chain answers, followed by processing the steps to reach the final answer.

***MM-Traj*** (Gao et al., 2024) constructs a dataset of query-trajectory pairs to train VLMs for tool use. GPT-4o-mini is first prompted with task seeds and a tool list to generate queries. It then receives each query and outputs commands for file generation: images are retrieved from a database, while documents are created via code generation. The model checks if the query is answerable using the available tools and files. If so, a strong agent generates the trajectory—comprising thoughts, code, and observations—in a zero-shot manner, which is added to the training set after passing the checking filter.

***HybridGen*** (Wang and Tan, 2025) leverages VLM-guided hybrid planning in a two-stage workflow—first decomposing human demonstrations into precision-critical segments such as object-centric pose transforms and automatable segments such as VLM-informed path planning, then amplifying them through format-agnostic, pose-only adaptations to yield vast, diverse, high-fidelity imitation learning datasets.

### 3.7 Regional Awareness & Visual Relations

***RegionGPT*** (Guo et al., 2024) improves regional visual comprehension in VLMs by an automated region caption generation pipeline. It first uses a VLM to produce a global image caption, then manually crops a region of interest and combines it with a crafted prompt—including the global caption and the region's class name—to request a region-level caption from the VLM. The resulting data can be used to fine-tune VLMs for region-specific captioning on input images with predefined regions.

***RelationVLM*** (Huang et al., 2024b) improves VLMs' visual relation comprehension by training on synthetically constructed data. It first collects image pairs with visual relations such as semantic, temporal association, or geometric transformation, using their attribute labels. Each image annotation is then fed to an LLM to generate the relation description, which the LLM subsequently uses to produce dialog-style question-answer pairs.

***VisMin*** (Awal et al., 2024) develops a synthetic data generation pipeline to train or evaluate VLMs on fine-grained visual understanding by applying minimal changes to image-caption pairs. In the first approach, an LLM edits object or attribute mentions in captions, while a segmentation model and an inpainting model apply corresponding changes to the image. In the second, the LLM generates prompts with layout information for a diffusion model, which then produces images; the layout is modified by changing object quantity or location before re-generation. A filtering step ensures image-caption consistency, followed by human verification for final data selection.

***RightThisWay*** (Liu et al., 2024c) enhances VLMs' ability to detect insufficient information by fine-tuning on synthetic image-caption pairs. Starting with an image-caption pair, it perturbs the image by shifting the bounding box and queries the VLM. If the VLM can still answer correctly, the perturbed pair is used for training; otherwise, the caption is marked as unanswerable before the data is added.

***COCONut-PanCap*** (Deng et al., 2025) constructs a dataset with detailed scene descriptions for the tasks of fine-grained captioning, segmentation-grounded captioning, and image generation. Its annotation pipeline begins with an image, applies segmentation masks, assigns a label to each mask, and prompts a VLM to draft a detailed caption, which is then human-edited and summarized by the VLM to produce the final image-caption pair.

***PRIST*** (Cai et al., 2025) presents a data generation pipeline using GPT-4o for pixel-level reasoning in multi-turn conversations, applicable to both training and evaluation. The task involves a multimodal model receiving an image with multi-turn questions, answering each, and finally segmenting the target in the image. The generation pipeline begins with an image, described by a VLM in terms of its visual elements. Each description is converted into a complex question, followed by the formation of a reasoning tree: the question is the root, single elements are the leaves, and sub-questions serve as intermediate, multi-level branches. This tree is used to create multi-turn QA data that progressively refines pixel-level target localization.

***SPARCL*** (Li and Li, 2025) improves VLMs' compositional understanding by training on synthetic data containing subtle pair-level differences. Given image-caption pairs, an LLM generates a positive caption—faithful yet explicitly artificial—and a negative caption with slight deviations. A text-to-image model, augmented with injected features, then produces the corresponding positive and negative images. The resulting dataset is used to train CLIP with contrastive learning via an adaptive margin loss, encouraging robust separation between positives, easy negatives, and hard negatives.

### 3.8 Video-Centric Tasks

***Inst-IT*** (Peng et al., 2024a) addresses instance-level comprehension in models. Given a video collection, it provides frames with visual prompts to highlight instances, feeds consecutive frame pairs to a model to generate captions for individual instances, entire images, and temporal changes, and uses these to produce a video-level summary and construct QA pairs for training or evaluation.

***Video-XL*** (Shu et al., 2024) addresses visual loss in long video comprehension. For each QA sample linked to a short video, it splits the video into clips, feeds them to a strong model to extract temporally ordered event clues, and generates QA data to train models for improved long video understanding.

***CogVLM2*** (Hong et al., 2024) proposes a video temporal QA generation pipeline for VLM post-training. For each video, frames are extracted and captioned using a VLM. Only frames with significant scene changes—determined by comparing captions via an LLM—are kept. An LLM is then prompted with few-shot examples to generate time-related QA pairs based on these captions.

***LLaVA-Video*** (Zhang et al., 2024c) focuses on data synthesis for key tasks in video instruction tuning. It adopts a bottom-up approach to generate detailed video descriptions, first processing short intervals and then generalizing to longer segments. An LLM, provided with the task, video descriptions, and few-shot examples, is then used to generate QA pairs.

***CogVideoX*** (Yang et al., 2024) proposes a diffusion model for text-to-video generation, maintaining text-video alignment for long videos. It uses a VLM and an LLM for caption generation: a VLM is prompted with short captions and frame images to generate diverse captions, which are then summarized by an LLM to produce a long video caption.

***VITED*** (Lu et al., 2025) tackles evidence localization and multi-step reasoning in video QA by using a VLM to extract segment-level question-relevant evidence, followed by an LLM with beam search to construct stepwise evidence chains. If the question is answerable using the constructed chain, the datapoint is used in the training stages of event distillation and answer generation.

### 3.9 User Interface & Web Design

***DreamStruct*** (Peng et al., 2024b) builds training data to improve VLMs in image-to-code conversion without initial annotations. It defines a set of principles and seed interface descriptions to prompt a VLM for diverse descriptions, which are then used with the principles to prompt an LLM for HTML code. After minor edits and the creation of internal images, the final output is prepared for production.

***Ferret-UI 2*** (Li et al., 2024f) introduces a multimodal model trained on synthetic data for advanced tasks in universal UI comprehension. It first collects raw screenshots depicting diverse usage scenarios with available bounding boxes. After filtering, a tool extracts bounding box data, which, along with the screenshots and task prompt, is passed to a strong VLM to generate QA data for tasks including functionality description of entire or partial image areas, perception of elements and layouts, and responses to user-interaction scenarios.

***ScreenAI*** (Baechler et al., 2024) employs synthetic data for a novel VLM architecture with core components comprising a vision encoder and an encoder-decoder language model. To enhance pretraining data diversity, it uses LLMs to generate varied tasks. Initially, it applies layout extractors, OCR, captioners, and icon classifiers to obtain a comprehensive screen schema description. This data is then input to the LLM to return general an screen navigation QA pairs, and screen summarization formats.

***WebSight*** (Laurençon et al., 2024) addresses the challenge of converting web screenshots to HTML using VLMs. To build a large-scale image-to-code dataset, it first employs a small LLM to generate diverse content for the desired website design. Then, a strong LLM with robust coding skills produces complete HTML and TailwindCSS code for each concept. A screenshot of the rendered full page is finally captured to fine-tune a VLM on this data.

***ILuvUI*** (Jiang et al., 2025) follows a LLaVA-style data synthesis approach to fine-tune a conversational VLM for user interface QA. Given a UI screenshot, it extracts the bounding boxes of visual elements and passes them to GPT to generate a caption. These are then used to prompt an LLM to produce QA data in single-step conversation and detailed description formats. Unlike LLaVA, it replaces reasoning with five specific tasks: identifying UI element types, listing potential actions on a UI, predicting the outcomes of actions, selecting elements for a goal, and goal-based planning.

***Flame*** (Ge et al., 2025) advances VLMs for frontend development by training on synthetic data. After collecting code components and modifying them to create snippets, these are expanded into complete frontend designs through: (1) Evolutional—an LLM enriches code via random in-depth or in-breadth growth; (2) Waterfall—an LLM defines tasks and generates code sequentially in a structured manner; and (3) Additive—an LLM enriches human-written code while preserving its core idea. Finally, after the LLM adds necessary configurations, checks the code, and generates its description, the code and description, along with the rendered image, are used for training.

### 3.10 Grounding & Personalization

***TWIST & SCOUT*** (Bhowmik et al., 2024) enhance VLMs' spatial visual grounding while retaining prior knowledge by: (1) introducing a twin expert module—frozen for image understanding and learnable for grounding—trained stepwise per question for incremental reasoning; and (2) generating synthetic training data by prompting an LLM with image-caption pairs to create "what" and "where" questions, which are then used with images to prompt a VLM for grounding data generation. Negative samples are also crafted from invalid queries to prevent hallucinations.

***SynGround*** (He et al., 2024) explores synthetic data to improve VLM outputs by grounding them to image regions. It proposes a simple pipeline: starting with an unlabeled image, a VLM generates a detailed description, which is sent to an LLM (to extract object names) and a diffusion model (to generate an accurate image). The resulting text and image are input into an object detector to obtain bounding boxes, forming image-text-box triplets for visual grounding tuning.

***PLVM*** (Pham et al., 2024) tackles the less-explored task of personalization in VLMs by introducing a data synthesis pipeline to compensate for limited datasets. The model is trained to answer a query given a reference image, a query image, and a question, enabling intuitive user-VLM interaction. Data is generated by prompting an LLM with templates for description generation, each of which is then used by a diffusion model to generate images conditioned on a face; outputs with low similarity to the descriptions or reference faces are filtered out. A VLM then answers predefined open- or close-ended questions based on the image and prompt.

***TRIG*** (Li et al., 2025b) tackles visual text grounding in text-rich images by proposing a synthetic data pipeline for training and evaluation. It first extracts OCR data, then uses an LLM—prompted with a question, answer, and box-text pairs—to identify the answer-containing box. If the answer is verifiable from that box, the data is used in two modes: OCR-free, with the input consisting of the question and instruction; and OCR-based, with the box-text pairs added to the prior, both considering the answer and the supporting box as output.

***CaT*** (An et al., 2025) personalizes VLMs using diverse positive and negative synthetic data. Positives are generated by a diffusion model trained on reference data. Negatives use the same model and a GPT-4o-built three-level tree per reference image: root (main object), dimensions (concept variations), and attributes (dimension states). GPT-4o forms image generation prompts via tree operations—altering the root for easy negative samples and modifying dimensions for hard ones.

### 3.11 Long-Context & Dialogue Handling

***GoingBeyondImagination*** (Zhan et al., 2024) enhances dialogue agents using visual descriptions in text format. The real dataset contains dialogues in an A-image-B format, where A and B are speakers and the image belongs to A. To enrich dialog acts and visual cues, synthetic data is generated by prompting an LLM with few-shot examples to produce visual descriptions as image alternatives. The VLM is then trained on entire triplets.

***FIRE*** (Li et al., 2024d) uses dialogue simulation for synthetic data generation in two steps. First, GPT-4V is prompted with an image-question-answer trio to simulate a teacher-student dialogue, where the student reasons over the image to answer, and the teacher provides grounded feedback. Second, two models fine-tuned on this data act as student and teacher, continuing the same setup to generate large-scale synthetic dialogues.

***StableLLaVA*** (Li et al., 2024e) enhances instruction tuning via synthetic image-dialogue data. The pipeline begins with text-only instructions. First, using few-shot demonstrations and a guided prompt, an LLM generates image generation prompts, which are used with a diffusion model to create custom images. Then, the LLM is prompted—again with instructions, the generated image prompt, and few-shot examples—to generate dialogues. The corresponding image-dialogue pairs are mapped and used for instruction tuning.

***Insight-V*** (Dong et al., 2024b) focuses on generating long-chain reasoning data. Given an image-question pair, a reasoner model produces step-by-step reasoning followed by a final summary. Summaries are retained after filtering, while reasoning paths are evaluated by a judge model—only those with high scores are kept. The final reasoning and summary data are then used for VLM training.

***LongWriter-V*** (Tu et al., 2025) tackles the lack of long output training data causing VLMs to generate incoherent long answers by synthesizing data in two folds for training. The first uses image-instruction pairs, filters those suitable for long outputs, and applies two-step prompting: generating a writing plan and completing its subtasks to produce long text. The second follows the same process but with multiple relevant images as input.

### 3.12 Chart Understanding

***ChartLlama*** (Han et al., 2023) points to the lack of high-quality instruction data for chart QA. Its data generation pipeline begins by prompting an LLM with desired concepts to generate tabular raw data and corresponding descriptions. Then, using the raw data and few-shot examples, the LLM produces chart code and descriptions, with the code rendered into chart images. Finally, the LLM uses the chart descriptions and raw data to generate instruction-tuning data, forming chart-question-answer triplets.

***EvoChart*** (Huang et al., 2025) addresses the lack of high-quality data for VLM chart comprehension. Its data synthesis starts with seed chart code generated by an LLM, followed by rendering chart images using a composable chart generator with varied configurations. The generated charts are then assessed, modified if needed, and used to create question-answer training pairs.

***RefChartQA*** (Vogel et al., 2025) addresses challenges in understanding interleaved visual-numerical relations in charts. For synthetic data, each image-question pair is processed either by a fine-tuned VLM for program-of-thought generation and step-by-step answering using chart metadata, or by passing the question and metadata to an LLM for direct answering; data from both paths is used if the generated answer matches the annotation.

***ChartCoder*** (Zhao et al., 2025b) addresses information loss in interpreting charts with textual descriptions. Its chart-to-code pipeline has two steps: first, given input instructions, an LLM generates related keywords and data, then uses them with few-shot examples for chart code generation, retaining only those with correct results; second, a subset of these is sampled, each code is decomposed and enriched by an LLM into a complete step-by-step program with textual guidance, and then concatenated.

***CoF*** (Li et al., 2025c) proposes a pipeline for fine-grained chart-based reasoning for both training and evaluation. Given JSON files with chart information, an LLM first enriches their content for accuracy and diversity. The enriched charts are rendered as images, and function-based programs are created from the enriched JSON. Using data segments as objects and fundamental operations, function chains are built as input-function-output triplets. These chains are used by an LLM to generate rationales, followed by the generation of chain-of-thought question-answer data.

### 3.13 3D Scene Comprehension

***ChatGarment*** (Bian et al., 2024) highlights the difficulty of editing and generating 3D garments using VLMs. It fine-tunes a VLM with instruction-image inputs and JSON outputs, which are rendered into simulated garments reflecting the applied instructions. For training data, a strong VLM generates JSON garment configurations—both part-level and whole-level—based on the provided image.

***MORE3D*** (Jiang et al., 2024) targets 3D design reasoning by constructing a simple pipeline that prompts GPT-4o with specific requirements, along with two images—one showing the current 3D scene and the other the ground truth—to generate question-answer pairs, where the answer is a rationale aimed at achieving the instruction goal.

***SpatialVLM*** (Chen et al., 2024a) enhances 3D spatial reasoning in VLMs. It begins with unlabeled images, applies semantic filtering to keep scene-level ones, and extracts 2D information such as region captions and object-level physical data. Captions are used to generate unambiguous questions, with answers derived from 3D bounding boxes obtained from the object-level information, forming synthetic image-question-answer training data.

***PiSA*** (Guo et al., 2025) tackles the low quality of 3D datasets through a data generation engine. It first enriches answers for image-question-answer triplets using a 3D multimodal model, then refines them with 2D VLMs by removing incorrect content based on similar provided 2D images. Finally, an iterative bootstrapping process injects more precise 3D knowledge for the given 3D input and question.

***3UR-LLM*** (Xiong et al., 2025) proposes a data generation pipeline for 3D scene understanding. It first prompts for descriptions of individual room parts based on images, then uses an LLM to generate a summary of the entire scene by concatenating these descriptions. Finally, it constructs and refines QA pairs for training purposes.

### 3.14 Image Captioning

***BLIP3-KALE*** (Awadalla et al., 2024) bridges synthetic captions and factual web alt-text. It starts with image-text pairs using the web alt-text, generates a synthetic caption via a VLM, and then refines it into a knowledge-augmented caption using an LLM fed with both the caption and the alt-text. To scale, a small VLM is trained on image-text inputs and their corresponding knowledge-augmented captions as outputs, and is then used to produce such captions for additional image-text pairs.

***VILA***$^2$ (Fang et al., 2024) introduces a method by which an open-source VLM can improve itself via synthetic data enhancement without relying on proprietary models. First, it uses a self-augmenting loop for pretraining, which trains the VLM with initial short image captions, then asks the trained model to generate a longer caption for each, and augments the text with real data to form new image-caption pairs, continuing this cycle. Secondly, it applies specialist augmenting by fine-tuning the model on small datasets for tasks such as spatial reasoning, grounding, and OCR. These specialists generate task-specific QA pairs, which are appended to the original data to form an enriched dataset for retraining the generalist VLM.

***CAPTURE*** (Dong et al., 2024a) proposes an image captioning evaluation metric and employs it to enhance caption quality for VLM training via a five-stage process using only open-source vision and language tools. A VLM first produces an overall caption as a skeleton. Next, it refines candidate regions to extract salient visual elements, followed by generating local captions to enrich the main caption and reduce hallucinations. A hallucination filtering step removes unreliable content, and finally, an LLM integrates the validated local captions into a high-quality, hallucination-suppressed description.

***SynthVLM*** (Liu et al., 2024f) builds a large image-caption dataset for VLM training. From the initial pairs, it filters low-quality captions and selects correlated pairs using a CLIP-based score. A diffusion model generates a high-quality image for each caption to replace the original. Final filtering combines the CLIP score (on resized images) with another metric assessing quality preservation after resizing.

### 3.15 Multiculturalism & Multilingualism

***X-LLaVA*** (Shin et al., 2024) builds a multilingual, multimodal instruction-tuning dataset. It first collects images with a specific number of main subjects, then uses GPT-4V to generate QA pairs in English, Korean, and Chinese. The question types include object-centric (focusing on main object details), location-centric (based on object positions via scene graphs), atmosphere-centric (capturing the holistic interplay of objects), and conversational (providing a deeper, image-level understanding).

***CultureVLM*** (Liu et al., 2025) enhances VLMs' cultural understanding by fine-tuning the model on a dataset created using the help of GPT-4o, for tasks like extracting cultural elements from Wikipedia, judging their relevance to specific classes and countries, generating detailed cultural concept introductions, and—most importantly—creating diverse question types and reasoning responses.

***MixCube*** (Kim et al., 2025) introduces a crosscultural evaluation benchmark. It gathers culturally identifiable web images and uses a diffusion model to inpaint them by altering the depicted ethnicities, retaining those with high CLIP similarity to the originals. The goal is to identify both the revealed culture and the cultural entity from an image. This dataset evaluates VLMs' cultural abilities by testing their resilience to injected cultural biases.

## 4 Future Directions

While extensive efforts have been made to bridge the contents of VLMs and synthetic data, several unexplored paths still offer great promise, and we outline a few ones that could spark future research.

**Customized Data Synthesis** Data for or by VLMs is often generated with fixed goals and limited customization. Training an open-source model for multimodal data or task generation with customizable properties—similar to (Nayak et al., 2024) for LLMs—would be highly valuable.

**In-Context Learning** As a training-free approach, it offers a valuable way to capture patterns via in-context demonstrations. Future work may focus on generating data as effective in-context examples and refining task prompts to ensure the examples instill better task understanding.

**Autonomous Driving** These systems need real-time understanding of dynamic environments, making them well-suited for robust VLMs. Due to the difficulty of collecting quality data, future research may focus on generating synthetic scenarios or using them to tune or evaluate VLMs.

**Video-Level Analytics** Many applications require reporting specific actions in videos. These can be addressed by methods that track various visual elements across frames to analyze their behavior and actions throughout the video.

**Dynamic Evaluation** Similar to approaches for LLMs (Kim et al., 2024), a multi-turn dynamic evaluation setup for VLMs can help avoid issues like contamination while simultaneously improving the model through synthetic data generated within the evaluation process.

**Low-Resource Settings** VLMs mainly advance in high-resource domains, while many regions lack labeled multimodal data. Synthetic pipelines can bridge this gap by generating data in underrepresented languages, cultures, and applications for further processing.

**Data Efficiency** With growing model sizes, maximizing data efficiency is crucial. While synthetic data is often generated at scale, applying quality and diversity measures can reduce data volume for better cost and resource use.

## 5 Conclusion

Vision-language models (VLMs) and synthetic data are becoming increasingly interdependent, driving significant advancements in multimodal AI. This paper presents the first comprehensive survey bridging VLMs and synthetic data. We explore how data generated using generative AI models can benefit VLMs, as well as how VLMs are involved in the synthetic data generation process. Through an extensive review of prior work, we survey the proposed methodologies and suggest potential directions for future research. Our survey establishes a structured foundation for bridging the concepts of VLMs and synthetic data, enabling customized solutions tailored to specific needs and objectives.

## References

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. 2025. Vision-language models do not understand negation. *arXiv preprint arXiv:2501.09425*.

Ruichuan An, Kai Zeng, Ming Lu, Sihan Yang, Renrui Zhang, Huitong Ji, Qizhe Zhang, Yulin Luo, Hao Liang, and Wentao Zhang. 2025. Concept-as-tree: Synthetic data is all you need for vlm personalization. *arXiv preprint arXiv:2503.12999*.

Anas Awadalla, Le Xue, Manli Shu, An Yan, Jun Wang, Senthil Purushwalkam, Sheng Shen, Hannah Lee, Oscar Lo, Jae Sung Park, et al. 2024. Blip3-kale: Knowledge augmented large-scale dense captions. *arXiv preprint arXiv:2411.07461*.

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. 2024. Vismin: Visual minimal-change understanding. *arXiv preprint arXiv:2407.16772*.

Gilles Baechler, Srinivas Sunkara, Maria Wang, Fedir Zubach, Hassan Mansoor, Vincent Etter, Victor Cărbune, Jason Lin, Jindong Chen, and Abhanshu Sharma. 2024. Screenai: A vision-language model for ui and infographics understanding. *arXiv preprint arXiv:2402.04615*.

Sami Baral, Li Lucy, Ryan Knight, Alice Ng, Luca Soldaini, Neil T Heffernan, and Kyle Lo. 2025. Drawedumath: Evaluating vision language models with expert-annotated students' hand-drawn math images. *arXiv preprint arXiv:2501.14877*.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. 2023. Mitigating open-vocabulary caption hallucinations. *arXiv preprint arXiv:2312.03631*.

Aritra Bhowmik, Mohammad Mahdi Derakhshani, Dennis Koelma, Martin R Oswald, Yuki M Asano, and Cees GM Snoek. 2024. Learning to ground vlms without forgetting. *arXiv preprint arXiv:2410.10491*.

Siyuan Bian, Chenghao Xu, Yuliang Xiu, Artur Grigorev, Zhen Liu, Cewu Lu, Michael J Black, and Yao Feng. 2024. Chatgarment: Garment estimation, generation and editing via large language models. *arXiv preprint arXiv:2412.17811*.

Dexian Cai, Xiaocui Yang, Yongkang Liu, Daling Wang, Shi Feng, Yifei Zhang, and Soujanya Poria. 2025. Pixel-level reasoning segmentation via multi-turn conversations. *arXiv preprint arXiv:2502.09447*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024b. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024c. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*.

Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. 2024. From the least to the most: Building a plug-and-play visual reasoner via data synthesis. *arXiv preprint arXiv:2406.19934*.

Hejie Cui, Lingjun Mao, Xin Liang, Jieyu Zhang, Hui Ren, Quanzheng Li, Xiang Li, and Carl Yang. 2024. Biomedical visual instruction tuning with clinician preference alignment. *arXiv preprint arXiv:2406.13173*.

Muzhi Dai, Jiashuo Sun, Zhiyuan Zhao, Shixuan Liu, Rui Li, Junyu Gao, and Xuelong Li. 2025. From captions to rewards (carevl): Leveraging large language model experts for enhanced reward modeling in large vision-language models. *arXiv preprint arXiv:2503.06260*.

Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, et al. 2024a. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv preprint arXiv:2410.17885*.

Xueqing Deng, Qihang Yu, Ali Athar, Chenglin Yang, Linjie Yang, Xiaojie Jin, Xiaohui Shen, and Liang-Chieh Chen. 2025. Coconut-pancap: Joint panoptic segmentation and grounded captions for fine-grained understanding and generation. *arXiv preprint arXiv:2502.02589*.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Y Zou, Kai-Wei Chang, and Wei Wang. 2024b. Enhancing large vision language models with self-training on image comprehension. *Advances in Neural Information Processing Systems*, 37:131369–131397.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. 2024a. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2024b. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*.

Yifan Du, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, Mingchen Cai, Ruihua Song, and Ji-Rong Wen. 2023. What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning. *arXiv preprint arXiv:2311.01487*.

Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay Krishna. 2024. Manipulate-anything: Automating real-world robots using vision-language models. *arXiv preprint arXiv:2406.18915*.

Yunhao Fang, Ligeng Zhu, Yao Lu, Yan Wang, Pavlo Molchanov, Jan Kautz, Jang Hyun Cho, Marco Pavone, Song Han, and Hongxu Yin. 2024. Vila 2: Vila augmented vila. *arXiv preprint arXiv:2407.17453*.

Deqing Fu, Tong Xiao, Rui Wang, Wang Zhu, Pengchuan Zhang, Guan Pang, Robin Jia, and Lawrence Chen. 2024. Tldr: Token-level detective reward model for large vision language models. *arXiv preprint arXiv:2410.04734*.

Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaojian Ma, Tao Yuan, Yue Fan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. 2024. Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage. *arXiv preprint arXiv:2412.15606*.

Tong Ge, Yashu Liu, Jieping Ye, Tianyi Li, and Chao Wang. 2025. Advancing vision-language models in front-end development via data synthesis. *arXiv preprint arXiv:2503.01619*.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*.

Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. 2024. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13806.

Zilu Guo, Hongbin Lin, Zhihao Yuan, Chaoda Zheng, Pengshuo Qiu, Dongzhi Jiang, Renrui Zhang, Chun-Mei Feng, and Zhen Li. 2025. Pisa: A self-augmented data engine and training strategy for 3d understanding with large models. *arXiv preprint arXiv:2503.10529*.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*.

Ruozhen He, Ziyan Yang, Paola Cascante-Bonilla, Alexander C Berg, and Vicente Ordonez. 2024. Learning from synthetic data for visual grounding. *arXiv preprint arXiv:2403.13804*.

Ahmed Heakl, Abdullah Sohail, Mukul Ranjan, Rania Hossam, Ghazi Ahmed, Mohamed El-Geish, Omar Maher, Zhiqiang Shen, Fahad Khan, and Salman Khan. 2025. Kitab-bench: A comprehensive multi-domain benchmark for arabic ocr and document understanding. *arXiv preprint arXiv:2502.14949*.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301*.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2024a. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9590–9601.

Zhe Hu, Jing Li, and Yu Yin. 2025. When words outperform vision: Vlms can self-improve via text-only training for human-centered decision making. *arXiv preprint arXiv:2503.16965*.

Zhe Hu, Yixiao Ren, Jing Li, and Yu Yin. 2024b. Viva: A benchmark for vision-grounded decision-making with human values. *arXiv preprint arXiv:2407.03000*.

Irene Huang, Wei Lin, Muhammad Jehanzeb Mirza, Jacob Hansen, Sivan Doveh, Victor Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuehne, Trevor Darrell, et al. 2024a. Conme: Rethinking evaluation of compositional reasoning for modern vlms. *Advances in Neural Information Processing Systems*, 37:22927–22946.

Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. 2025. Evochart: A benchmark and a self-training approach towards real-world chart understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3680–3688.

Zhipeng Huang, Zhizheng Zhang, Zheng-Jun Zha, Yan Lu, and Baining Guo. 2024b. Relationvlm: Making large vision-language models understand visual relations. *arXiv preprint arXiv:2403.12801*.

Donggon Jang, Yucheol Cho, Suin Lee, Taehyeon Kim, and Dae-Shik Kim. 2025. Mmr: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation. *arXiv preprint arXiv:2503.13881*.

Yuanfeng Ji, Chongjian Ge, Weikai Kong, Enze Xie, Zhengying Liu, Zhengguo Li, and Ping Luo. 2023. Large language models as automated aligners for benchmarking vision-language models. *arXiv preprint arXiv:2311.14580*.

Xueying Jiang, Lewei Lu, Ling Shao, and Shijian Lu. 2024. Multimodal 3d reasoning segmentation with complex scenes. *arXiv preprint arXiv:2411.13927*.

Yue Jiang, Eldon Schoop, Amanda Swearngin, and Jeffrey Nichols. 2025. Iluvui: Instruction-tuned language-vision modeling of uis from machine conversations. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 861–877.

Eunsu Kim, Juyoung Suk, Seungone Kim, Niklas Muennighoff, Dongkwan Kim, and Alice Oh. 2024. Llm-as-an-interviewer: Beyond static testing through dynamic llm evaluation. *arXiv preprint arXiv:2412.10424*.

Jun Seong Kim, Kyaw Ye Thu, Javad Ismayilzada, Junyeong Park, Eunsu Kim, Huzama Ahmad, Na Min An, James Thorne, and Alice Oh. 2025. When tom eats kimchi: Evaluating cultural bias of multimodal large language models in cultural mixture contexts. *arXiv preprint arXiv:2503.16826*.

Pranavi Kolouju, Eric Xing, Robert Pless, Nathan Jacobs, and Abby Stylianou. 2025. good4cir: Generating detailed synthetic captions for composed image retrieval. *arXiv preprint arXiv:2503.17871*.

Hugo Laurençon, Léo Tronchon, and Victor Sanh. 2024. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029.*

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11286–11315.

Andrew Li, Rahul Thapa, Rahul Chalamala, Qingyang Wu, Kezhen Chen, and James Zou. 2025a. Smir: Efficient synthetic data pipeline to improve multi-image reasoning. *arXiv preprint arXiv:2501.03675.*

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Guanzhen Li, Yuxi Xie, and Min-Yen Kan. 2024a. Mvp-bench: Can large vision–language models conduct multi-level visual perception like humans? *arXiv preprint arXiv:2410.04345.*

Haoxin Li and Boyang Li. 2025. Enhancing vision-language compositional understanding with multimodal synthetic data. *arXiv preprint arXiv:2503.01167.*

Huayang Li, Siheng Li, Deng Cai, Longyue Wang, Lemao Liu, Taro Watanabe, Yujiu Yang, and Shuming Shi. 2023b. Textbind: Multi-turn interleaved multimodal instruction-following in the wild. *arXiv preprint arXiv:2309.08637.*

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. 2024b. Vlrewardbench: A challenging benchmark for vision-language generative reward models. *arXiv preprint arXiv:2411.17451.*

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. 2024c. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. *arXiv preprint arXiv:2410.09421.*

Ming Li, Ruiyi Zhang, Jian Chen, Jiuxiang Gu, Yufan Zhou, Franck Dernoncourt, Wanrong Zhu, Tianyi Zhou, and Tong Sun. 2025b. Towards visual text grounding of multimodal large language model. *arXiv preprint arXiv:2504.04974.*

Pengxiang Li, Zhi Gao, Bofei Zhang, Tao Yuan, Yuwei Wu, Mehrtash Harandi, Yunde Jia, Song-Chun Zhu, and Qing Li. 2024d. Fire: A dataset for feedback integration and refinement evaluation of multimodal models. *arXiv preprint arXiv:2407.11522.*

Yanda Li, Chi Zhang, Gang Yu, Wanqi Yang, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. 2024e. Enhanced visual instruction tuning with synthesized image-dialogue data. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14512–14531.

Zhangheng Li, Keen You, Haotian Zhang, Di Feng, Harsh Agrawal, Xiujun Li, Mohana Prasad Sathya Moorthy, Jeff Nichols, Yinfei Yang, and Zhe Gan. 2024f. Ferret-ui 2: Mastering universal user interface understanding across platforms. *arXiv preprint arXiv:2410.18967.*

Zijian Li, Jingjing Fu, Lei Song, Jiang Bian, Jun Zhang, and Rui Wang. 2025c. Chain of functions: A programmatic pipeline for fine-grained chart reasoning data. *arXiv preprint arXiv:2503.16260.*

Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. 2025d. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189.*

Tuo Liang, Zhe Hu, Jing Li, Hao Zhang, Yiren Lu, Yunlai Zhou, Yiran Qiao, Disheng Liu, Jeirui Peng, Jing Ma, et al. 2025a. When'yes'meets' but': Can large models comprehend contradictory humor through comparative reasoning? *arXiv preprint arXiv:2503.23137.*

Xiwen Liang, Min Lin, Weiqi Ruan, Rongtao Xu, Yuecheng Liu, Jiaqi Chen, Bingqian Lin, Yuzheng Zhuang, and Xiaodan Liang. 2025b. Structured preference optimization for vision-language long-horizon task planning. *arXiv preprint arXiv:2502.20742.*

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *CoRR*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253.*

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

Jihao Liu, Xin Huang, Jinliang Zheng, Boxiao Liu, Jia Wang, Osamu Yoshie, Yu Liu, and Hongsheng Li. 2024b. Mm-instruct: Generated visual instructions for large multimodal model alignment. *arXiv preprint arXiv:2406.19736.*

Li Liu, Diji Yang, Sijia Zhong, Kalyana Suma Sree Tholeti, Lei Ding, Yi Zhang, and Leilani Gilpin. 2024c. Right this way: Can vlms guide us to see more to answer questions? *Advances in Neural Information Processing Systems*, 37:132946–132976.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023c. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. 2024d. Best practices and lessons learned on synthetic data. *arXiv preprint arXiv:2404.07503*.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*.

Siyuan Liu, Jiawei Du, Sicheng Xiang, Zibo Wang, and Dingsheng Luo. 2024e. Relep: A novel framework for real-world long-horizon embodied planning. *arXiv preprint arXiv:2409.15658*.

Zheng Liu, Hao Liang, Xijie Huang, Wentao Xiong, Qinhan Yu, Linzhuang Sun, Chong Chen, Conghui He, Bin Cui, and Wentao Zhang. 2024f. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Yujie Lu, Yale Song, William Wang, Lorenzo Torresani, and Tushar Nagarajan. 2025. Vited: Video temporal evidence distillation. *arXiv preprint arXiv:2503.12855*.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. 2025. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*.

Ji Ma, Wei Suo, Peng Wang, and Yanning Zhang. 2024a. C3l: Content correlated vision-language instruction tuning data generation via contrastive learning. *arXiv preprint arXiv:2405.12752*.

Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. 2024b. Visa: Retrieval augmented generation with visual source attribution. *arXiv preprint arXiv:2412.14457*.

Zixian Ma, Jianguo Zhang, Zhiwei Liu, Jieyu Zhang, Juntao Tan, Manli Shu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Caiming Xiong, et al. Taco: Learning multi-modal models to reason and act with synthetic chains-of-thought-and-action. In *Workshop on Reasoning and Planning for Large Language Models*.

Samarth Mishra, Kate Saenko, and Venkatesh Saligrama. 2025. Enhancing compositional reasoning in vision-language models with synthetic preference data. *arXiv preprint arXiv:2504.04740*.

Nihal V Nayak, Yiyang Nan, Avi Trost, and Stephen H Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. *arXiv preprint arXiv:2402.18334*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Wujian Peng, Lingchen Meng, Yitong Chen, Yiweng Xie, Yang Liu, Tao Gui, Hang Xu, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. 2024a. Inst-it: Boosting multimodal instance understanding via explicit visual prompt instruction tuning. *arXiv preprint arXiv:2412.03565*.

Yi-Hao Peng, Faria Huq, Yue Jiang, Jason Wu, Xin Yue Li, Jeffrey P Bigham, and Amy Pavel. 2024b. Dreamstruct: Understanding slides and user interfaces via synthetic data generation. In *European Conference on Computer Vision*, pages 466–485. Springer.

Chau Pham, Hoang Phan, David Doermann, and Yunjie Tian. 2024. Personalized large vision-language models. *arXiv preprint arXiv:2412.17610*.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. 2024a. Cogcom: Train large vision-language models diving into details through chain of manipulations. *arXiv preprint arXiv:2402.04236*.

Yayun Qi, Hongxi Li, Yiqi Song, Xinxiao Wu, and Jiebo Luo. 2024b. How vision-language tasks benefit from large pre-trained models: A survey. *arXiv preprint arXiv:2412.08158*.

Han Qiu, Jiaxing Huang, Peng Gao, Qin Qi, Xiaoqin Zhang, Ling Shao, and Shijian Lu. 2024. Longhalqa: Long-context hallucination evaluation for multimodal large language models. *arXiv preprint arXiv:2410.09962*.

Tianyuan Qu, Longxiang Tang, Bohao Peng, Senqiao Yang, Bei Yu, and Jiaya Jia. 2025. Does your vision-language model get lost in the long video sampling dilemma? *arXiv preprint arXiv:2503.12496*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. 2023. Visual chain of thought: bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*.

Dongjae Shin, HyeonSeok Lim, Inho Won, Changsu Choi, Minjun Kim, Seungwoo Song, Hangyeol Yoo, Sangmin Kim, and Kyungtae Lim. 2024. X-llava: Optimizing bilingual large vision-language alignment. *arXiv preprint arXiv:2403.11399*.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2024. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*.

Xin Su, Man Luo, Kris W Pan, Tien Pei Chou, Vasudev Lal, and Phillip Howard. 2024. Sk-vqa: Synthetic knowledge generation at scale for training context-augmented multimodal llms. *arXiv preprint arXiv:2406.19593*.

Guohao Sun, Can Qin, Huazhu Fu, Linwei Wang, and Zhiqiang Tao. 2024. Stllava-med: Self-training large language and vision assistant for medical question-answering. *arXiv preprint arXiv:2406.19973*.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. 2025. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. *arXiv preprint arXiv:2502.13383*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.

Shangqing Tu, Yucheng Wang, Daniel Zhang-Li, Yushi Bai, Jifan Yu, Yuhao Wu, Lei Hou, Huiqin Liu, Zhiyuan Liu, Bin Xu, et al. 2025. Longwriter-v: Enabling ultra-long and high-fidelity generation in vision-language models. *arXiv preprint arXiv:2502.14834*.

Alexander Vogel, Omar Moured, Yufan Chen, Jiaming Zhang, and Rainer Stiefelhagen. 2025. Refchartqa: Grounding visual answer on chart images through instruction tuning. *arXiv preprint arXiv:2503.23131*.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. 2024a. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5309–5317.

Guankun Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, et al. 2025. Endochat: Grounded multimodal large language model for endoscopic surgery. *arXiv preprint arXiv:2501.11347*.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*.

Ke Wang, Jiahui Zhu, Minjie Ren, Zeming Liu, Shiwei Li, Zongye Zhang, Chenkai Zhang, Xiaoyu Wu, Qiqi Zhan, Qingjie Liu, et al. 2024b. A survey on data synthesis and augmentation for large language models. *arXiv preprint arXiv:2410.12896*.

Liqiong Wang, Teng Jin, Jinyu Yang, Ales Leonardis, Fangyi Wang, and Feng Zheng. 2024c. Agri-llava: Knowledge-infused large multimodal assistant on agricultural pests and diseases. *arXiv preprint arXiv:2412.02158*.

Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. 2024d. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. *arXiv preprint arXiv:2406.14194*.

Wensheng Wang and Ning Tan. 2025. Hybridgen: Vlm-guided hybrid planning for scalable data generation of imitation learning. *arXiv preprint arXiv:2503.13171*.

Yuhao Wang, Zhiyuan Zhu, Heyang Liu, Yusheng Liao, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024e. Drawing the line: Enhancing trustworthiness of mllms through the power of refusal. *arXiv preprint arXiv:2412.11196*.

Ian Wu, Sravan Jayanthi, Vijay Viswanathan, Simon Rosenberg, Sina Pakazad, Tongshuang Wu, and Graham Neubig. 2024a. Synthetic multimodal question generation. *arXiv preprint arXiv:2407.02233*.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. 2025. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*.

Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, et al. 2024b. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. *arXiv preprint arXiv:2406.10900*.

Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. 2023. Embodied task planning with large language models. *arXiv preprint arXiv:2307.01848*.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv preprint arXiv:2411.02712*.

Haomiao Xiong, Yunzhi Zhuge, Jiawen Zhu, Lu Zhang, and Huchuan Lu. 2025. 3ur-llm: An end-to-end multimodal large language model for 3d scene understanding. *arXiv preprint arXiv:2501.07819*.

Zhe Xu, Daoyuan Chen, Zhenqing Ling, Yaliang Li, and Ying Shen. 2025. Mindgym: Enhancing vision-language models via synthetic self-challenging questions. *arXiv preprint arXiv:2503.09499*.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, Di Jin, Yu Cheng, Qifan Wang, and Lifu Huang. 2024. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. *arXiv preprint arXiv:2402.11690*.

Yue Yang, Ajay Patel, Matt Deitke, Tanmay Gupta, Luca Weihs, Andrew Head, Mark Yatskar, Chris Callison-Burch, Ranjay Krishna, Aniruddha Kembhavi, et al. 2025. Scaling text-rich image understanding via code-guided synthetic multimodal data generation. *arXiv preprint arXiv:2502.14846*.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.

Jingwei Yi, Junhao Yin, Ju Xu, Peng Bao, Yongliang Wang, Wei Fan, and Hao Wang. 2025. Imageref-vl: Enabling contextual image referencing in vision-language models. *arXiv preprint arXiv:2501.12418*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Xiaomin Yu, Pengxiang Ding, Wenjie Zhang, Siteng Huang, Songyang Gao, Chengwei Qin, Kejian Wu, Zhaoxin Fan, Ziyue Qiao, and Donglin Wang. 2025. Unicorn: Text-only data synthesis for vision language model training. *arXiv preprint arXiv:2503.22655*.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. 2024. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li. 2023. Halle-control: controlling object hallucination in large multimodal models. *arXiv preprint arXiv:2310.01779*.

Haolan Zhan, Sameen Maruf, Ingrid Zukerman, and Gholamreza Haffari. 2024. Going beyond imagination! enhancing multi-modal dialogue agents with synthetic visual descriptions. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 420–427.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024a. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, et al. 2024b. Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model. *arXiv preprint arXiv:2407.07053*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024c. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Henry Hengyuan Zhao, Pan Zhou, and Mike Zheng Shou. 2024. Genixer: Empowering multimodal large language model as a powerful data generator. In *European Conference on Computer Vision*, pages 129–147. Springer.

Xiangyu Zhao, Shengyuan Ding, Zicheng Zhang, Haian Huang, Maosong Cao, Weiyun Wang, Jiaqi Wang, Xinyu Fang, Wenhai Wang, Guangtao Zhai, et al. 2025a. Omnialign-v: Towards enhanced alignment of mllms with human preference. *arXiv preprint arXiv:2502.18411*.

Xuanle Zhao, Xianzhen Luo, Qi Shi, Chi Chen, Shuo Wang, Wanxiang Che, Zhiyuan Liu, and Maosong Sun. 2025b. Chartcoder: Advancing multimodal large language model for chart-to-code generation. *arXiv preprint arXiv:2501.06598*.

Wenliang Zhong, Weizhi An, Feng Jiang, Hehuan Ma, Yuzhi Guo, and Junzhou Huang. 2024. Compositional image retrieval via instruction-aware contrastive learning. *arXiv preprint arXiv:2412.05756*.

Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. 2024a. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, pages 260–278. Springer.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. 2024b. Megapairs: Massive data synthesis for universal multimodal retrieval. *arXiv preprint arXiv:2412.14475*.

Yiyang Zhou, Zhaoyang Wang, Tianle Wang, Shangyu Xing, Peng Xia, Bo Li, Kaiyuan Zheng, Zijian Zhang, Zhaorun Chen, Wenhao Zheng, et al. Anyprefer: An automatic framework for preference data synthesis. In *The Thirteenth International Conference on Learning Representations*.

Meng Ziyang, Yu Dai, Zezheng Gong, Shaoxiong Guo, Minglong Tang, and Tongquan Wei. 2024. Vga: Vision gui assistant-minimizing hallucinations through image-centric fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1261–1279.