
SURVEY · ARXIV:2506.04788 · READING GROUP

Towards LLM-Centric Multimodal Fusion

A survey on integration strategies and techniques

Jisu An · Junseok Lee
Seoul National University

Jeoungeun Lee
UC San Diego

Yongseok Son
Chung-Ang University

MOTIVATION

Everyone builds MLLMs the same way. Nobody agrees on the wiring.

The recipe is settled: take a **frozen pre-trained LLM**, bolt on a **frozen modality encoder**, and train something small in between.

That "something in between" is where all the design decisions live — and it is the least systematically documented part of the literature.

Existing surveys catalogue components, datasets and benchmarks. Few of them ask **why** a given module was placed where it was.

This paper reads 125 MLLMs (2021–2025) through one lens: how does a non-text signal get into the language embedding space?

125

MLLMs surveyed

2021–25

publication window

3

classification axes

The same component, two different intentions

Prior taxonomies classify by what the module is. But an MLP is an MLP — its role depends entirely on what the authors wanted from it.

MLP · LLaVA-1.5

Project features into the LLM's embedding space

Purpose is alignment. Token count is unchanged; the geometry of the space is what moves.

MLP · Wang et al. 2024

Reduce the number of image tokens

Purpose is compression. Same layer type, an entirely different problem being solved.

Classify by contextual function, not by layer type — otherwise every taxonomy collapses the moment a component does double duty.

Three orthogonal axes

01

Architecture

How, and where, the signal enters

Abstraction

Projection

Semantic embedding

Cross-attention

× fusion level:
early / intermediate / hybrid

02

Representation

What shape the shared space takes

Joint

Coordinated

Hybrid

Fusion inside one network, or
alignment between two?

03

Training

In how many passes, against what
loss

Single-stage

Two-stage

Multi-stage

× objective: LM, contrastive,
reconstruction, DICE

Orthogonal by design — any mechanism can be applied at any level, under any training regime.

BACKGROUND

Anatomy of an MLLM

encoder

Frozen, modality-specific

ViT / CNN for images, Wav2Vec 2.0 for audio. CLIP encoders arrive already language-aligned; ImageBind pushes toward one space for everything.

connector

The only part that really trains

A few million parameters carrying the entire burden of translating one representation into another. This is the survey's subject.

llm

Frozen, or lightly tuned

LLaMA, Vicuna and friends used as reasoning engines. Kept frozen to preserve pretrained knowledge and to keep training affordable.

AXIS 01

Architectural strategies for modality fusion

Four contextual mechanisms — abstraction, projection, semantic embedding, cross-attention — crossed with three fusion levels.

Q-FORMER

Abstraction and semantics, depending on the paper

LLAVA

Projection at its most minimal, fused early

FLAMINGO

Resampler plus gated cross-attention, fused mid-stack

Four things a connector can be asked to do

Abstraction

Control how many tokens arrive

Projection

Move them into the right space

Semantic embedding

Add high-level meaning before entry

Cross-attention

Let the LLM reach out and query

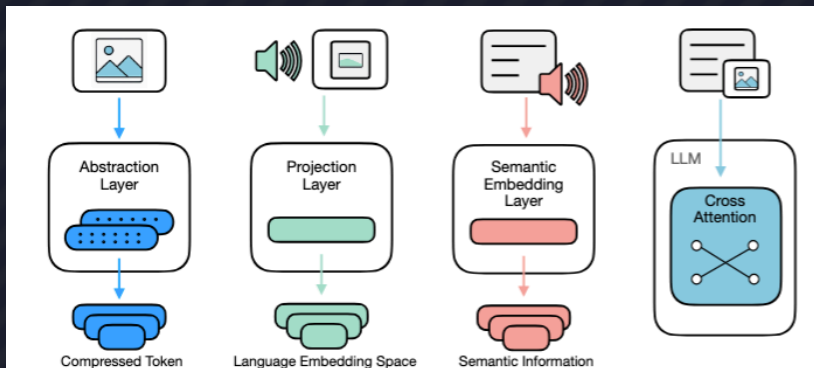


Figure 1: Proposed contextual fusion mechanisms

One component can occupy more than one box — a Q-Former abstracts in BLIP-2 and carries semantics in InstructBLIP.

Abstraction — an information bottleneck

A high-resolution ViT emits hundreds of patch tokens, and the count varies with input size. Abstraction fixes the length and pays down the cost.

Perceiver Resampler

Fixed-length learnable queries
cross-attend to the features

Flamingo, Kosmos-1, mPLUG-Owl2,
Idefics2

Q-Former

Same idea, but self- and cross-
attention layers alternate

BLIP-2, Video-LLaMA, VideoChat

C-Abstractor / convolution

Answers query-based
information loss; keeps spatial
locality

MM1, Honeybee, SEAL

Trade-off: every token you delete is context the LLM will never see.

The Q-Former

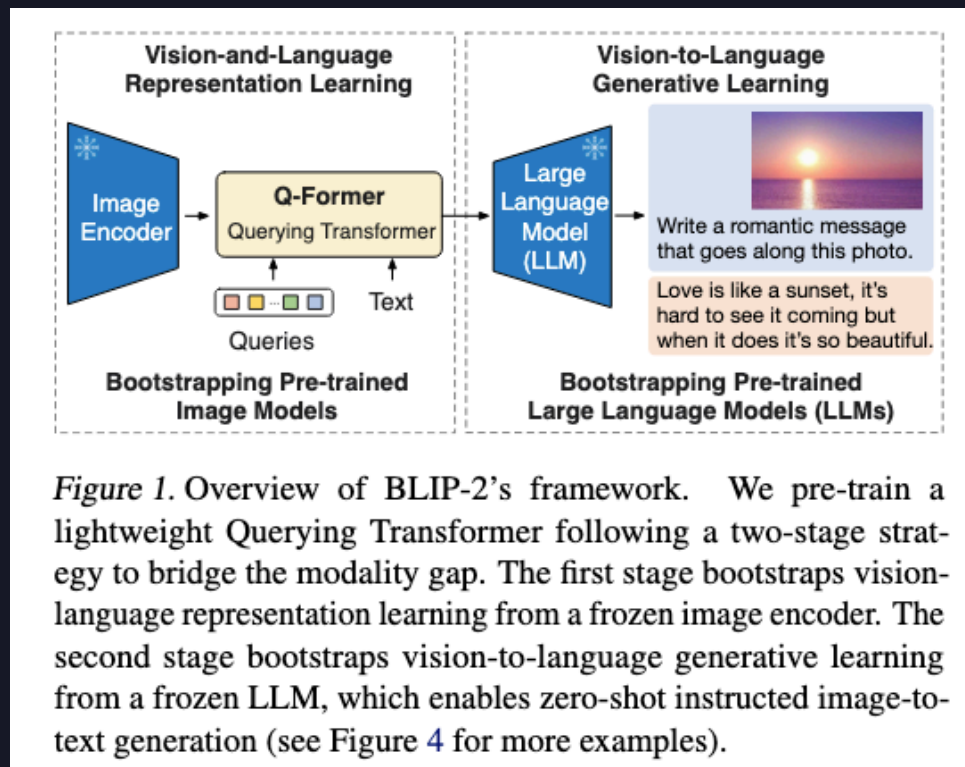


Figure 1. Overview of BLIP-2's framework. We pre-train a lightweight Querying Transformer following a two-stage strategy to bridge the modality gap. The first stage bootstraps vision-language representation learning from a frozen image encoder. The second stage bootstraps vision-to-language generative learning from a frozen LLM, which enables zero-shot instructed image-to-text generation (see Figure 4 for more examples).

Projection — the cheapest thing that works

Map encoder features into the language embedding space, leave the LLM's weights alone.
Cheap, knowledge-preserving, and consequently everywhere.

Linear

Minimum resources, minimum capacity

LLaVA · VideoChat · SEED

MLP

LLaVA-1.5's one-line change: linear → MLP lifted performance outright

LLaVA-1.5 · CogVLM · LISA

Transformer

Capacity to spare when the modality gap is wide

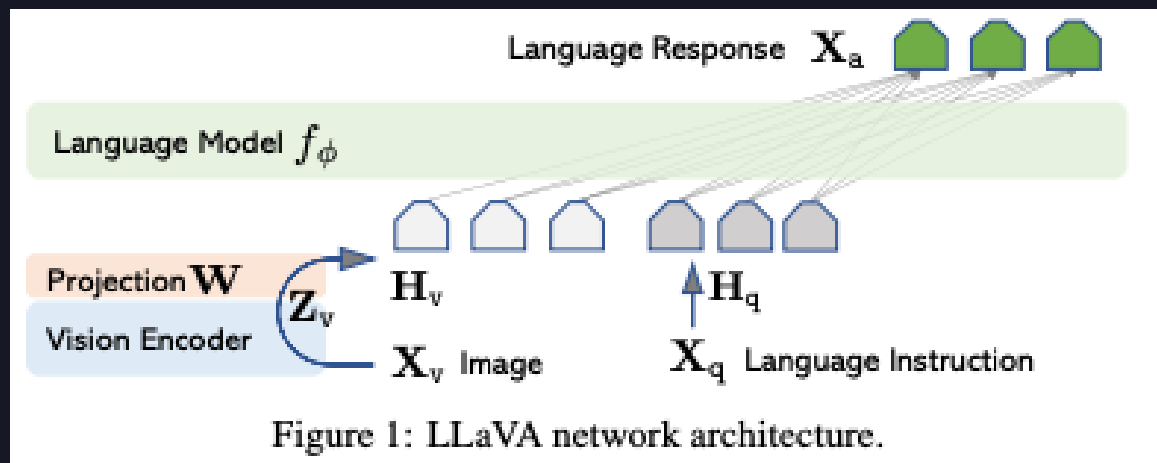
VILA · ManipLLM · PMC-VQA

Q-Former

Instruction-aware projection — the double-duty case again

X-InstructBLIP

The LLaVA architecture



Semantic embedding — adding meaning, not just format

Abstraction shrinks; projection re-bases. Semantic embedding does something else: it injects high-level information into the features before they ever reach the LLM.

learnable queries as instructions

Q-Former queries decide which semantics get extracted

InstructBLIP goes further and feeds the text instruction into the Q-Former alongside the queries, so the visual summary is conditioned on the question being asked.

BLIP-2 · InstructBLIP · BLIVA · MA-LMM

structural priors

Convolution to put spatial bias back in

Cambrian-1 adds spatial inductive bias with learnable queries plus convolution-like cross-attention; AnomalyGPT uses plain convolution for the same end.

Cambrian-1 · AnomalyGPT · LHRs-Bot

Cross-attention — let the LLM do the reaching

Instead of handing the LLM a fixed set of tokens, give it the ability to query the modality dynamically at every generation step.

Reuse the internal layers

Transformer LLMs already contain attention machinery — feed non-text features straight into it.

Flamingo pipes embeddings into gated internal cross-attention.

CogVLM widens the internal Q/K/V matrices for vision.

mPLUG-Owl2 generates modality-specific keys and values.

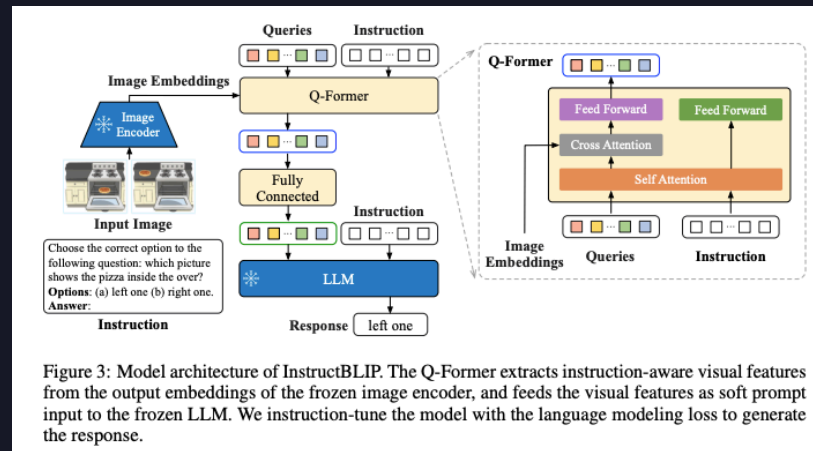
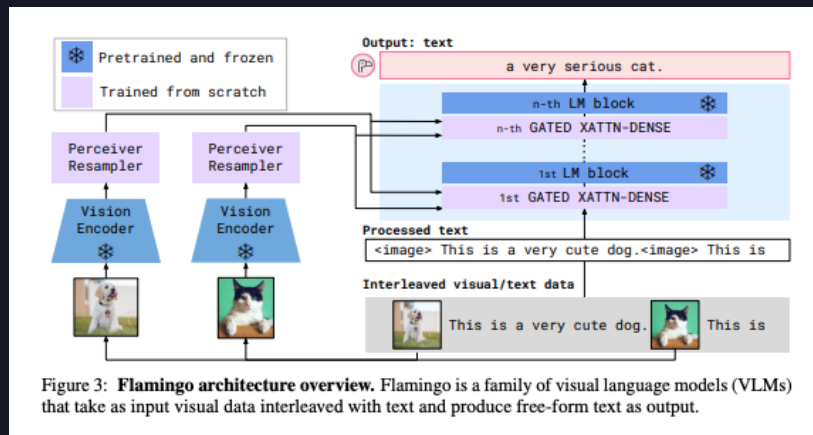
Or bolt on new ones

Add external cross-attention rather than touching the backbone's own layers.

LLaMA-VID adds separate layers between text queries and vision embeddings.

mPLUG-Owl3 runs text self-attention and vision-text cross-attention in parallel, merged by adaptive gates.

Perceiver Resampler + Q-Former



Where the merge happens

Early — before the LLM

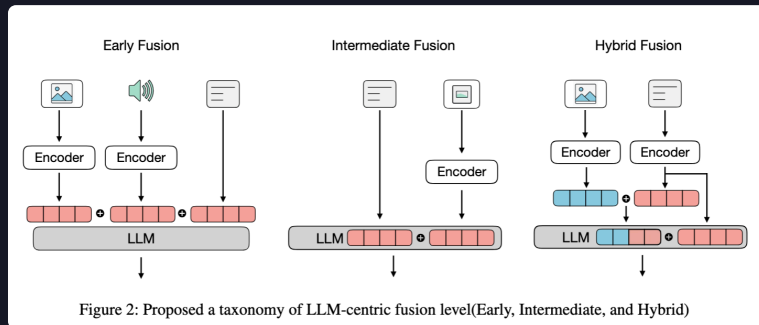
Features become tokens, then join the sequence. Efficient, integrity-preserving, simple. LLaVA, BLIP-2, Kosmos-1, LLaMA-VID

Intermediate — inside the stack

Layer-wise interaction; the model re-queries visual detail at each step. Flamingo, CogVLM, Cambrian-1, LLaMA-Adapter V2

Hybrid — both

Align early for global layout, inject mid-stack for fine detail. CogAgent, ManipLLM



AXIS 02

Representation of multimodal data

One shared space, two aligned spaces, or both in sequence — a choice driven by the task, not by the architecture.

Fuse, align, or align then fuse

Joint

one shared space

Everything becomes a token in the same transformer, fused by self- and cross-attention.

Good at fine-grained VQA, captioning, dialogue.

Weak at missing modalities and retrieval.

LLaVA · MiniGPT-4 · Flamingo · Qwen-VL

Coordinated

two spaces, pulled together

Separate encoders, contrastive loss. CLIP-style: matching pairs converge, mismatched pairs diverge.

Good at retrieval, zero-shot, swappable encoders.

Weak at token-level grounding.

CLIP · HACL · DPE

Hybrid

align, then fuse

Contrastively align modular encoders first, then merge through a Q-Former or cross-attention block.

Good at both search and grounding.
Cost two training regimes to tune.

BLIP-2 · MoVA · Video-LLaMA

The open question: can a model switch between these modes at inference, per input and per task?

AXIS 03

Training strategies and objectives

How many passes, over which data, updating which components — and against what loss.

One pass, two, or many

Single-stage

Frozen · InstructBLIP ·
BLIP

Train integration in one shot on enormous caption corpora. Simple pairwise data often isn't enough on its own, so practitioners mix in task data.

Two-stage

the default · LLaVA-1.5 ·
BLIP-2 · CogVLM

Stage 1 — alignment: lots of cheap caption pairs, LLM frozen, train the connector.

Stage 2 — instruction tuning: a little expensive curated data.

Uses both data regimes for what each is good at, and freezing in stage 1 blunts catastrophic forgetting.

Multi-stage

Qwen-VL · LongVILA ·
LHRS-Bot

Finer control: change which components unfreeze per stage, vary the dataset per stage, or expand data curriculum-style from easy to hard.

TABLE 1

◆ intermediate ◇ hybrid fusion ▲ coordinated ★ hybrid repr. otherwise early / joint

The map, applied

Model	Abstraction	Projection	Semantic embedding	Cross-attention
AnomalyGPT ★	—	Linear	Convolution	—
BLIP-2 ★	—	Linear	Q-Former	—
Cambrian-1 ◆	—	—	Cross-attention (conv-like)	—
CogAgent ◇	—	MLP	—	Within model
Flamingo ◆	Perceiver Resampler	—	—	Within model
InstructBLIP	—	Linear	Q-Former	—
Kosmos-1	Perceiver Resampler	MLP (last ViT layer)	—	—
LHRS-Bot	—	—	Perceiver Resampler	—
LLaMA-Adapter V2 ◆	—	Linear	—	Within model (self-attn)
LLaVA	—	Linear	—	—
LLaVA-1.5	—	MLP	—	—
MM1	C-Abstractor	—	—	—
mPLUG-Owl2 ◆	Visual Abstractor	—	—	Modality Adaptive Module
mPLUG-Owl3 ◆	—	Linear	—	Within model (HAT block)
PMC-VQA	—	MLP / Transformer	—	—
SEAL ▲	Convolution	MLP	—	—
SEED ★	—	Linear	Q-Former	—
VideoChat	Q-Former	Linear	—	—
VILA	—	Linear / Transformer	—	—
VisionLLM	—	Transformer	—	—
X-InstructBLIP	—	Q-Former + Linear	—	—

Three coordinates, one design space

- 01 Any MLLM can be located by three coordinates: mechanism × level, representation, training regime. They vary independently.
- 02 Classify by intent, not by module. The same Q-Former abstracts in one paper and carries semantics in the next.
- 03 The field has converged hard: early fusion, joint representation, two-stage training, LM loss. The corners of the space are largely unexplored.