Q-JEPA: PREDICTING QUESTION-RELEVANT VISUAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Acquiring visual representations that capture question-relevant regions is central to visual understanding. Existing approaches obtain such representations through explicit image manipulation, attention-guided visual selection, or autoregressive latent generation, requiring additional visual processing or intermediate generation within the language model. As the transition from a full image to a question-relevant local view can be considered as a zoom-in prediction in video, we propose Q-JEPA, which draws on V-JEPA's video-pretrained spatiotemporal representations and predictive priors to infer question-relevant target-view representations directly from the full-image context. These representations are predicted independently of the LLM's generation process, providing an effective and flexible visual interface with the potential to reduce inference overhead. Specifically, Q-JEPA introduces a Question-Guided Predictor that builds on the pretrained predictor with global cross-attention module to incorporate the full question context and focus cross-attention module to further associate target-view features with object- and relation-bearing contents in the question, thereby guiding prediction toward the relevant visual evidence. In addition, a three-stage training strategy progressively enables the predictor with the ability to produce question-relevant visual representations and subsequently aligns these representations with a downstream LLM for visual understanding. Evaluations show that Q-JEPA achieves 64.10% on GQA and 79.03% on PhysObjects, outperforming Qwen2.5-VL-7B by 3.72 and 19.58 percentage points, respectively, while reducing image inference latency by 22.2%.

1 INTRODUCTION

Recent multimodal large language models (MLLMs) (Alayrac et al., 2022; Li et al., 2023a; Zhu et al., 2023; Liu et al., 2024a; Bai et al., 2023; Peng et al., 2023; Bai et al., 2025; Team et al., 2023; Xu et al., 2025; Yan et al., 2026) have demonstrated remarkable capabilities in understanding and reasoning over multimodal information, particularly visual information, which provides a rich representation of the physical world. A common paradigm for enabling visual understanding in MLLMs is to equip a large language model (LLM) with a well-established visual encoder, such as CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023), or a custom-designed visual encoder, such as that used in Qwen 3 (Bai et al., 2023), to extract visual representations from images. However, when answering a specific question, humans do not typically rely exclusively on global visual information. Instead, we naturally attend to regions or objects that are relevant to the question or task, enabling more accurate and efficient visual understanding. This observation has motivated recent efforts toward question-guided target-region reasoning in MLLMs.

As shown in Figure 1, existing approaches can be roughly categorized into three groups. (a) Explicit image-space reasoning aims to manipulate the input image in pixel space. For example, by cropping or warping the image to focus on relevant regions (Shen et al., 2025; Zheng et al., 2026; Hong et al., 2026; Wang et al., 2025; Liu et al., 2024b; Sun et al., 2025; Su et al., 2025; Zhang et al., 2026). Although these tool-augmented or agent-based approaches can be effective, they typically require multiple calls to external vision tools, resulting in additional computational overhead and increased inference latency. (b) Attention-driven visual focusing exploits the MLLM's own ability to identify where to look by manipulating or analyzing its attention maps (Zhang et al., 2025; Gao et al., 2025; Dalal et al., 2026; Zhong et al., 2025; Wu et al., 2026). However, attention maps can vary substantially across layers and different MLLM architectures, which may limit the generalization and

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

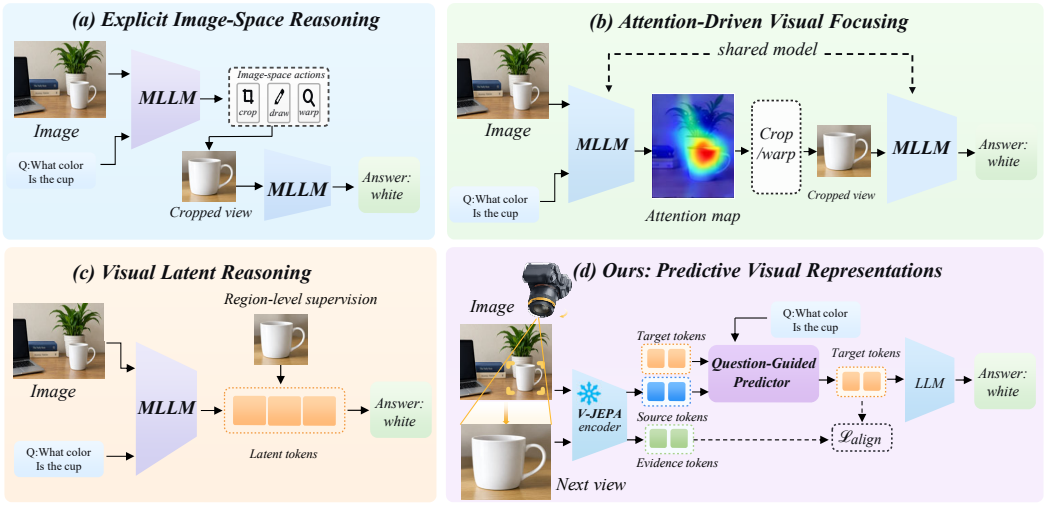

Figure 1: **Comparison of question-guided visual focusing paradigms**. Existing methods obtain question-relevant evidence through explicit image-space operations, attention-driven visual selection, or latent visual reasoning within the multimodal model. In contrast, Q-JEPA formulates visual focusing as a **zoom-in prediction problem** by directly predicting the question-relevant target-view representation in the latent space from the full-image context and question, and provides the predicted target visual features for downstream language-model reasoning.

robustness of such approaches. Moreover, obtaining and processing attention maps often requires additional computation or forward passes through the MLLM, further increasing the inference cost. (c) Visual latent reasoning introduces target regions as additional learnable latent queries within the MLLM (Li et al., 2026; Xu et al., 2026). Although these methods avoid explicit image-space manipulation, they still rely on the MLLM's visual representation space to encode target regions, which may limit the richness of the predicted representations for fine-grained and spatially localized visual understanding.

Unlike these approaches, **we formulate target-region reasoning as a zoom-in prediction problem in video**, and leverage the powerful spatiotemporal representations and predictive priors learned by V-JEPA (Assran et al., 2025; Mur-Labadia et al., 2026) from large-scale video data. Specifically, rather than explicitly manipulating pixels or relying on architecture-specific attention maps, we aim to predict the visual representation of a question-relevant target region directly in latent space, conditioned on the full-image context. The predicted target representation can then serve as question-relevant input visual features for LLMs, providing fine-grained visual information for downstream visual understanding. Building on this insight, we propose Q-JEPA, which introduces a Question-Guided Predictor (QGP) on top of a pretrained V-JEPA predictor to predict question-relevant target-region representations. To incorporate the question into the prediction process, the QGP is equipped with two complementary cross-attention modules. A global cross-attention module that incorporates the full question context into the prediction process, and a focus cross-attention module that further associates the predicted target-view features with object- and relation-bearing information in the question, guiding the predictor toward relevant visual evidence.

Explicitly parsing object- and relation-bearing expressions from the question at inference time would require an additional external language processing component. To eliminate this dependency during inference, we adopt a three-stage training strategy to progressively train the model to move from explicit target specification to implicit target discovery. In the first stage, the image question is fed into global cross-attention, while an explicit object- or relation-level focus phrase is provided to the focus cross-attention module, allowing the predictor to learn how to transform full-scene features into focused evidence representations under explicit target guidance. In the second stage, we replace both forms of explicit conditioning with the full question and introduce a FocusGate that learns a soft routing prior over question tokens, shifting the focus decision from externally specified cues to internal and question-conditioned selection. Finally, an understanding-oriented joint adaptation stage jointly trains the predictor and the vision–language interface to enable effective use of the predicted representations for downstream visual understanding.

In summary, our contributions are as follows: 1) We formulate target region reasoning as a zoom-in prediction problem in video and propose Q-JEPA, which deploys a question-conditional predictor on a pretrained V-JEPA predictor to predict question-relevant target region representations in latent space from the full-image context, providing fine-grained visual evidence to support effective and efficient downstream LLM understanding. 2) We equip the predictor with global cross-attention to incorporate the full question context and focus cross-attention to associate target-view features with object- and relation-bearing information, enabling more accurate question-guided prediction of relevant visual evidence. 3) We propose a three-stage training strategy that progressively enables the predictor to identify and predict question-relevant visual representations and align them with downstream LLMs. 4) Extensive experiments on GQA (Hudson & Manning, 2019) and PhysObjects (Gao et al., 2024) demonstrate superior accuracy and favorable parameter efficiency and inference speed compared with competing methods. We further extend Q-JEPA to short video understanding, demonstrating its applicability beyond single-image tasks.

2 Related Work

Explicit Image-Space Reasoning. Previous works based on explicit image-space reasoning acquire visual evidence by manipulating the input image or constructing intermediate views. ZoomEye (Shen et al., 2025) explores an image through a hierarchy of zoomed regions, while Visuo-Think (Wang et al., 2025) combines visual and textual reasoning with look-ahead tree search. DeepEyes (Zheng et al., 2026) learns tool-assisted visual reasoning through reinforcement learning, and Pixel Reasoner (Su et al., 2025) combines instruction tuning with curiosity-driven reinforcement learning to encourage operations such as zooming and frame selection. DeepEyesV2 (Hong et al., 2026) further extends tool integration toward agentic multimodal reasoning, while Thyme (Zhang et al., 2026) generates executable code for image processing and computational operations. These methods explicitly acquire task-relevant visual evidence through image-space operations, whereas Q-JEPA predicts the target-view representation directly in latent space.

Attention-Driven Visual Focusing. Attention-driven methods leverage internal VLM signals to identify question-relevant visual regions. ViCrop (Zhang et al., 2025) exploits attention or gradient maps to identify informative regions, while ICoT (Gao et al., 2025) incorporates attention-driven visual selection into interleaved visual-textual reasoning. FOCUS (Zhong et al., 2025) derives object relevance maps from the key-value cache for fine-grained VQA. AttWarp (Dalal et al., 2026) uses attention-guided image warping to redistribute visual resolution, while ConFoThinking (Wu et al., 2026) consolidates query-sensitive attention signals for region extraction. These methods rely on internal attention signals for visual focusing, while Q-JEPA performs target-region prediction without explicitly extracting attention maps.

Visual Latent Reasoning. Visual latent reasoning represents intermediate visual evidence through continuous latent states. Latent Visual Reasoning (LVR) (Li et al., 2026) trains the language model to autoregressively generate latent states that reconstruct question-relevant visual tokens and interleaves them with textual generation. Semantic-Enriched Latent Visual Reasoning (SLVR) (Xu et al., 2026) further enriches region-centric latents with fine-grained attribute supervision and aligns them across multiple questions through Multi-query Group Relative Policy Optimization. These methods explore latent representations as an alternative to explicit image-space visual reasoning, while Q-JEPA leverages a pretrained predictive visual model for target-region representation prediction.

Predictive Visual Representations. Joint-Embedding Predictive Architectures (JEPAs) (LeCun, 2022) learn visual representations by predicting target features from contextual observations rather than reconstructing pixels. V-JEPA (Bardes et al., 2024) extends this predictive paradigm to video and learns spatiotemporal representations from large-scale video data, with subsequent works (Assran et al., 2025; Mur-Labadia et al., 2026) further improving its representation and predictive capabilities. Motivated by the human ability to zoom in on question-relevant regions, we formulate question-guided target-region reasoning as a zoom-in prediction problem in video and build Q-JEPA upon a pretrained V-JEPA predictor.

3 Methodology

We formulate target-region reasoning as a zoom-in prediction problem in video. The goal is to predict the representation of a question-relevant target view from the full-image context and the ques-

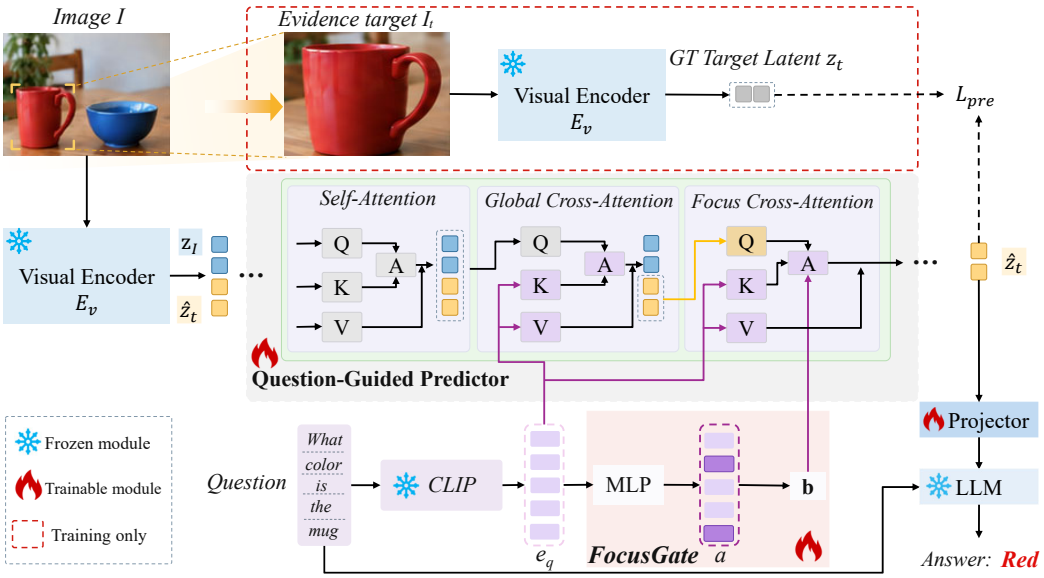

Figure 2: **Overview of Q-JEPA.** Built upon pretrained V-JEPA 2.1, Q-JEPA uses a frozen visual encoder to extract full-image features $\mathbf{z}_I$. A Question-Guided Predictor (QGP) then predicts question-relevant target-view representations from these features, conditioned on the question. QGP augments the pretrained self-attention blocks with global cross-attention to incorporate the complete question context and focus cross-attention to emphasize object- and relation-bearing cues through a learned FocusGating module. A lightweight projector aligns the predicted representations with a downstream LLM for visual understanding.

tion. This formulation naturally aligns with the predictive paradigm of V-JEPA 2.1 (Mur-Labadia et al., 2026), allowing us to leverage its powerful spatiotemporal representations and predictive priors learned from large-scale video data. Based on this insight, we build Q-JEPA on top of V-JEPA 2.1 for question-guided target-region reasoning. In this section, we first introduce the architecture of V-JEPA 2.1 and then present Q-JEPA, including its question-conditional predictor with global and focus cross-attention modules, followed by the three-stage training strategy.

3.1 Preliminaries

V-JEPA 2.1 (Mur-Labadia et al., 2026) is an extension of the video-based Joint-Embedding Predictive Architecture (V-JEPA) (Bardes et al., 2024), which learns visual representations by predicting target features from contextual observations rather than reconstructing pixels. It consists of a visual encoder E_θ, a target encoder $\mathrm{E}_{\bar{\theta}}$, and a predictor P_ψ, where θ and ψ denote the learnable parameters. Given a context observation α and a target observation β, the model is trained to predict the target representation from the contextual representation by minimizing

$$\min_{\theta,\psi} \left\| P_\psi\left(E_\theta(\alpha)\right) - \mathrm{sg}\left(E_{\bar{\theta}}(\beta)\right) \right\|_1, \tag{1}$$

where sg denotes the stop-gradient operation, and $E_{\bar{\theta}}$ is an exponential moving average (EMA) of the visual encoder E_θ.

3.2 Q-JEPA

3.2.1 Overview

To leverage the powerful spatiotemporal representations and predictive priors of V-JEPA 2.1, Q-JEPA is built on top of the pretrained V-JEPA 2.1 model. As shown in Figure 2, we freeze both the visual encoder and the target encoder, which are collectively denoted as $\mathrm{E_v}$ for simplicity. Q-JEPA differs from the original V-JEPA 2.1 in its definitions of the context and target. In V-JEPA 2.1, α denotes the contextual observations, while β denotes the masked target region. In Q-JEPA, we instead

define α as the full-image context $\mathbf{I}$ and β as the question-relevant target view $\mathbf{I_t}$, corresponding to the region on which we aim to zoom in. Detailed illustration refer to Appendix A Figure 4.

As shown in Figure 2, the Q-JEPA pipeline consists of three main components. First, the frozen visual encoder $\mathrm{E_v}$ extracts the latent representation $\mathbf{z_I} = \mathrm{E_v}(\mathbf{I})$ from the full-image context $\mathbf{I}$. Secondly, a question-guided predictor (QGP) predicts the target representation $\hat{\mathbf{z}}_\mathbf{t}$ conditioned on $\mathbf{z_I}$ and the question embedding $\mathbf{e_q}$, where $\mathbf{e_q}$ is obtained from a frozen text encoder such as CLIP (Radford et al., 2021). Finally, to align the predicted target representation $\hat{\mathbf{z}}_\mathbf{t}$ with the visual feature space expected by a downstream LLM, we introduce a lightweight and learnable projector. Following Qwen2.5-VL (Bai et al., 2025), the projector first spatially merges $\hat{\mathbf{z}}_\mathbf{t}$ by a factor of 2 using average pooling, followed by MLP layers.

To train Q-JEPA, we propose a three-stage training strategy that progressively enables the predictor to identify and predict question-relevant visual representations and aligns the predicted representations with a downstream LLM for visual understanding.

3.2.2 Question-Guided Predictor

The Question-Guided Predictor (QGP) builds on the pretrained V-JEPA 2.1 predictor, whose backbone consists primarily of self-attention blocks. To guide target-latent prediction toward question-relevant visual evidence, we augment the predictor with two complementary modules: a global cross-attention module and a focus cross-attention module.

Global Cross-Attention Module (GCAM). The GCAM uses the hidden state corresponding to the full-image context features and the target latent tokens as queries, with the question embeddings $\mathbf{e}_q$ serving as keys and values. It integrates the complete question context into the visual token sequence, leading the target-view prediction conditioned on the question.

Focus Cross-Attention Module (FCAM). While the complete question provides contextual information, specific words identifying objects, attributes, or relations often offer more direct cues for visual focusing, such as "mug" and "color" in Figure 2. We therefore introduce FCAM to emphasize these cues when updating the target latent tokens. This module uses only target tokens as queries and the question embeddings as keys and values. To prioritize informative question tokens, we introduce an MLP-based **FocusGate** module, denoted by FG, which maps each question token embedding to a scalar relevance score. Given $\mathbf{e}_q = [e_0; \ldots; e_{L-1}]$, where L is the sequence length. The routing weights are computed as

$$a_i = \frac{\exp\left(\mathrm{FG}(e_i)\right)}{\sum_{j=0}^{L-1} \exp\left(\mathrm{FG}(e_j)\right)}, \qquad i \in \{0, \ldots, L-1\}. \tag{2}$$

We convert these weights into an additive attention bias, $\mathbf{b} = [b_1, \ldots, b_L]$, where $b_i = \log(La_i + \epsilon)$ and $\epsilon = 10^{-6}$ ensure numerical stability. The focus-attention weights are then computed as

$$A_f = \mathrm{softmax}\left(\frac{Q_t K_q^\top}{\sqrt{d_k}} + \mathbf{b}\right), \tag{3}$$

where Q_t and K_q denote the projected target queries and question keys, respectively. The dimension of each attention head is denoted by d_k.

3.2.3 Three-Stage Learning Strategy

We propose a three-stage training strategy to progressively develop the predictor's ability to produce question-relevant visual representations and align them with a downstream LLM for visual understanding. The three stages are **explicit focus-guided warmup**, **implicit focus-guided learning**, and **LLM alignment**.

Stage 1: Explicit focus-guided warmup. In this stage, we bypass the FocusGating module and provide the focus cross-attention module with explicit object- and relation-bearing cues extracted from the question. These cues can be identified using syntactic analysis, such as dependency parsing. We encode the extracted words with the text encoder to obtain the focus embeddings $\mathbf{e}_f$, which serve as keys and values for focus cross-attention. This explicit guidance helps the predictor associate target latent features with relevant question content without requiring it to identify informative

tokens within the full question. Meanwhile, the global cross-attention module continues to receive the complete caption embeddings.

In this stage, we adopt the feature prediction objective in Equation 1, which can be adaptively expressed as

$$\mathcal{L} = \mathcal{L}_{\text{pred}} = \left\| \mathrm{P}_\psi \left(\mathrm{E}_v(\mathbf{I}), \mathbf{e}_q, \mathbf{e}_f \right) - \mathrm{sg}\left(\mathrm{E}_v(\mathbf{I}_t) \right) \right\|_1, \tag{4}$$

where only the parameters ψ of the question-guided predictor are optimized in this stage.

Stage 2: Implicit focus-guided learning. In this stage, we enable the FocusGating module and replace the explicit focus embeddings with the complete question embeddings as the input to focus cross-attention. To train FocusGating to identify informative question tokens, we supervise its routing weights a_i using the annotated object- and relation-bearing token positions.

Let $\mathcal{K}_{\text{pos}} \subseteq \{0, \ldots, L-1\}$ denote the set of annotated token positions, with $N = |\mathcal{K}_{\text{pos}}| \leq L$. Given the routing weights a_i from Equation 2, we define the total routing mass assigned to these positions and their normalized routing distribution as

$$s_{\text{pos}} = \sum_{i \in \mathcal{K}_{\text{pos}}} a_i, \qquad p_i = \frac{a_i}{s_{\text{pos}}}, \quad i \in \mathcal{K}_{\text{pos}}. \tag{5}$$

The focus loss is then defined as

$$\mathcal{L}_{\text{focus}} = \max\left\{0, \log\tau - \log s_{\text{pos}}\right\} + \lambda_D \, D_{\text{KL}}\left(\mathcal{U}_{\text{pos}} \parallel p_i\right), \quad i \in \mathcal{K}_{\text{pos}}, \tag{6}$$

where $\mathcal{U}_{\text{pos}}$ is the uniform distribution over the annotated positions, assigning probability $1/N$ to each position. $\lambda_D = 0.1$.

The first term encourages the annotated tokens to receive a total routing mass of at least τ, where

$$\tau = \max\left\{u, \, \min\left\{\gamma_0, \, u + \gamma_1(1-u)\right\}\right\}, \qquad u = \frac{N}{L}, \tag{7}$$

with $\gamma_0 = 0.60$ and $\gamma_1 = 0.35$. Here, u represents the routing mass that the annotated tokens would receive under a uniform distribution over the question. The adaptive threshold τ encourages greater emphasis on these tokens while accounting for their proportion in the sequence. The KL divergence term D_{KL} encourages a balanced allocation of routing mass among the annotated tokens, discouraging concentration on only a subset of the relevant cues.

In this stage, we jointly optimize the target feature prediction and focus objectives:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{focus}} \mathcal{L}_{\text{focus}}, \tag{8}$$

where $\lambda_{\text{focus}} = 0.005$ controls the contribution of the focus loss.

Stage 3: LLM alignment. In the final stage, we introduce a lightweight projector to align the question-relevant visual representations predicted by Q-JEPA with the input embedding space of a downstream LLM. We supervise visual understanding using a token-level cross-entropy loss, denoted by $\mathcal{L}_{\text{llm}}$. We retain the feature prediction and focus losses to preserve target-view prediction and question-guided focusing during alignment. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{llm}} + \lambda_{\text{pred}} \mathcal{L}_{\text{pred}} + \lambda_{\text{focus}} \mathcal{L}_{\text{focus}}, \tag{9}$$

where $\lambda_{\text{pred}} = 0.5$ and $\lambda_{\text{focus}} = 0.005$ control the contributions of the feature prediction and focus losses, respectively.

4 Experiments

4.1 Experiment Details

Experimental Settings. We implement Q-JEPA on top of V-JEPA 2.1 (Mur-Labadia et al., 2026) with a ViT-L/16 visual encoder. For question embedding extraction, we use the ViT-B/32 text encoder from CLIP (Radford et al., 2021). The downstream LLM is the language component of Qwen2.5-VL-7B-Instruct. All components of these pretrained models are kept frozen during training. The input resolution is 384×384, and the batch size is 32. All experiments are conducted on eight NVIDIA RTX 4090 GPUs.

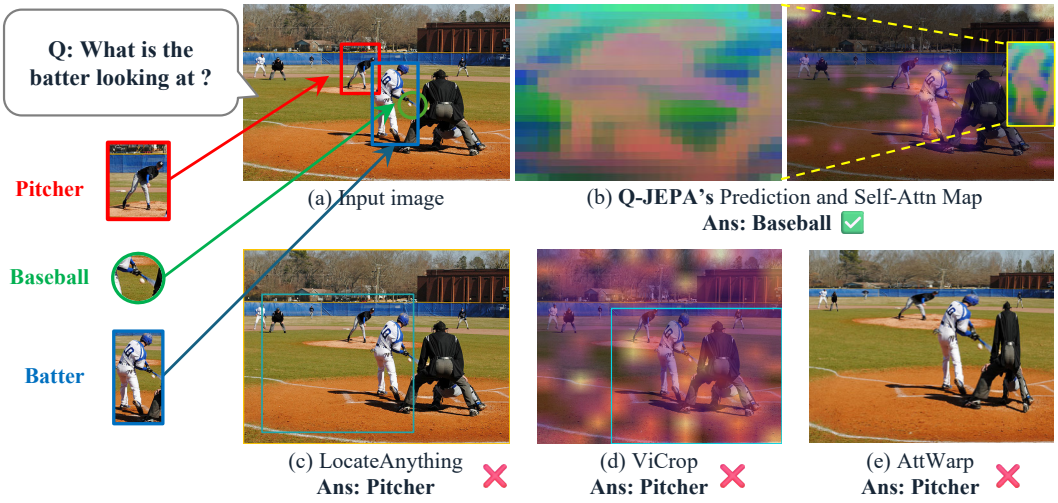

Figure 3: **Qualitative Comparisons.** (a) is the input image. (b) shows the visualization of Q-JEPA's latent prediction (left) and self-attention map (right), which highlights the predicted target region in the full image. To better visualize the predicted target view, we coarsely resize the predicted representation to the input spatial resolution (highlighted with a yellow box). (c), (d), and (e) are the grounding result, attention region, and warp result of LocateAnything (Wang et al., 2026), ViCrop (Zhang et al., 2025), and ttWarp (Dalal et al., 2026), respectively. These results demonstrate that Q-JEPA predicts more accurate question-relevant representations, enabling more reliable answer generation. In contrast, the attended regions of the other three methods are relatively broad and fail to isolate the small target object, *baseball*, resulting in the same incorrect answer, *pitcher*.

Datasets. In the first training stage, Q-JEPA is trained on Flickr30k (Plummer et al., 2015; Young et al., 2014) with 374,352 samples. In the second and third stages, we use the training sets of GQA (Hudson & Manning, 2019) and PhysObjects (Gao et al., 2024), containing 943,000 and 18,740 samples, respectively. GQA primarily covers general visual understanding and reasoning, while PhysObjects focuses on understanding physical properties of objects. For evaluation, we use the GQA test-dev set, which contains 12,578 samples, and the PhysObjects test set, which contains approximately 868 questions covering material, transparency, contents, liquid containment, and sealed-state understanding.

Evaluation Metrics. We report exact-match accuracy for GQA. For PhysObjects, predictions are selected by conditional log-likelihood over property-specific candidate answers.

4.2 Comparison Experiments

We compare Q-JEPA with Qwen2.5-VL-7B and representative approaches from three paradigms: explicit image-space reasoning, attention-driven visual focusing, and visual latent reasoning. Specifically, we include PixelReasoner (Su et al., 2025), Thyme (Zhang et al., 2026), ViCrop (Zhang et al., 2025), AttWarp (Dalal et al., 2026), and LVR (Li et al., 2026) with eight reasoning steps. To provide a more comprehensive comparison, we additionally consider the representative grounding method LocateAnything (Wang et al., 2026), which first grounds the question-relevant regions and then feeds the grounded regions to Qwen2.5-VL-7B for visual understanding.

We evaluated the compared methods using publicly released code and checkpoints where available. For AttWarp (Dalal et al., 2026), the authors provide a LLaVA-1.5-7B-based (Liu et al., 2024a) example but do not release their Qwen2.5-VL-7B attention-extraction implementation. We therefore implemented the Qwen2.5-VL-7B variant following the paper, using the released LLaVA example as a reference.

Quantitative Comparisons. Table 1 compares Q-JEPA with the baseline VLM and representative methods across three visual reasoning paradigms. Q-JEPA achieves the highest accuracy on both GQA and PhysObjects, reaching 64.10% and 79.03%, respectively. Compared with Qwen2.5-VL-7B, these results correspond to absolute improvements of 3.72 and 19.58 percentage points, demonstrating the effectiveness of the proposed question-relevant visual representations for downstream

Table 1: **Effectiveness, efficiency, and model size comparisons on GQA and PhysObject.** Q-JEPA demonstrates favorable performance across all three aspects. Row colors denote base VLMs, explicit image-space reasoning, attention-driven visual focusing, visual latent reasoning, and our predictive visual representations. Parameter counts include only the visual modules and exclude the downstream LLM.

Methods	GQA (%) ↑	PhysObjects (%) ↑	Visual Params (B) ↓	Time (s) ↓
Qwen2.5-VL-7B	60.38	59.45	0.677	0.171
LocateAnything (ECCV 2026)	56.01	60.83	1.107	0.454
PixelReasoner (NeurIPS 2025)	60.73	63.25	0.677	1.864
Thyme (ICLR 2026)	60.76	50.69	0.677	4.655
ViCrop (ICLR 2025)	61.00	66.24	0.677	0.410
AttWarp (ICLR 2026)	60.37	58.64	0.677	0.225
LVR (8 Steps) (ICLR 2026)	57.09	54.61	0.677	0.496
Q-JEPA	**64.10**	**79.03**	**0.421**	**0.133**

visual understanding. LocateAnything (Wang et al., 2026) achieves lower accuracy than Qwen2.5-VL-7B on GQA but modestly improves performance on PhysObjects, suggesting that explicit region localization alone does not consistently improve visual understanding. Q-JEPA, in comparison, predicts question-relevant target-view features conditioned on full-image context and improves performance on both benchmarks. Q-JEPA also outperforms explicit image-space reasoning methods, including PixelReasoner (Su et al., 2025) and Thyme (Zhang et al., 2026); attention-driven visual focusing methods, including ViCrop (Zhang et al., 2025) and AttWarp (Dalal et al., 2026); and the visual latent reasoning method LVR (Li et al., 2026). Relative to ViCrop, the strongest competing method on both benchmarks, Q-JEPA improves accuracy by 3.10 percentage points on GQA and 12.79 percentage points on PhysObjects. These comparisons support predictive visual focusing as an effective approach to providing question-relevant visual evidence for LLM-based reasoning.

Qualitative Comparisons. Figure 3 illustrates Q-JEPA's ability to capture fine-grained, question-relevant visual evidence. Answering "What is the batter looking at?" requires identifying the small baseball near the batter. Q-JEPA's attention map emphasizes the batter's local interaction with the ball, and its predicted target-view representation supports the correct answer, *Baseball*. In contrast, LocateAnything and ViCrop select broader regions, while AttWarp redistributes image resolution without sufficiently isolating the relevant evidence; all three methods incorrectly answer **Pitcher**. This example highlights the value of predicting focused visual representations for distinguishing small, task-relevant objects from prominent contextual distractors.

Efficiency and Model Size Analyses. Table 1 compares inference time and visual-module parameter counts, excluding the downstream LLM. Among methods with reported measurements, Q-JEPA achieves the lowest inference time of 0.133 s, reducing latency by 22.2% relative to Qwen2.5-VL-7B. It is also approximately $3.08\times$ faster than ViCrop and $3.73\times$ faster than LVR. Meanwhile, Q-JEPA uses 0.421 B visual-module parameters, compared with 0.677 B for Qwen2.5-VL-7B, a reduction of 37.8%. These results show that Q-JEPA improves accuracy while requiring fewer visual-module parameters and achieving lower inference latency.

4.3 Ablation Studies

To evaluate the contributions of individual components in Q-JEPA, we conduct ablation studies on GQA, as summarized in Table 2.

Effectiveness of the Pretrained Predictor. With explicit focus-guided warmup disabled, initializing the predictor from pretrained weights improves accuracy from 61.16% to 62.75%, a gain of 1.59 percentage points. This comparison supports the benefit of transferring video-pretrained predictive priors to question-guided visual representation learning.

Effectiveness of Focus Cross-Attention (FCA). With the pretrained predictor and Stage 1 training enabled, adding FCA without the focus loss improves accuracy from 62.22% to 63.77%, a gain of

Table 2: **Ablation study results on GQA.** We examine the contributions of predictor pretraining, Stage 1 explicit focus-guided warmup, focus cross-attention, and the focus loss $\mathcal{L}_{\text{focus}}$. The full Q-JEPA configuration achieves the highest accuracy of **64.10%**, supporting the effectiveness of the combined design.

Pretrained Predictor	Stage1 Explicit Focus	Focus Cross-Attention	L_{focus}	GQA
		✓	✓	61.16%
✓		✓	✓	62.75%
✓	✓			62.22%
✓	✓	✓		63.77%
✓	✓	✓	✓	**64.10%**

1.55 percentage points. This result supports the effectiveness of a dedicated cross-attention module for incorporating question content into target-view prediction.

Effectiveness of the Focus Loss. Adding $\mathcal{L}_{\text{focus}}$ to the configuration with the pretrained predictor, Stage 1 training, and FCA increases accuracy from 63.77% to 64.10%. This gain of 0.33 percentage points suggests that explicit supervision of the routing weights further improves question-guided visual focusing.

Effectiveness of Stage 1 Training. With the pretrained predictor, FCA, and focus loss enabled, introducing explicit focus-guided warmup improves accuracy from 62.75% to 64.10%, a gain of 1.35 percentage points. This result supports progressively transitioning from explicit semantic cues to implicit cue identification, rather than learning question-guided focusing directly from the full question.

4.4 Extension to Video Understanding

To evaluate the applicability of Q-JEPA beyond single-image understanding, we extend it to video clips. Specifically, the visual encoder E_v extracts features from the sampled frames, which are concatenated to form a visual context sequence. This sequence is concatenated with the target latent tokens and passed to the question-guided predictor, conditioned on the question embeddings. The predictor produces a single question-relevant target-view representation for the entire clip, which is subsequently projected into the downstream LLM's input embedding space for answer generation.

We evaluate this extension on a subset of STAR (Wu et al., 2021) comprising 1,111 samples. Each clip is represented by four frames obtained using one of two sampling strategies: (1) Raw-Uniform-4F (STAR-Raw), which uniformly samples four frames within the official clip interval, and (2) AG-Time-4F (STAR-AG), which uses annotation-assisted keyframe sampling. We report four-choice accuracy as a percentage. Answers are generated using greedy decoding with a maximum of 16 output tokens, and invalid or conflicting responses are counted as incorrect.

Q-JEPA achieves accuracies of 75.25% on STAR-Raw and 73.36% on STAR-AG, compared with 67.96% and 69.67% for Qwen2.5-VL-7B, yielding gains of 7.29 and 3.69 percentage points, respectively. These improvements hold under both sampling strategies, even though Q-JEPA represents the four-frame context through a single predicted target view. The results support the applicability of predictive visual focusing to video understanding.

5 Conclusion

We presented Q-JEPA, a framework that formulates question-guided visual focusing as a zoom-in prediction task in latent space. By adapting a pretrained V-JEPA predictor with global and focus cross-attention, Q-JEPA predicts question-relevant target-view representations from full-image context. A three-stage learning strategy progressively transitions from explicit focus guidance to implicit cue identification and aligns the predicted representations with a downstream LLM. Experiments on GQA and PhysObjects demonstrate improved accuracy with fewer visual-module parameters and lower inference latency, while the extension to STAR supports the applicability of this predictive interface to video understanding.

AI USE STATEMENT

We follow ICLR's policy and use LLMs solely as general-purpose assistive tools. In this work, LLMs were used to polish the writing of selected paragraphs and to suggest improvements to code. They were not involved in research ideation, experimental design, or the development of the core technical contributions.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Jinze Bai, Shuai Bai, Yun Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. doi: 10.48550/arXiv.2309.16609.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. doi: 10.48550/arXiv.2502.13923. URL https://arxiv.org/abs/2502.13923.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

Dwip Dalal, Gautam Vashishtha, Utkarsh Mishra, Jeonghwan Kim, Madhav Kanda, Hyeonjeong Ha, Svetlana Lazebnik, Heng Ji, and Unnat Jain. Constructive distortion: Improving MLLMs with attention-guided image warping. In *International Conference on Learning Representations*, 2026.

Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19520–19529, 2025.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. DeepEyesV2: Toward agentic multimodal model. In *International Conference on Learning Representations*, 2026.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Yann LeCun. A path towards autonomous machine intelligence, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. Position paper.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023a.

Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. FACTUAL: A benchmark for faithful and consistent textual scene graph parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6377–6390, Toronto, Canada, July 2023b. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-acl.398.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-Spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024b.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mahmoud Assran, Koustuv Sinha, Michael Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shumin Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. doi: 10.48550/arXiv.2306.14824.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. doi: 10.48550/arXiv.2103.00020. URL https://arxiv.org/abs/2103.00020.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. ZoomEye: Enhancing multimodal LLMs with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6602–6618, 2025.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel Reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 9130–9159, 2025.

Guangyan Sun, Mingyu Jin, Zhenting Wang, Chenglong Wang, Siqi Ma, Qifan Wang, Tong Geng, Yingnian Wu, Yongfeng Zhang, and Dongfang Liu. Visual agents as fast and slow thinkers. In *International Conference on Learning Representations*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jing Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katherine Millican, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. doi: 10.48550/arXiv.2312.11805.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, Jan Kautz, Lei Zhang, and Zhiding Yu. Locateanything: Fast and high-quality vision-language grounding with parallel box decoding. *arXiv preprint arXiv:2605.27365*, 2026.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. VisuoThink: Empowering LVLM reasoning with multimodal tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21707–21719, 2025.

Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/5ef059938ba799aaa845e1c2e8a762bd-Paper-round2.pdf.

Zhaodong Wu, Haochen Xue, Qi Cao, Wenqi Mo, Yu Pei, Wenqi Xu, Jionglong Su, and Yang Liu. ConFoThinking: Consolidated focused attention driven thinking for visual question answering. *arXiv preprint arXiv:2603.00165*, 2026.

Tianrun Xu, Guo Chen, Yifan Li, Yizhou Xi, Zhen Mu, Rui Wang, Tianren Zhang, Hongyang Gao, and Feng Chen. OURO: A self-bootstrapped framework for enhancing multimodal scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18240–18251, 2025.

Tianrun Xu, Yue Sun, Qixun Wang, Jingyi Lu, Yuan Wang, Tianren Zhang, Longteng Guo, Fengyun Rao, Jing Lyu, Feng Chen, and Jing Liu. Semantic-enriched latent visual reasoning. In *International Conference on Machine Learning*, 2026.

Shen Yan, Ji Tong, Haochen Xue, Xiaoyu Tang, Yifan Wang, Kai Shi, Guanhua Zhang, Rong Li, and Yang Zou. Act wisely: Cultivating meta-cognitive tool use in agentic multimodal models. *arXiv preprint arXiv:2604.08545*, 2026. doi: 10.48550/arXiv.2604.08545.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. doi: 10.48550/arXiv.2303.15343. URL https://arxiv.org/abs/2303.15343.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images. In *International Conference on Learning Representations*, 2026.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "Thinking with Images" via reinforcement learning. In *International Conference on Learning Representations*, 2026.

Liangyu Zhong, Fabio Philipp Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. FOCUS: Internal MLLM representations for efficient fine-grained visual question answering. In *Advances in Neural Information Processing Systems*, volume 38, pp. 107014–107051, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xi Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. doi: 10.48550/arXiv.2304.10592.

APPENDIX

A V-JEPA V.S. Q-JEPA

From temporal prediction to a virtual target view. The pretrained V-JEPA 2.1 predictor maps encoded context observations and learned target queries to features at a masked or future target position. In Q-JEPA, the same predictive interface is repurposed for question-conditioned visual evidence. Unlike the original EMA teacher/student pretraining setup, the Q-JEPA implementation uses one frozen ViT-L/16 encoder for both the source image and its training-only evidence crop. It processes a 384×384 image into a 24×24 grid of 1,024-dimensional source features. The predictor receives these 576 source tokens and 576 learned target queries. Source tokens occupy the observed temporal position, whereas target queries occupy a second, *virtual* temporal position in the positional encoding. This second position is a readout slot for a question-relevant view; it does not denote an observed second image, a generated RGB frame, or physical future prediction. For the four-frame video extension, four independently encoded source grids are concatenated in temporal order and the 576 target queries occupy the next virtual position.

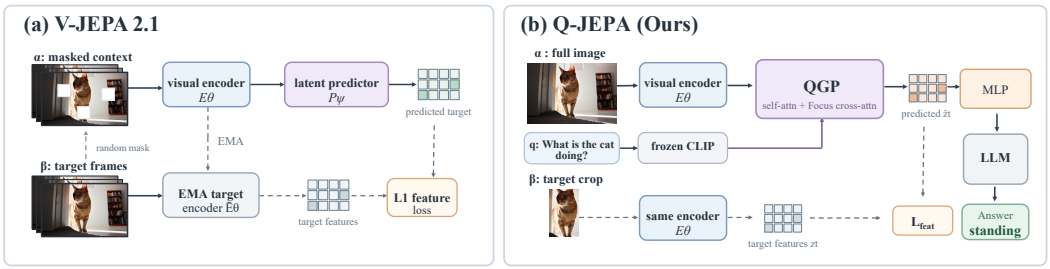

Figure 4: **Comparison of V-JEPA 2.1 and our Q-JEPA**

Predictor computation. The Q-JEPA predictor has 12 Transformer blocks, a hidden width of 384, and 12 attention heads. Let $S \in \mathbb{R}^{N_s \times d}$ and $T \in \mathbb{R}^{576 \times d}$ denote the source and target tokens after projection, and let $E_q \in \mathbb{R}^{L \times d}$ be frozen CLIP ViT-B/32 question-token embeddings after the learned text projection. Each block first applies self-attention to $[S; T]$. Global cross-attention then uses *all* source and target hidden states as queries and question tokens as keys and values. Focus cross-attention uses *only target* hidden states as queries; its keys and values are again the question tokens. Thus, the two text-attention branches differ in their query set and, for Focus, in the learned routing bias. The final predictor output is a 576×1024 target-feature grid.

B DETAILS FOR DATASET CONSTRUCTION

Overview. We construct the training and evaluation data for Q-JEPA from four sources: Flickr30k Entities, GQA, PhysObjects, and STAR. Flickr30k is used for explicit focus-conditioned pretraining, where each training example pairs a full image and caption with an object- or relation-level focus phrase and its corresponding evidence region. GQA is then used for question-conditioned transfer, where the explicit focus phrase is removed and the original question becomes the only textual input. PhysObjects and STAR further extend the training and evaluation to physical-property reasoning and video question answering, respectively. Table 3 summarizes the processed data used in this study. The reported counts refer to the constructed protocols used by our project rather than the complete original benchmarks.

B.1 FLICKR30K: FOCUS-CONDITIONED PRETRAINING DATA

Source data. Stage 1 is constructed from the training split of Flickr30k Entities (Plummer et al., 2015). Each training example contains a full image, its caption, an explicit object- or relation-level focus phrase, and a corresponding evidence region. The complete caption is provided to the Global text branch, whereas the extracted focus phrase is provided to the Focus branch.

Table 3: Prepared data used in this study. Counts describe the processed protocols used by our project rather than the complete original benchmarks. STAR evaluation uses a subset of the official validation split rather than the hidden test split.

Data role	Train	Development	Reported evaluation
Flickr30k region–caption pairs	374,352	feature validation	—
GQA balanced questions	943,000	132,062	12,578 test-dev
PhysObjects prepared questions	18,740	1,032	868
STAR Interaction_T1 questions	4,593	501	1,111

Object focus construction. We retain caption-linked object phrases with annotated bounding boxes from six core entity categories: people, animals, vehicles, instruments, clothing, and body parts. When one phrase is associated with multiple bounding boxes, their union is used as a single evidence region. This procedure produces 278,904 object-focused training examples.

Relation focus construction. To construct relation-level focus examples, caption relations are parsed using the FACTUALSCENEGRAPH model (Li et al., 2023b). The argument entities of each parsed relation are then matched to the grounded Flickr entities. We use a minimum entity-matching score of 0.45 and discard ambiguous matches when the difference between the two highest matching scores is below 0.05. At most eight successfully grounded relations are retained for each caption record. This procedure produces 95,448 relation-focused examples.

Target-view construction. For each retained object or relation, the corresponding evidence bounding box is expanded by 8% to include a small amount of surrounding context, clipped to the image boundary, and resized to the input resolution of the frozen visual encoder. The full image and the resulting evidence crop form a pseudo-temporal source–target pair in feature space: the full image provides the observed source context, while the crop provides the latent target-view representation to be predicted.

Labels. The object or relation phrase serves as the explicit focus text, while the corresponding evidence crop provides the target visual feature. In total, Stage 1 contains 278,904 object examples and 95,448 relation examples, yielding 374,352 region–caption training pairs. This number counts constructed training examples rather than unique Flickr30k images.

B.2 GQA: QUESTION-CONDITIONED TRAINING DATA

Input format. Stage 2 uses 943,000 prepared questions from the GQA balanced training split. Unlike Stage 1, the explicit object- or relation-level focus phrase is no longer provided as model input. Instead, the original question is encoded by the frozen CLIP text encoder and supplied to both the Global and Focus cross-attention branches. The predictor must therefore infer the question-relevant visual evidence from the question itself.

Evidence-target construction. For each training example, an evidence region is constructed from the available GQA training annotations. The corresponding crop is encoded by the frozen visual encoder and used as the target representation for latent feature prediction. The evidence crop and its coordinates serve only as training supervision; they are not provided to the predictor as visual input and are not required at inference time.

Focus labels. For training examples with valid entity annotations, the corresponding entity-bearing phrases are aligned with the question tokens and retained as focus labels. These labels supervise the FocusGate routing distribution, encouraging the model to assign higher routing mass to question tokens that identify the relevant entities. Importantly, the labels are used only as supervision: the model still receives the original full question as its textual input, without an additional explicit focus phrase. Examples without valid focus labels do not contribute to the auxiliary focus loss. No focus annotations are required during inference.

Development and evaluation. We use 132,062 GQA validation questions for development, including a fixed 10,012-question diagnostic subset. The reported image-QA results are evaluated on all 12,578 questions in the balanced test-dev set. Answers are scored using strict string exact match under the archived evaluation script. The test-dev split does not provide the evidence boxes used by the training pipeline.

B.3 PHYSOBJECTS: PHYSICAL-PROPERTY QUESTION ANSWERING

Processed protocol. Our prepared PhysObjects protocol contains 18,740 training annotations, 1,032 development questions, and 868 evaluation questions over 417 full-scene images. The evaluation set contains five physical properties: material (356), transparency (303), contents (114), liquid containment (64), and sealed state (31). This constitutes a five-property, full-scene evaluation protocol rather than the complete original object-centric benchmark.

Answer selection. For each property, we use a fixed candidate answer set. The conditional token log-probabilities of each candidate are summed, and the candidate with the largest total score is selected. EOS probabilities are not included and no length normalization is applied.

Evaluation. We report micro accuracy over all evaluation questions as the main PhysObjects metric. The reported Q-JEPA checkpoint obtains 79.03% micro accuracy over the 868 questions.

B.4 STAR: FOUR-FRAME VIDEO QUESTION ANSWERING

Processed protocol. For video question answering, we use only the Interaction_T1 portion of STAR (Wu et al., 2021). The prepared training, development, and held-out splits contain 4,593, 501, and 1,111 questions, respectively. The held-out split contains 744 videos drawn from the official validation set, and videos are disjoint across the three data roles. The resulting numbers therefore correspond to our processed Interaction_T1 protocol rather than the complete STAR benchmark or its hidden test set.

Frame sampling. We evaluate two four-frame sampling strategies on the same 1,111 held-out questions. Raw-Uniform-4F uniformly samples four frames within the official clip interval and serves as the annotation-free setting. AG-Time-4F instead uses annotation-assisted keyframe times and is reported separately as a diagnostic setting that exploits additional temporal annotation information. The 35 held-out questions for which a valid training-style evidence crop cannot be constructed remain in the evaluation denominator.

Video prediction interface. The four sampled frames are independently processed by the frozen visual encoder. Their feature grids are concatenated before a single question-conditioned predictor readout. The predictor produces one 576-token target-view grid for the entire four-frame input, which is subsequently spatially merged into 144 tokens for the frozen language model. Candidate answers are introduced only after target-view prediction; the predictor itself receives the question but not the answer choices.

Evaluation. STAR answers are decoded with a fixed four-choice parser. Invalid or conflicting outputs are counted as incorrect, and greedy answer generation is limited to 16 tokens. Raw-Uniform-4F and AG-Time-4F are evaluated on exactly the same 1,111 questions and are reported separately.

C VISUALIZATION OF THE PREDICTED RESULT OF QGP

Shared PCA of source, evidence, and predicted features. The frozen encoder independently maps a full GQA-validation image and each annotation-derived evidence crop to a $24 \times 24 \times 1024$ feature grid after resizing the encoder input to 384×384. For one image with two questions, write the source grid as S, the two crop grids as G_1, G_2, and the corresponding predicted grids as P_1, P_2, each flattened to 576×1024. We normalize every token independently over its 1,024 channels, using

$$\mathcal{N}(z) = \frac{z - \mathrm{mean}(z)}{\sqrt{\mathrm{var}(z) + 10^{-5}}}, \qquad X = [\mathcal{N}(S); \mathcal{N}(G_1); \mathcal{N}(G_2)] \in \mathbb{R}^{1728 \times 1024}. \tag{10}$$

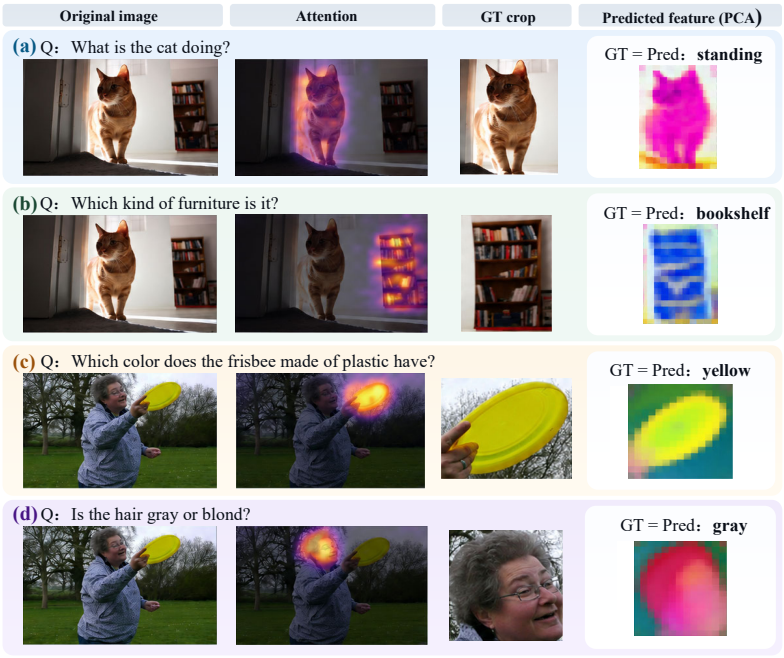

Figure 5: Visualization of question-relevant evidence. For each example, we show the original image, the cross-attention map from the question-guided predictor, the ground-truth crop, and the PCA visualization of the predicted target-view features. Q-JEPA successfully focuses on the correct visual evidence for diverse question types, including object state, object category, attribute recognition, and fine-grained human attributes.

PCA is fitted to the centered reference matrix X: its three leading eigenvectors form $U \in \mathbb{R}^{1024\times 3}$. The source is included once, and *neither prediction is used to fit the PCA*. All source, crop, and prediction tokens in this two-question group are projected using the same reference mean and U.

largest-absolute loading positive. For each component, the 0.5th and 99.5th percentiles of the projected *reference* tokens define one shared color range; clipped, normalized components 1–3 become the red, green, and blue channels. We fit a separate transform for the other image pair, so color values are comparable within a pair but not across the two pairs.

Display geometry and attention. The PCA output remains a 24×24 color grid. For display only, nearest- neighbor interpolation expands the source grid to the original-image pixel extent and the crop/prediction grids to the corresponding evidence-crop pixel extent. The original RGB crop itself retains its native pixels. In particular, the predicted-feature panel borrows the *ground-truth crop's display dimensions*; the predictor outputs neither that box nor its scale. The spatial heatmap is a separate diagnostic of predictor target-to-source *self-attention*, not Focus cross-attention over question words. It averages predictor layers 8–11, all 12 heads, and all 576 target queries, then conditions on the attention mass directed to source tokens before overlaying the 24×24 map on the original image.

D Limitations and Discussions

Our evaluations demonstrate that Q-JEPA can predict question-relevant visual representations for both images and short video clips, providing a compact and informative contents for LLM-based visual understanding. While a single predicted target view is effective for the evaluated image understanding tasks, it may compress temporally distributed evidence too aggressively when applied to video clips. Q-JEPA improves over the baseline under both video sampling strategies, but capturing multiple events and their temporal relationships may require richer representations. These observations motivate extending the current single-view formulation to multiple complementary target

views. Future work will explore question-conditioned predictive rollouts that generate a sequence of target-view representations to capture evolving visual evidence and support multi-step temporal reasoning.