Hybrid Token Compression for Vision-Language Models

Jusheng Zhang[1] Xiaoyang Guo[1] Kaitong Cai[1] Qinhan Lv[1] Yijia Fan[1]
Wenhao Chai[2] Jian Wang[3] Keze Wang[1]

[1]Sun Yat-sen University [2]Princeton University [3]Snap Inc.

Abstract

*Vision-language models (VLMs) have transformed multimodal reasoning, but feeding hundreds of visual patch tokens to LLMs incurs quadratic computational costs, straining memory and context windows. Traditional approaches face a trade-off: continuous compression dilutes high-level semantics like object identities, while discrete quantization loses granular details such as textures. We challenge this by introducing **HTC-VLM**, a hybrid framework that disentangles semantics and appearance through dual channels, i.e., a continuous pathway for fine-grained details via ViT patches and a discrete pathway for symbolic anchors using MGVQ quantization projected to four tokens. These are fused into a 580-token hybrid sequence and compressed to one token via a disentanglement attention mask and <voco> bottleneck, ensuring efficient, grounded representations. HTC-VLM achieves an average performance retention of **87.2%** across seven benchmarks (GQA, VQAv2, MMBench, MME, POPE, SEED-Bench, ScienceQA-Image), outperforming the leading continuous baseline at **81.0%** with a 580-to-1 compression ratio. Attention analyses show the compressed token prioritizes the discrete anchor, validating its semantic guidance.*
https://github.com/jushengzhang/HybridToken-VLM

1. Introduction

Vision–language models (VLMs) increasingly rely on large sets of patch-level visual tokens (N=576 for a single ViT image) to supply rich perceptual information to a large language model (LLM) [11, 17, 25, 28, 37, 45, 46, 53]. While effective, this dense coupling imposes a prohibitive quadratic attention cost $\mathcal{O}((N+L)^2)$, rapidly exhausting GPU memory and context budgets [9, 10, 36, 44, 46]. A natural question thus arises: *Can a VLM retain semantically useful visual information when the entire image is compressed to only a few tokens or even to a single one?*

Existing attempts have split into two directions that exhibit complementary failure modes [12, 18, 34, 40, 43, 48, 51, 54, 55]. *Continuous compression* projects the full patch sequence into a single dense vector, reducing latency but inevitably collapsing high-level semantics as mutual information $I(v_c; S)$ drops. Conversely, *full discretization* (e.g., VQ codebooks) preserves categorical semantics but discards fine-grained continuous cues (pose, texture, deformation), creating a granularity gap that limits detailed reasoning [17, 21, 51, 53, 58]. These behaviors are often interpreted as an unavoidable efficiency–fidelity trade-off.

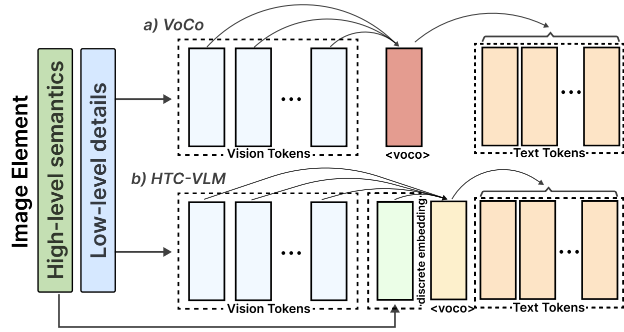

Figure 1. Vision-token compression. (a) VoCo-LLaMA collapses 576 patches into one <voco> token, losing semantic structure. (b) HTC-VLM adds 4 discrete semantic tokens and compresses all into one <voco> token, preserving semantics and visual detail.

We revisit this problem from a representational perspective. By compressing all 576 ViT tokens into a *single* latent, we expose a structural bottleneck: *A one-token continuous bottleneck cannot simultaneously encode discrete semantics and continuous visual details [36, 43].* This motivates us to examine how much structure must be preserved *before* compression. A key observation emerges: inserting a *minimal* set of discrete semantic anchors prior to the bottleneck restores the high-level scaffolding required by downstream reasoning, while continuous tokens retain complementary fine-grained detail. This leads to **HTC-VLM**, a hybrid compression architecture (Fig. 1) that prepends a small number of discrete semantic tokens to the continuous patch tokens and compresses them jointly into a single <voco> latent. Unlike VoCo-style methods, which compress only continuous tokens, our approach explicitly decomposes visual information into (i) **seman-**

tic anchors (discrete) and (ii) **detail carriers** (continuous) *before* fusion, following the principle that *single-token compression remains expressive only if semantics and details are disentangled prior to compression*. Empirically, HTC-VLM retains **87.2%** of full-model performance across seven benchmarks—outperforming the best continuous-compression baseline (**81.0%**) under the same 580-to-1 compression ratio. Attention analyses further show that the compressed latent selectively attends to the discrete anchors, validating their role as interpretable semantic carriers.

Our **main** contributions are three-fold: i) **Representational analysis of the bottleneck.** We identify the expressiveness gap in single-token visual compression and show that semantics and continuous details cannot be jointly preserved within a purely continuous latent; ii) **Hybrid semantic–detail decomposition.** We introduce a principled framework that injects a minimal number of discrete semantic anchors before compression, enabling disentangled fusion of discrete semantics and continuous appearance in the hybrid latent; iii) **A practical and scalable hybrid VLM architecture.** We instantiate this principle as HTC-VLM, achieving state-of-the-art retention (87.2%) under extreme 580-to-1 compression, with analyses showing the hybrid latent consistently preserves interpretable semantics and fine-grained cues.

2. Related Work

Vision-Language Models and Token Efficiency. Modern VLMs such as LLaVA [25], Qwen-VL [2], and GPT-4V [35] rely on hundreds of visual tokens (e.g., 576 from ViT-L/14) to enable strong multimodal alignment. This dense design, however, incurs quadratic attention cost and motivates reducing the visual token budget. Existing token-efficiency methods, including token merging [3, 5], patch dropping [30–32, 38], and redundancy-aware selection [24, 33], operate entirely in the *continuous* feature space. While they effectively reduce computation, these methods degrade rapidly under extreme compression (e.g., 1–4 tokens), where continuous features collapse and lose semantic structure.

Continuous vs. Discrete Compression. A complementary direction compresses vision features after the encoder. Continuous approaches such as pooling, attention aggregation, Q-Former [8], and VoCo-LLaMA [49] map all patches to a single dense embedding, but often suffer from *semantic dilution* when diverse patches are averaged into a unimodal vector. Conversely, discrete visual tokenizers (e.g., VQ-VAE [43], MoVQ [57], MGVQ [16]) produce compact and interpretable codes that preserve high-level semantics, but inevitably lose fine-grained appearance because quantization removes continuous variation. Neither paradigm preserves both high-level semantics and low-level details under a single-token bottleneck.

Hybrid Representation Learning. Recent studies suggest separating semantic and appearance information can improve multimodal representations, but existing approaches either require large token budgets or do not target extreme compression [12, 17, 21, 36, 43, 53, 58]. HTC-VLM differs by explicitly **disentangling** visual information into a discrete semantic channel and a continuous detail channel, and fusing them through a **single-token bottleneck** equipped with a disentanglement attention mask. This hybrid architecture simultaneously avoids semantic dilution and granularity loss, enabling one-token representations that remain structured, semantically stable, and detail-preserving.

3. Problem Formulation

The Dilemma of Visual Representation. A vision-language model (VLM) seeks to model the conditional distribution $p_\theta(Y \mid I, T)$, where an image I and a textual instruction T jointly generate a coherent textual response Y [4, 6]. In architectures like LLaVA, the image is decomposed into $N = 576$ patch embeddings using a pretrained vision encoder $\mathcal{E}_v$ (e.g., ViT-L/14) and projected into the LLM's embedding space via a trainable projector $\mathcal{P}_v$:

$$V = \{v_1, \ldots, v_N\} = \mathcal{P}_v\big(\mathcal{E}_v(I)\big), \quad V \in \mathbb{R}^{576 \times d_{\text{model}}}, \tag{1}$$

where $d_{\text{model}} = 4096$ corresponds to the embedding dimension of modern LLMs [26, 42, 46]. This representation facilitates multimodal alignment by mapping visual features into a shared semantic space, but its high dimensionality introduces profound computational and informational challenges, as detailed below.

3.1. The Scaling Imperative and Compression Objective

The LLM's self-attention mechanism [1, 20], defined as $A = \text{softmax}\big(\frac{QK^T}{\sqrt{d_k}}\big)V$, scales quadratically with the total sequence length, $\mathcal{O}((N + L)^2)$, where L is the text token count and d_k is the attention head dimension (typically $d_k = 64$). For $N = 576$ patch embeddings, this results in a per-layer memory complexity of $\Theta(N^2 d_{\text{model}})$, approximately 1.07×10^9 floating-point operations for $d_{\text{model}} = 4096$ on a single layer, rapidly exhausting GPU memory (e.g., 24GB VRAM) and saturating context windows (e.g., 4096 tokens). This quadratic bottleneck, exacerbated by multi-head attention across h heads, motivates compressing V to a single token, reducing the visual term to $\mathcal{O}(L^2)$ and yielding a theoretical speedup of $576^2 \approx 3.3 \times 10^5$ in attention computations. The optimal compressor $\mathcal{C}$ is thus formulated as:

$$\mathcal{C}^\star = \arg\min_{\mathcal{C}} \mathbb{E}_{(I,T,Y)\sim\mathcal{D}} \left[\mathcal{L}\big(Y, \mathrm{VLM}\big(T, \mathcal{C}\big(\mathcal{E}_v(I)\big)\big)\big)\right], \quad (2)$$

where $\mathcal{L}$ is typically the cross-entropy loss $\mathcal{L} = -\sum_{y\in Y} \log p_\theta(y \mid T, V_c)$, and $\mathcal{D}$ is the joint distribution over (I, T, Y). This optimization problem, however, reveals a critical trade-off: compressing to $|V_c| = 1$ risks diminishing the information content $I(V_c; Y)$, necessitating a balance between computational efficiency and representational fidelity, as explored next.

3.2. The Representation Dilemma: Theoretical and Practical Trade-offs

The compression challenge manifests in two paradigms, each with distinct limitations. *Continuous compression* transforms V into a single vector $v_c = \mathcal{C}_{\text{cont}}(V) \in \mathbb{R}^{d_{\text{model}}}$, often via global pooling or attention aggregation. Information-theoretically, this process reduces the entropy $H(V)$ to $H(v_c)$, where the mutual information $I(v_c; S)$ with high-level semantics S (e.g., object identities, spatial relations) diminishes. This *semantic dilution* arises because averaging convolves diverse patch distributions into a unimodal representation, lowering $H(v_c)$ below the threshold needed for disambiguation. For instance, averaging patches of a 'dog" and a cat" yields v_c with insufficient entropy to distinguish species, forcing the LLM to rely on ambiguous prior distributions, i.e., leading to errors in tasks like object classification (see 4.1.1 for our approach to mitigate this). In contrast, *discrete compression* via vector quantization maps I to indices $k = \arg\min_j \|f(I) - c_j\|_2^2$, where $\{c_j\}_{j=1}^K$ is a codebook of size K. This preserves interpretability by clustering semantic modes, but introduces a *granularity gap* due to quantization noise $\epsilon = f(I) - c_k$. The mutual information $I(k; D)$ with low-level details D (e.g., texture gradients, pose angles) is reduced, as continuous feature variance is discretized into discrete bins. Practically, this manifests when a Golden Retriever on grass and a Poodle on sand map to the same k, erasing contextual cues critical for fine-grained tasks like pose estimation or texture recognition (see 4.1.2 for our hybrid resolution). Neither paradigm optimizes the joint information $I(V_c; S, D)$ while minimizing redundancy $I(S; D \mid V_c)$, highlighting the need for a disentangled representation.

3.3. A Guiding Question for Disentangled Compression

The trade-off between semantic dilution and granularity gap suggests that a compact representation must disentangle S and D to maximize their joint contribution $I(V_c; S) + I(V_c; D)$ while reducing conditional dependence $I(S; D \mid V_c)$. This requires a representation where the compressed V_c acts as a sufficient statistic for both S and D, satisfying the Markov condition $S \perp D \mid V_c$. This insight frames our central inquiry:

> *How can we craft an ultra-compact visual representation that* ***disentangles*** *high-level, discrete semantics from low-level, continuous appearance, thereby escaping both semantic dilution and the granularity gap?*

This demands a hybrid framework where complementary channels encode orthogonal information, preserving diversity across S and D. HTC-VLM, introduced in Section 4, proposes a dual-channel architecture with a theoretically grounded bottleneck to achieve this disentanglement, as detailed below.

4. Method: Realizing Disentanglement with HTC-VLM

HTC-VLM tackles the guiding question from 3.3 by disentangling high-level semantics S (e.g., object categories, spatial layouts) and low-level details D (e.g., textures, poses) into distinct channels, fused through a disentanglement bottleneck that compresses the representation into a single token. This design emerged from the dilemma in 3.2: initial experiments compressing V (Eq. 1) to a single continuous token revealed a collapse in mutual information $I(v_c; S) \to 0$ due to variance loss, as the entropy $H(v_c) \ll H(V)$ failed to capture semantic diversity. Iterative trials with discrete augmentations, guided by information-theoretic metrics, identified that a single token from a vector quantizer (VQ) restored $I(v_d; S)$, inspiring HTC-VLM's architecture [12, 41]. This section elucidates the theoretical underpinnings, i.e., rooted in variational inference and attention dynamics, and practical implementation, validated by enhanced information retention (cross-referenced to experimental results).

> **Core Idea.** HTC-VLM disentangles visual information into a continuous channel for D and a discrete channel for S, fused via a disentanglement bottleneck to optimize the joint information $I(V_c; S, D)$ while minimizing redundancy $I(S; D \mid V_c)$.

4.1. Exploratory Decomposition and Channel Design

Our development began with a compression experiment on V, where reducing it to one token via averaging or attention pooling diminished $H(V)$ to $H(v_c)$, losing semantic structure as $I(v_c; S) \approx 0$ (3.2). To address this, we explored discrete representations, evaluating multiple VQ models. MGVQ [15], with its multi-group quantization (8 groups, 16384 codebook size, 16x downsampling), emerged as optimal due to its ability to cluster diverse semantic

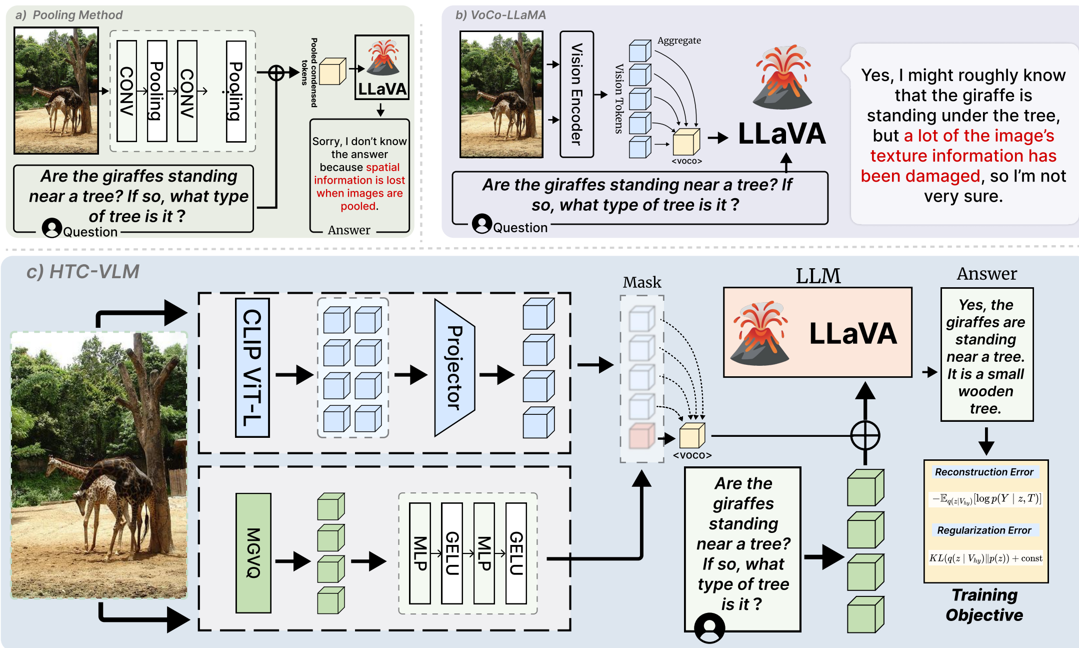

Figure 2. Comparison of visual token compression strategies. (a) **Pooling Method:** visual embeddings are averaged or pooled before being fused with text inputs. (b) **VoCo-LLaMA:** compresses 576 visual tokens into a single <voco> token. (c) **HTC-VLM (ours):** introduces a hybrid representation with a continuous channel (D) encoding 576 patch embeddings and a discrete channel (S) generating 4 semantic tokens via MGVQ. The hybrid sequence $[v_d; V]$ is compressed into a trainable <voco> token under the disentanglement mask M_{hy}, producing latent z that preserves both semantics and fine-grained details.

modes, decomposing I into dual channels that maximize $I(V; D) + I(q; S)$.

4.1.1. Continuous Channel: Encoding Low-Level Details D

To preserve the information content $I(V; D)$ and counter the granularity gap (3.2), we employ a pretrained vision encoder $\mathcal{E}_v$ (CLIP ViT-L/14) and a trainable linear projector $\mathcal{P}_v$ to generate a sequence of $N = 576$ patch embeddings:

$$V = \{v_i\}_{i=1}^{576} = \mathcal{P}_v\big(\mathcal{E}_v(I)\big), \quad v_i \in \mathbb{R}^{4096}, \tag{3}$$

where $\mathcal{P}_v : \mathbb{R}^{d_{\text{vision}}} \to \mathbb{R}^{4096}$ is a learned linear transformation aligning patch features with the LLM's embedding space. This high-dimensional manifold captures fine-grained details like texture gradients and pose variations, ensuring V retains a rich representation of D with entropy $H(V) \propto \log |\mathcal{M}_D|$, where $\mathcal{M}_D$ is the detail manifold.

4.1.2. Discrete Channel: Encoding High-Level Semantics S

To restore $I(q; S)$ and mitigate semantic dilution (3.2), we leverage MGVQ to quantize I into a feature vector $q \in \mathbb{R}^{14112}$, reflecting its multi-group structure. This is projected to a sequence of discrete anchor tokens $v_d \in \mathbb{R}^{4\times 4096}$ via a two-layer MLP $\mathcal{P}_d$ with GELU activations:

$$q = \mathcal{Q}(I), \quad v_d = \mathcal{P}_d(q) = \text{GELU}\big(W_2 \cdot \text{GELU}\big(W_1 \cdot q\big)\big), \tag{4}$$

where $W_1 \in \mathbb{R}^{16384\times 14112}$ and $W_2 \in \mathbb{R}^{16384\times 16384}$ are weight matrices, and $\text{GELU}(x) = x\Phi(x)$ (where Φ is the Gaussian CDF) introduces non-linearity. MGVQ's quantization minimizes the reconstruction error $\mathbb{E}\|I - \mathcal{Q}^{-1}(q)\|_2^2$, clustering S into discrete modes (e.g., 'dog on grass'), with v_d serving as a low-dimensional anchor that preserves $I(v_d; S) \approx H(S)$ under codebook constraints.

4.2. Fusion and Disentanglement Bottleneck: Theoretical Framework

The channels are fused by prepending v_d to V, forming a hybrid sequence:

$$V_{hy} = [v_d; V] \in \mathbb{R}^{580\times 4096}, \tag{5}$$

followed by a trainable <voco> token. The Disentanglement Attention Mask M_{hy} is defined over the full input $X = [V_{hy}; \texttt{<voco>}; W]$, where W are text embeddings:

$$M_{hy}(i,j) = \begin{cases} 0, & \text{if } x_i \in W \text{ and } x_j \in V_{hy}, \\ -\infty, & \text{if } x_i, x_j \in V_{hy} \text{ and } i \neq j \text{ (self-attention within } V_{hy}), \\ 1, & \text{otherwise}, \end{cases} \tag{6}$$

This mask ensures text attends only to <voco>, while <voco> integrates V_{hy}. Theoretically, this bottleneck approximates a variational autoencoder (VAE), where <voco> represents a latent variable z with posterior $p(z \mid V_{hy})$. The objective is to minimize the Kullback-Leibler divergence $KL(p(V_{hy} \mid z) \| p(V_{hy}))$, guided by v_d to disentangle S and D. The evidence lower bound (ELBO) for this process is:

$$\log p(Y \mid T, I) \geq \mathbb{E}_{q(z|V_{hy})} [\log p(Y \mid z, T)] - KL(q(z \mid V_{hy}) \| p(z)), \tag{7}$$

where $q(z \mid V_{hy}) = p(\texttt{<voco>} \mid V_{hy}; \theta)$ is learned, and $p(z)$ is a prior (e.g., $\mathcal{N}(0, I)$). The compression ratio of 580-to-1 is achieved by optimizing z to maximize the mutual information:

$$\mathcal{I}(z; V_{hy}) = \mathbb{E}_{p(V_{hy})} \left[\log \frac{p(V_{hy}, z)}{p(V_{hy}) p(z)} \right], \tag{8}$$

where v_d biases z toward S, and V contributes D, reducing $I(S; D \mid z)$ via M_{hy}'s constraint.

Disentanglement Bottleneck. The <voco> token compresses V_{hy} into a latent z, with v_d enforcing disentangled encoding of S and D via M_{hy}, optimizing $\mathcal{I}(z; S) + \mathcal{I}(z; D)$.

4.3. Training Objective and Optimization Dynamics

The training objective is the expected autoregressive loss:

$$\mathcal{L}_{\text{HTC}} = -\mathbb{E}_{p(I,T,Y)} \left[\textstyle\sum_{i=1}^{|Y|} \log p_\theta(y_i \mid y_{<i}, \texttt{<voco>}, T; M_{hy}) \right], \tag{9}$$

where M_{hy} shapes the gradient flow. This loss can be decomposed into a variational lower bound, aligning with the ELBO (Eq. 7):

$$\mathcal{L}_{\text{HTC}} \approx -\mathbb{E}_{q(z|V_{hy})} [\log p(Y \mid z, T)] + KL(q(z \mid V_{hy}) \| p(z)) + \text{const}, \tag{10}$$

where the first term is the reconstruction error, and the second term regularizes z. The mask M_{hy} constrains $q(z \mid V_{hy})$ to depend on <voco>, with v_d acting as a prior anchor for S. Gradient dynamics reveal that $\frac{\partial \mathcal{L}}{\partial v_d}$ enhances semantic clustering (e.g., maximizing $I(v_d; S)$), while $\frac{\partial \mathcal{L}}{\partial V}$ refines D's variance, achieving a disentangled latent space. This optimization leverages the bottleneck to outperform single-channel baselines by preserving $I(\texttt{<voco>}; S, D)$, as validated in subsequent experiments.

Training Pressure Redirected. The mask M_{hy} reroutes gradients to enforce a disentangled latent z, optimizing $I(\texttt{<voco>}; S, D)$ via variational inference.

Algorithm 1 HTC-VLM Forward Pass

Require: Image I, Text T, components ($\mathcal{E}_v$, $\mathcal{P}_v$, $\mathcal{Q}$, $\mathcal{P}_d$, $\mathcal{E}_t$, $\mathcal{L}_{\text{LLM}}$)

$V \leftarrow \mathcal{P}_v(\mathcal{E}_v(I))$ ▷ Generate 576 patch embeddings
$q \leftarrow \mathcal{Q}(I)$ ▷ MGVQ quantizes to 14112 features
$v_d \leftarrow \mathcal{P}_d(q)$ ▷ MLP projection to discrete embedding
$V_{hy} \leftarrow [v_d; V]$ ▷ Construct 580-token hybrid
$W \leftarrow \mathcal{E}_t(T)$ ▷ Encode text
$X \leftarrow [V_{hy}; \texttt{<voco>}; W]$ ▷ Integrate with <voco> token
$M_{hy} \leftarrow$ CreateDisentanglementAttentionMask(X) ▷ Apply disentanglement mask
Logits $\leftarrow \mathcal{L}_{\text{LLM}}(X, M_{hy})$ ▷ Compute logits
return Logits

5. Experiments

5.1. Experimental Setup

To ensure a fair and direct comparison, our experimental setup, including the training data, architectural backbone, and evaluation protocols, strictly follows that of VoCo-LLaMA [49]. We evaluate our model, HTC-VLM, on a comprehensive suite of seven popular visual understanding benchmarks: GQA [14], VQAv2 [13], MMBench [27], MME [50], POPE [22], SEED-Bench [19], and ScienceQA (Image) [29]. The performance of the baseline models, including the *Upper Bound* (the original VLM without compression), the *Lower Bound* (compression without specific training), Q-Former [20], and Avg. Pool [23], are directly cited from the VoCo-LLaMA [49] study to provide a consistent and comprehensive frame of reference. For fairness, we reproduce the results of VoCo-LLaMA [49] under the same setting. Additionally, we introduce the VQA$^{\textit{text}}$ [39] and MMVet [52] benchmarks to further evaluate the model's performance on text and visual understanding tasks. For a comprehensive comparison, we compare HTC-VLM with methods such as ToMe [3], FastV [7], PDrop [47], and SparseVLM[56].

5.2. Experiment Results

The main results, presented in Table 1, demonstrate that our proposed Disentangled Hybrid Visual Representation is highly effective. HTC-VLM consistently surpasses all previous vision compression baselines, including Q-Former [20], Average Pooling [23], and the state-of-the-art VoCo-LLaMA [49] model. Different from VoCo-LLaMA, which directly compresses 576 image patch tokens into a single voco token, our method first augments the patch sequence with four additional discrete semantic tokens generated by a vector quantizer, and then compresses the $4 + 576$ tokens into a single voco token. This design explicitly supplements high-level semantics before compression. As

Table 1. Comparison of HTC-VLM with previous vision compression approaches on common visual understanding benchmarks. All methods reduce 576 tokens to one. "Avg." refers to the average of per-benchmark performance retention rates, calculated as (Result - Lower Bound) / (Upper Bound - Lower Bound) for each benchmark. Our hybrid approach attains the best results..

Model	Tokens	GQA	VQA^{v2}	MMBench	MME^{P}	POPE	SEED	SQA^{I}	Avg. (%)
Upper Bound	576	61.1 100%	77.7 100%	64.0 100%	1487.2 100%	85.0 100%	57.9 100%	66.5 100%	- 100%
Q-Former [20]	1	51.1 57.3%	63.4 70.5%	51.7 53.2%	1079.7 75.2%	77.3 49.0%	47.2 34.5%	62.7 60.8%	- 57.2%
Avg. Pool [23]	1	52.9 65.0%	65.0 79.6%	55.5 68.1%	1210.3 81.0%	79.1 63.8%	50.3 25.8%	62.2 65.2%	- 64.1%
VoCo-LLaMA [49]	1	57.4 84.2%	71.8 83.8%	57.9 85.4%	1241.4 71.7%	81.5 88.7%	48.8 56.7%	66.3 96.6%	- 81.0%
HTC-VLM (ours)	**1 (hybrid)**	**57.6** **85.0%**	**72.4** **85.5%**	**60.0** **90.4%**	**1265.2** **74.5%**	**82.8** **92.9%**	**49.8** **61.4%**	**67.7** **120.7%**	- **87.2%**
Lower Bound	1	37.7 0%	41.2 0%	22.3 0%	617.3 0%	53.9 0%	36.9 0%	60.7 0%	- 0%

a result, HTC-VLM achieves an average performance retention rate of 87.2%, establishing a new benchmark for highly compressed visual representations. This strongly supports our central hypothesis: the discrete token effectively recovers structured, high-level semantic concepts that are inevitably lost during continuous compression, pushing the model's performance closer to the uncompressed Upper Bound. While purely continuous compression like VoCo-LLaMA [49] is powerful, our hybrid strategy demonstrates that explicitly disentangling and injecting semantic information prior to compression is a superior solution.

Table 2 shows that across different token budgets (192, 128, 64 tokens), HTC-VLM consistently maintains performance levels close to or above competing compression methods. While it does not always achieve the absolute highest average retention at 192 or 128 tokens, it remains competitive with ToMe [3], FastV [7], PDrop [47], and SparseVLM[56], and notably surpasses all of them at the 64-token setting, retaining 89.8% of the original performance. The per-benchmark results indicate that HTC-VLM performs well on both semantic-heavy tasks, such as GQA and VQA$^{\text{text}}$, and detail-intensive tasks, like MME and POPE, reflecting the effectiveness of the hybrid design in preserving both high-level semantic information and fine-grained visual details under strong compression.

5.3. Analysis of Semantic Decoupling and Compression

Representation Probing To quantitatively validate our core hypothesis, i.e., HTC-VLM successfully decouples high-level semantics (S) from low-level details (D) through its hybrid architecture, and ultimately integrates them effectively in the compression bottleneck, we design a comprehensive Representation Probing experiment. This experiment aims to assess the model's internal representations' ability to encode specific types of information. Specifically, we extract three different 4096-dimensional intermediate representations from the trained and frozen HTC-VLM model for probing: In this approach, we consider three types of representations: 1) Discrete semantic representation (v_d): The four discrete semantic representations generated by the discrete channels are averaged using pooling, followed by voco encoding; 2) Continuous detail representation (V^-): The 576 continuous image block labels are averaged using pooling, followed by voco encoding; 3) Compressed hybrid representation (z_{voco}): The final vector output by the <voco> token. To enable targeted evaluation of different levels of information, we selected approximately 200 samples from each of the VQAv2 and GQA visual question answering datasets, focusing on samples with semantic features (S class) and detail features (D class), totaling 776 samples. Each class of samples is accompanied by a closed label set, allowing for representation decoding using a linear classification head.

Table 3 shows that the compressed hybrid representation (z_{voco}) achieves the best performance in both task categories (30.70% for the detail task and 26.67% for the semantic task), which validates that the voco compression mechanism in HTC-VLM can retain both semantic and detail information within a single token, achieving effective information fusion. Although the discrete semantic representation (v_d) slightly underperforms the continuous detail representation (V^-) in overall accuracy, this phenomenon is consistent with its design objective: v_d is formed by only 4 tokens, carrying much less information than V^-, which is aggregated from 576 tokens. Therefore, its slightly weaker performance in closed-set classification is expected. However, v_d demonstrates stable decodability in both semantic

Table 2. **Comparison of token compression methods under varying token budgets.** Vanilla, with 576 visual tokens, serves as the upper bound for each benchmark. The table reports per-benchmark results and average performance retention (%) for different token lengths (192, 128, 64), highlighting how compression affects performance across tasks.

Method	GQA	MMB	MME	POPE	SQA^I	SEED	VQA^{text}	MMVet	Avg.(%)
Upper Bound, 576 Tokens									
Vanilla	61.9 100%	64.6 100%	1864 100%	85.9 100%	69.5 100%	60.3 100%	58.3 100%	30.9 100%	100%
192 Tokens									
ToMe [3]	52.4 84.7%	53.3 82.4%	1343 72.1%	62.8 73.1%	59.6 85.8%	50.9 84.4%	49.1 84.4%	27.2 88.0%	88.9%
FastV [7]	52.6 85.0%	61.0 94.4%	1605 86.1%	64.8 75.4%	69.1 99.4%	52.1 86.4%	52.5 90.1%	26.7 86.4%	87.9%
PDrop [47]	57.1 92.2%	63.2 97.8%	1766 94.7%	82.3 95.8%	70.2 101.0%	54.7 90.7%	56.1 96.2%	30.5 98.7%	95.9%
SparseVLM[56]	59.5 96.1%	64.1 99.2%	1787 95.9%	85.3 99.3%	68.7 98.8%	58.7 97.3%	57.8 99.1%	33.1 107.1%	99.1%
VoCo-LLaMA [49]	61.4 99.2%	56.3 87.2%	1596 85.6%	84.5 98.4%	66.6 95.8%	51.1 84.7%	50.6 86.8%	27.2 88.0%	90.7%
HTC-VLM	62.4 100.8%	59.3 91.8%	1687 90.5%	85.1 99.1%	66.8 96.1%	52.8 87.6%	52.0 89.2%	30.4 98.4%	94.2%
128 Tokens									
ToMe [3]	52.4 84.7%	53.3 82.4%	1343 72.1%	62.8 73.1%	59.6 85.8%	50.9 84.4%	49.1 84.4%	27.2 88.0%	81.9%
FastV [7]	49.6 80.1%	56.1 86.8%	1490 79.9%	53.4 62.2%	68.6 98.7%	48.1 79.8%	50.5 86.6%	26.3 85.1%	82.4%
PDrop [47]	56.0 90.5%	61.1 95.4%	1664 89.3%	82.3 95.8%	69.9 100.6%	53.3 88.4%	55.1 94.5%	30.8 99.7%	94.3%
SparseVLM[56]	58.4 94.3%	64.5 99.8%	1746 93.7%	85.0 99.0%	68.6 98.7%	58.2 96.5%	56.7 97.3%	29.0 93.9%	96.7%
VoCo-LLaMA [49]	61.5 99.4%	56.4 87.3%	1640 88.0%	84.5 98.4%	66.6 95.8%	50.5 83.7%	51.7 88.7%	29.7 96.1%	92.2%
HTC-VLM	61.8 99.8%	60.5 93.7%	1629 87.4%	84.5 98.4%	67.9 97.7%	52.4 86.9%	51.9 89.0%	30.2 97.7%	93.8%
64 Tokens									
ToMe [3]	48.6 78.5%	43.7 67.5%	1138 61.1%	52.5 61.1%	50.0 71.9%	44.0 73.0%	45.3 77.8%	24.1 78.0%	71.1%
FastV [7]	46.1 74.5%	47.2 73.1%	1255 67.3%	38.2 44.5%	68.7 98.8%	43.7 72.5%	47.8 82.0%	19.6 63.4%	72.0%
PDrop [47]	41.9 67.7%	33.3 51.6%	1092 58.6%	55.9 65.1%	69.2 99.6%	40.0 66.3%	45.9 78.7%	30.7 99.4%	73.4%
SparseVLM[56]	53.8 86.9%	60.1 93.0%	1589 85.2%	77.5 90.2%	69.8 100.4%	52.2 86.6%	53.4 91.6%	24.9 80.6%	89.3%
VoCo-LLaMA [49]	60.2 97.3%	57.7 89.3%	1623 87.1%	83.4 97.1%	67.7 97.4%	50.0 82.9%	51.1 87.7%	24.1 78.0%	89.6%
HTC-VLM	60.3 97.4%	59.1 91.5%	1618 86.8%	83.6 97.3%	66.3 95.4%	50.6 83.9%	50.7 87.0%	24.4 79.0%	89.8%

Task Type	$z_{\textbf{voco}}$	v_d	$\bar{V}$
D-10 (Detail)	30.70%	25.44%	27.19%
S-10 (Semantic)	26.67%	20.83%	26.67%

Table 3. Probing Top-1 accuracy (%) of discrete (v_d), continuous ($\bar{V}$), and hybrid (z_{voco}) representations on semantic (S-10) and detail (D-10) tasks.

and detail tasks (20.83% and 25.44%, respectively), indicating that although compact, it still possesses the ability to jointly encode both semantic and detail information, reflecting higher information compression efficiency.

Furthermore, V^- slightly outperforms v_d in D-10 (27.19% vs. 25.44%), which aligns with its role as the lower-level visual path. In contrast, V^- and z_{voco} show similar performance in S-10 (26.67%), suggesting that semantic features do permeate through the continuous channel.

Overall, these results validate the semantic decoupling hypothesis of HTC-VLM: the discrete channel tends to encode high-level semantics, the continuous channel retains low-level details, and the voco compression mechanism strikes a balance and fusion between the two.

5.3.1. Attention Analysis

To evaluate how HTC-VLM retains high-level semantic information while maintaining compact visual representations, we analyzed the attention patterns of the compressed <voco> token and the performance of different compression strategies. On a subset of 16 test samples from the MME benchmark dataset, we visualized the attention distribution of the <voco> token over the input tokens. In

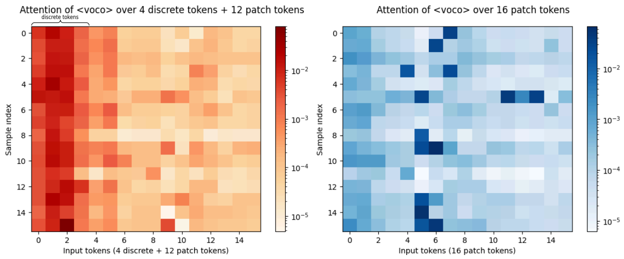

Figure 3. Comparison of compression strategies and their effect on visual token attention. **Left:** Attention heatmap of the <voco> token in HTC-VLM over 4 discrete semantic token plus the first 12 image patch tokens for 16 test samples from the MME benchmark. **Right:** Attention heatmap of the <voco> token in the original VoCo-LLaMA [49] model over the first 16 image patch tokens for the same 16 test samples.

HTC-VLM, the input sequence consists of "4 discrete semantic tokens" and the "first 12 continuous image block tokens from the original 576 image block tokens", with the discrete semantic tokens located at the front of the sequence. In the generated heatmaps, each row corresponds to a test sample, and each column corresponds to an input token (the first four columns are discrete tokens, and the following 12 columns are image block tokens). Figure 3 (left) shows that for the first four columns corresponding to the discrete tokens, the attention values are consistently much higher than those of most subsequent image block tokens. This indicates that the <voco> token effectively utilizes the high-level semantic information encoded in the discrete tokens.

For comparison, Figure 3 (right) visualizes the attention distribution of the original VoCo-LLaMA [49] model (pure continuous compression). In this model, the <voco> token distributes attention more evenly across the image block tokens, lacking the focused semantic guidance provided by the discrete tokens. This contrast highlights the role of "discrete semantic anchors" in guiding the compression process and retaining critical high-level information.

5.3.2. Ablation Experiment

To validate the design of HTC-VLM and trace its performance gains, we conduct ablations over three components: (i) hybrid vs. non-hybrid compression, (ii) the number of discrete tokens, and (iii) fusion strategy. **Hybrid vs. Non-hybrid.** Our key hypothesis is that hybrid representations outperform purely continuous or purely discrete ones. We compare: (i) **HTC-VLM**: compressing 576 continuous tokens with 4 discrete semantic tokens into one <voco> token; (ii) **Continuous-only**: VoCo-LLaMA-style compression of 576 continuous tokens; (iii) **Discrete-only**: compressing the 441 MGVQ-discrete tokens. As Table 4 shows, continuous-only retains 81.0%, while discrete-only drops to ∼33.0%. HTC-VLM reaches 87.2%, confirming the advantage of hybrid representations. **Number of Discrete Tokens (N_d).** We then vary the number of discrete tokens. Using one token (N_d=1) yields 83.9%; two (N_d=2) improves to 84.9%; the best performance is at N_d=4 with 87.2%. Increasing to eight (N_d=8) slightly degrades performance (84.6%), indicating that excessive discrete tokens introduce redundancy. Thus, a small set of semantic tokens strikes the best balance of expressiveness and compactness. **Fusion Strategy.** Finally, we compare ways to fuse discrete and continuous tokens. The default *pre-fusion* (placing discrete tokens first) performs best. *Post-fusion* and *mean fusion* alternatives both underperform, with mean fusion showing the largest drop due to dilution of the semantic guiding signal. This verifies the design choice of positioning discrete tokens as a semantic anchor.

Table 4. Ablation study on different configurations of HTC-VLM. Performance retention (%) is reported relative to the full model.

Configuration	Retention (%)
Hybrid vs. Non-Hybrid	
Discrete-Only (441 tokens)	33.3
Continuous-Only (576 tokens)	81.0
HTC-VLM (4 + 576 tokens)	**87.2**
Number of Discrete Tokens (N_d)	
$N_d = 1$	83.9
$N_d = 2$	84.9
$N_d = 4$ **(ours)**	**87.2**
$N_d = 8$	84.6
Fusion Strategy	
Pre-fusion (ours)	**87.2**
Post-fusion	84.6
Mean fusion	84.6

6. Limitations and Conclusion

Limitations. HTC-VLM focuses on single-image compression and has not yet explored multi-image or video settings, where temporal cues may interact with the hybrid token design. In addition, the discrete semantic anchors are produced by an external VQ tokenizer; jointly learning them with the VLM may further improve adaptability.**Conclusion.**This work presents HTC-VLM, a disentangled hybrid compression framework that injects a small set of discrete semantic tokens before compressing both semantic and continuous visual information into a single <voco> token. By preserving high-level structure and low-level details, HTC-VLM achieves state-of-the-art performance retention under a 580-to-1 compression ratio. Our analysis and ablations confirm that disentangling semantics and details is key to stable, efficient visual representations. We hope this study inspires future work on scalable and interpretable multimodal token compression.

7. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62276283, in part by the China Meteorological Administration's Science and Technology Project under Grant CMAJBGS202517, in part by Guangdong-Hong Kong-Macao Greater Bay Area Meteorological Technology Collaborative Research Project under Grant GHMA2024Z04, in part by Fundamental Research Funds for the Central Universities, Sun Yat-sen University under Grant 23hytd006, in part by Guangdong Provincial High-Level Young Talent Program under Grant RL2024-151-2-11, and in part by the Key Development Project of the Artificial Intelligence Institute, Sun Yat-sen University.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 2

[3] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 2, 5, 6, 7

[4] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey, 2024. 2

[5] Qingqing Cao, Bhargavi Paranjape, and Hannaneh Hajishirzi. Pumer: Pruning and merging tokens for efficient vision language models, 2023. 2

[6] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023. 2

[7] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 5, 6, 7

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2

[9] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. 1

[10] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022. 1

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1

[12] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878, 2021. 1, 2, 3

[13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 5

[14] 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*, pages 6700–6709, 2019. 5

[15] Mingkai Jia, Wei Yin, Xiaotao Hu, Jiaxin Guo, Xiaoyang Guo, Qian Zhang, Xiao-Xiao Long, and Ping Tan. Mgvq: Could vq-vae beat vae? a generalizable tokenizer with multi-group quantization. *arXiv preprint arXiv:2507.07997*, 2025. 3

[16] Mingkai Jia, Wei Yin, Xiaotao Hu, Jiaxin Guo, Xiaoyang Guo, Qian Zhang, Xiao-Xiao Long, and Ping Tan. Mgvq: Could vq-vae beat vae? a generalizable tokenizer with multi-group quantization, 2025. 2

[17] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Comput. Surv.*, 54(10s), 2022. 1, 2

[18] Yuri Kuratov, Mikhail Arkhipov, Aydar Bulatov, and Mikhail Burtsev. Cramming 1568 tokens into a single vector and back again: Exploring the limits of embedding space capacity. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19323–19339, Vienna, Austria, 2025. Association for Computational Linguistics. 1

[19] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 5

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *In-*

ternational conference on machine learning, pages 19730–19742. PMLR, 2023. 2, 5, 6

[21] Jindong Li, Yali Fu, Jiahong Liu, Linxiao Cao, Wei Ji, Menglin Yang, Irwin King, and Ming-Hsuan Yang. Discrete tokenization for multimodal llms: A comprehensive survey, 2025. 1, 2

[22] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 5

[23] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 5, 6

[24] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations, 2022. 2

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 1, 2

[26] 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*, pages 26296–26306, 2024. 2

[27] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 5

[28] Yang Liu, Yao Zhang, Yixin Wang, Feng Hou, Jin Yuan, Jiang Tian, Yang Zhang, Zhongchao Shi, Jianping Fan, and Zhiqiang He. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6): 7478–7498, 2024. 1

[29] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 5

[30] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 2

[31] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024.

[32] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 2

[33] Yue Ma, Zexuan Yan, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, et al. Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. *arXiv preprint arXiv:2509.16630*, 2025. 2

[34] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers for image classification. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 12–21, 2023. 1

[35] OpenAI. Gpt-4o system card, 2024. 2

[36] Lorenzo Papa, Paolo Russo, Irene Amerini, and Luping Zhou. A survey on efficient vision transformers: Algorithms, techniques, and performance benchmarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 7682–7700, 2024. 1, 2

[37] 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, 2021. 1

[38] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification, 2021. 2

[39] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 5

[40] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6), 2022. 1

[41] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. 3

[42] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 2

[43] Aaron van den Oord, Oriol Vinyals, and Koray

Kavukcuoglu. Neural discrete representation learning, 2018. 1, 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 1

[45] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020. 1

[46] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey, 2023. 1, 2

[47] Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024. 5, 6, 7

[48] Chenyu Yang, Xizhou Zhu, Jinguo Zhu, Weijie Su, Junjie Wang, Xuan Dong, Wenhai Wang, Lewei Lu, Bin Li, Jie Zhou, Yu Qiao, and Jifeng Dai. Vision model pre-training on interleaved image-text data via latent compression learning. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024. 1

[49] Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, and Yansong Tang. Voco-llama: Towards vision compression with large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29836–29846, 2025. 2, 5, 6, 7, 8

[50] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12): nwae403, 2024. 5

[51] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation, 2024. 1

[52] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 5

[53] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024. 1, 2

[54] Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. CF-VLM: Counterfactual vision-language fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[55] Jusheng Zhang, Kaitong Cai, Qinglin Zeng, Ningyuan Liu, Stephen Fan, Ziliang Chen, and Keze Wang. Failure-driven workflow refinement, 2025. 1

[56] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024. 5, 6, 7

[57] Chuanxia Zheng, Long Tung Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation, 2022. 2

[58] Yifan Zhong, Fengshuo Bai, Shaofei Cai, Xuchuan Huang, Zhang Chen, Xiaowei Zhang, Yuanfei Wang, Shaoyang Guo, Tianrui Guan, Ka Nam Lui, Zhiquan Qi, Yitao Liang, Yuanpei Chen, and Yaodong Yang. A survey on vision-language-action models: An action tokenization perspective, 2025. 1, 2