GazaVHR: AI-Driven Legally Grounded Conflict Harm Documentation

Nesibe Şebnem Paluluoğlu [* 1]   **Dilara Zeynep Gürer** [* 1]   **Muhammed Furkan Akıncı** [1]   **Mustafa Taha Koçyiğit** [1]

Abstract

We present GazaVHR, a vision-language model (VLM)-annotated dataset for fine-grained analysis of potential human rights violations in Gaza conflict imagery. Sourced from 145,662 conflict-related tweets, our pipeline integrates vision-language models, vision encoders, and semantic clustering to generate structured annotations with minimal manual intervention. Beginning with 176,731 raw images, a multi-stage filtering (content rules, deduplication, semantic clustering) identifies 13,834 visually unique instances that are most likely conflict-relevant. To ensure legal relevance, we align results with the Kanıt (Evidence) dataset: 231 expert-curated images grounded in the Rome Statute of the International Criminal Court (ICC Articles 5–8). This framework refines the dataset to 4,603 high-confidence images likely indicative of conflict-related harm. While our work highlights AI's potential to systematize human rights documentation at scale, we acknowledge limitations in reduced manual oversight and biases inherent to LLM-based annotation and hashtag-driven social media data.

1. Introduction

Monitoring human rights violations (HRVs) in conflict zones like Gaza is challenging due to limited on-the-ground access. Social media has become a crucial source of real-time, user-generated visual content, yet the lack of structured, annotated datasets hinders research and advocacy.

We introduce **GazaVHR** (Visual Human Rights Violations Dataset), a curated dataset of images from the Gaza conflict, designed to support machine learning research in visual HRV detection. To address issues of scale, bias, and anno-

tation ethics, we propose an automated, modular pipeline using promptable vision-language models (VLMs) with targeted filtering and deduplication.

Our main contributions are (1) A VLM-assisted annotation framework for efficient and ethical curation of conflict imagery. (2) The release of **GazaVHR**, a filtered and categorized dataset for visual HRV and humanitarian analysis. (3) A preliminary dataset analysis revealing latent themes and biases, with future directions for deeper auditing.

By integrating scalable AI with ethical filtering, we aim to advance the intersection of machine learning and human rights, offering a foundation for future conflict and humanitarian research.

2. Related Work

Advances in AI and social media analysis have enabled researchers to extract insights from publicly shared content, particularly in crisis response, public sentiment, and human rights. This section reviews prior work on two fronts: leveraging user-generated content (UGC) for social understanding, and detecting human rights violations (HRVs) using machine learning and computer vision.

2.1. User-Generated Content

UGC is a valuable resource for understanding public behavior online. On TikTok, hashtags like #teenmentalhealth show how adolescents and therapists discuss mental health via personal stories and humor, though often lacking scientific grounding (Lau et al., 2025). On Twitter, sentiment analysis and embedding-based models have been used to track well-being during crises, with modern NLP methods outperforming traditional tools (Wang et al., 2025). In tourism, online reviews help uncover drivers of user satisfaction, highlighting the utility of personal narratives (Meneses et al., 2025).

Building on this work, we focus on visual UGC from conflict zones to enable automated HRV analysis.

2.2. Human Rights Violation Detection

Machine learning and computer vision have shown promise in detecting HRVs using diverse sources like social media,

---

[1]The Institute for Data Science and Artificial Intelligence, Bogazici University, Istanbul, Turkey. Correspondence to: Nesibe Şebnem Paluluoğlu <sebnem.paluluoglu@bogazici.edu.tr>, Dilara Zeynep Gürer <dilara.gurer@std.bogazici.edu.tr>.
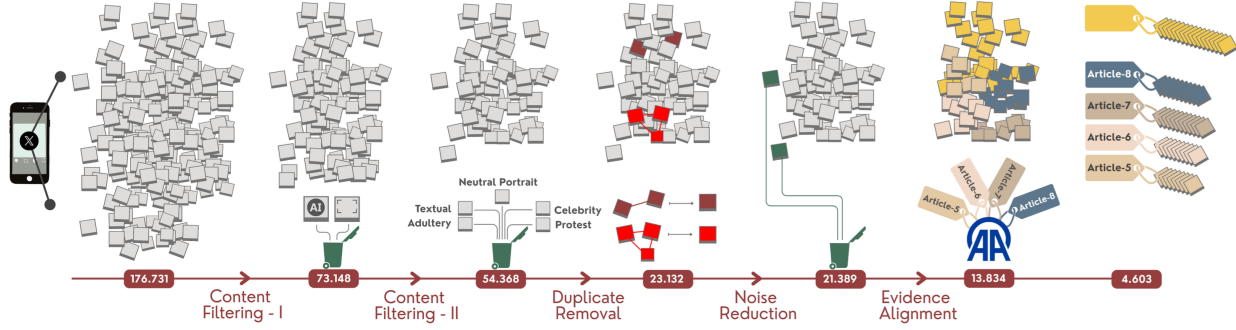
Figure 1. The GazaVHR dataset creation pipeline: (1) **Content Filtering** categorizes images into media types (AI-generated, screenshots) and thematic classes (protest, celebrity, textual, adultery, neutral portrait); (2) **Duplicate Removal** removes exact and near-identical copies; (3) **Noise Reduction** identifies and removes irrelevant content; (4) **Evidence Alignment** aligns images with the Kanıt dataset. This multi-stage process refines social media imagery into legally relevant evidence of potential human rights violations.

web images, and satellite data. Social platforms such as Twitter and Telegram have been used to classify HRV-related content with high accuracy (up to 98%) using ML and deep learning models (Pilankar, 2022; Nemkova et al., 2023).

Several benchmark datasets, including HRUN and HRA, support visual HRV recognition. Studies show that CNNs, especially with transfer learning, perform well (e.g., 88.10% mAP) in identifying abuse-related imagery, with object- and scene-level features enhancing results (Kalliatakis et al., 2017a; 2019; Kalliatakis, 2019). Satellite imagery has also been used to verify conflict reports. In Sudan, feature extraction helped detect burned villages, complementing manual satellite analysis (O'Connell & Young, 2014). Citizen-shared images and social media posts are increasingly central to HRV documentation. When paired with domain-adapted computer vision tools and large datasets, these sources provide critical real-time evidence (Kalliatakis et al., 2017b).

Overall, integrating deep learning, natural language processing, and remote sensing offers a robust framework for detecting human rights violations (HRVs). A cornerstone of our approach is the use of powerful vision-language models (VLMs), which enable high-precision zero-shot annotation and categorization of conflict imagery without task-specific training. The next section outlines our data collection and filtering strategy.

3. Methodology

3.1. Data Collection

3.1.1. SOCIAL MEDIA DATA ACQUISITION

To examine visual user-generated content (UGC) related to the Gaza conflict, we collected image-containing tweets posted between October 7, 2023, and March 10, 2025. The collection process began by identifying trending topics using

the publicly accessible Twitter Trending Archive (Twitter Trending Archive) which tracks daily top trending hashtags globally and by region. A Python script automatically extracted the top 5 trending hashtags daily from global and Turkey-specific lists in the archive.

Two researchers then manually selected Gaza-conflict-relevant hashtags using: (i) direct keyword relevance, (ii) contextual alignment with conflict events, and (iii) exclusion of unrelated/spam tags. Inter-rater agreement was high and discrepancies were resolved by a third reviewer.

We stratified hashtags by popularity. We sampled 4,000 image-containing tweets from top 15 hashtags each and further sampled up to 2,000 image-containing tweets from each of the next 44 hashtags.

For efficiency, we engaged a third-party data partner to retrieve tweets. Through our data partner, we were able to access 145,662 tweets. Fewer than theoretical maxima were retrieved since some niche hashtags contained limited image-containing tweets. From these tweets, we extracted 176,731 raw images (videos excluded) (see Appendix A).

3.1.2. REFERENCE DATASET: KANIT

The Kanıt ("evidence" in Turkish) dataset forms the legal foundation of our methodology—a curated collection of 231 expert-verified images captured by Anadolu Agency photojournalists during the Gaza conflict since October 2023(Anadolu Agency). This dataset provides three essential grounding functions:

First, it establishes direct legal connections to the Rome Statute of the International Criminal Court (ICC Articles 5–8), documenting violations including genocide, crimes against humanity, and war crimes. Second, it offers verified documentation of Geneva Convention breaches, such as banned munitions deployment. Third, its evidentiary authority is recognized through citation in Amnesty Inter-

national reports and submission as evidence in ICJ Case No. 2023/77 (South Africa v. Israel) (International Court of Justice, 2023).

By anchoring our analysis to this legally validated reference, we ensure alignment with internationally recognized standards of human rights documentation.

3.2. Preprocessing Pipeline

Our four-stage pipeline refined 176,731 images to 4,603 high-confidence instances through:

1. **Content Filtering** removing non-photographic and irrelevant images
2. **Duplicate Removal** eliminating exact/near-identical copies
3. **Noise Reduction** identifying noise and outliers
4. **Evidence Alignment** with legal evidence standards.

This progressive approach maximized relevance while minimizing manual intervention (Fig. 1).

3.2.1. CONTENT FILTERING

We implemented a two-stage VLM classification process to progressively filter out non-conflict imagery:

Stage 1: Media Authenticity Filtering Using task-specific prompts (see Appendix B), we categorized all images into:

- `AI-generated/drawing`: Synthetic or illustrated content
- `screenshot`: Digital interface captures
- `realPhoto`: Authentic photographs

Only `realPhoto` images progressed to Stage 2, filtering out 73,148 images (41.4% of raw data). A manual review of a randomly chosen hashtag of 2,581 images produced a Negative Predictive Value (NPV) of 99.65%, confirming that almost every image the classifier filtered out was in fact irrelevant.

Stage 2: Conflict Relevance Classification The remaining 103,582 images were classified into fine-grained categories after an iterative review process, during which we identified six commonly occurring types:

- `celebrity`: Non-conflict public figures
- `protest`: Demonstrations from all around the world
- `adultery`: Explicit/inappropriate content
- `textual`: Text-dominated visuals
- `neutralPortrait`: Personal photography
- `conflictRelevant`: Potential violation imagery

Across all hashtags, 54,368 images classified as `conflictRelevant` were retained (69.2% reduction from the initial dataset). Manual review of the same tag revealed that across all irrelevant groups (636 images in

total), 19 were found to be misclassified, yielding a NPV of 97.01%, further indicating the classifier's strong, though not perfect, ability to exclude non-conflict-relevant content.

3.2.2. DUPLICATE REMOVAL

To eliminate redundant content while preserving contextual variations, we applied complementary deduplication techniques at two levels of granularity. First, we removed exact duplicates by computing SHA-256 hashes of raw image bytes. This approach identified and eliminated identical copies regardless of format or metadata, removing 15,114 exact duplicates. Second, we addressed near-duplicates using pretrained ResNet-18 model (He et al., 2016) feature extraction. By calculating cosine similarity between image embeddings with a threshold of 0.9 (empirically optimized to balance redundancy reduction against viewpoint preservation), we filtered near-identical content while retaining valuable variations such as different angles of the same scene and temporally sequential captures. This step removed 16,122 near-duplicates while maintaining visually distinct perspectives.

Combined, these techniques reduced the dataset by 31,236 images (57.5%), leaving 23,132 unique visuals for subsequent analysis.

3.2.3. NOISE REDUCTION

To overcome limitations of rule-based filtering, we employed unsupervised clustering on VLM vision-encoder generated semantic embeddings. This approach identified visually coherent groups while preserving flexibility for emerging patterns, moving beyond rigid predefined categories.
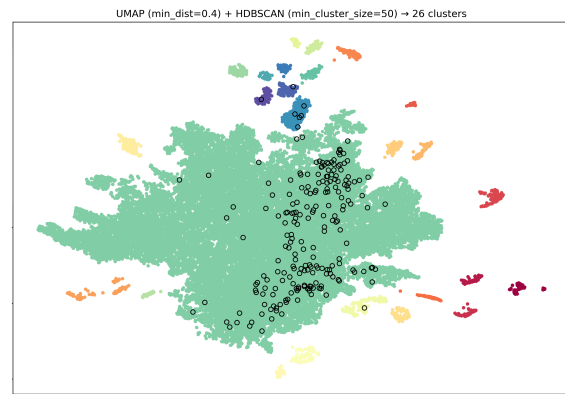

UMAP (min_dist=0.4) + HDBSCAN (min_cluster_size=50) → 26 clusters

Figure 2. Initial clustering results showing the significant majority of Kanıt reference images shown with hallow circles grouped into a large cluster while outlier clusters are visible.

The process began with feature extraction using InternViT-6B (Chen et al., 2024) to generate 3200-dimensional embed-

dings. We then applied UMAP (McInnes et al., 2018) for dimensionality reduction with a minimum distance parameter of 0.4, followed by HDBSCAN (Campello et al., 2013) clustering configured with a minimum cluster size of 50. This combination allowed us to identify natural groupings in the data without imposing predetermined cluster counts.

Initial results (Fig. 2) revealed two critical insights: while significant portion of the Kanıt reference images were grouped into a single cluster - indicating insufficient granularity for contextual separation - clear outlier clusters simultaneously emerged containing various types of non-conflict imagery. Through manual inspection, we identified prevalent noise categories relating to residual AI-generated content and irrelevant scenes (such as furniture or maritime imagery).

By systematically removing these identified outlier clusters, we eliminated 1,743 images (7.5% of the input), significantly enhancing topical coherence while retaining 21,389 visually distinct candidates for further refinement.

3.2.4. EVIDENCE ALIGNMENT

The final clustering stage served dual objectives: (1) identifying social media content visually and contextually aligned with verified evidence in the Kanıt dataset, and (2) discovering potential violation patterns beyond those currently documented. To achieve this, we optimized our clustering pipeline specifically to associate social media images with Kanıt's article-level provenance while maintaining sensitivity to novel patterns.

Through parameter exploration focused on cluster granularity, we implemented a configuration (UMAP min_dist=0.05, HDBSCAN min_cluster_size=35) that balanced two criteria: sufficient separation of Kanıt images by source article, and preservation of visually similar content regardless of provenance. This generated 135 semantically coherent clusters where 13,834 candidate images belong to conflict related clusters (see Figure 3).

Within this structure, 4,603 images (33.3% of clustered content) co-clustered with Kanıt reference images (see Appendix D), serving as potential supplementary evidence for documented violations. The remaining 9,231 images formed distinct clusters not containing Kanıt references, representing visually cohesive patterns that may indicate undocumented violation incidents. This dual-path outcome both strengthens existing evidence documentation and identifies promising directions for future human rights investigation.

3.3. Vision-Language Models

Our annotation pipeline leverages OpenGVLab's state-of-the-art open-weight models, selected for their leading performance on vision-language benchmarks at publication
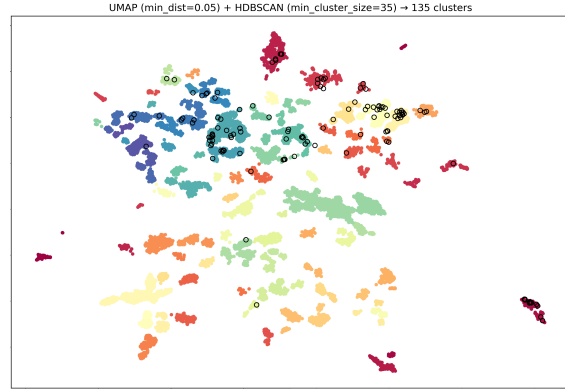

Figure 3. Optimized clustering yielding 135 distinct clusters with clear separation. Enables precise Kanıt dataset alignment through contextual grouping.

time. The infrastructure comprises two components: (i) The **InternViT 6B V2.5** vision encoder (Wang et al., 2024) provides exceptionally rich semantic representations through its 6-billion parameter architecture - the largest openly available vision transformer during our development period. This high-capacity model generates 3200-dimensional embeddings that capture fine-grained visual concepts essential for clustering nuanced conflict imagery. (ii) The **InternVL2.5 78B** vision-language model (Chen et al., 2024) integrates this powerful encoder with a fine tuned Qwen2.5 72B (Team, 2024) language model. This combined architecture enables zero-shot classification through prompt engineering, allowing us to categorize images without task-specific training. The model's joint training approach balances contrastive learning for image-text matching with generative capabilities for contextual understanding. In order to improve classification accuracy we apply a confidence thresholding (≥ 0.7) and automated filtering rules based on VLM outputs.

4. Discussion and Future Work

GazaVHR demonstrates the potential of vision-language models for analyzing conflict-related imagery at scale, offering a valuable resource for humanitarian and legal contexts. While the dataset benefits from structured filtering and expert-aligned references, it lacks manual verification, which may affect precision.

Future work will further examine the 4,603 images related to the Kanıt dataset, followed by expert review to validate and assess potential human rights violations. We will also examine the remaining image clusters to identify additional instances of potential human rights violations not covered in the Kanıt dataset, using VLM-based detection and expert comparison.

Acknowledgements

References

Anadolu Agency. The evidence. https://aakitap.com.tr/kanit. Accessed: 2025-05-23.

Campello, R. J., Moulavi, D., and Sander, J. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172. Springer, 2013.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

International Court of Justice. The republic of south africa institutes proceedings against the state of israel and requests the court to indicate provisional measures. https://www.icj-cij.org/node/203395, 2023. Accessed: 2025-05-30.

Kalliatakis, G. *Visual Recognition of Human Rights Violations*. PhD thesis, University of Essex, 2019.

Kalliatakis, G., Ehsan, S., Fasli, M., Leonardis, A., Gall, J., and McDonald-Maier, K. D. Detection of human rights violations in images: Can convolutional neural networks help? *arXiv preprint arXiv:1703.04103*, 2017a.

Kalliatakis, G., Ehsan, S., and McDonald-Maier, K. D. A paradigm shift: Detecting human rights violations through web images. *arXiv preprint arXiv:1703.10501*, 2017b.

Kalliatakis, G., Ehsan, S., Leonardis, A., Fasli, M., and McDonald-Maier, K. D. Exploring object-centric and scene-centric cnn features and their complementarity for human rights violations recognition in images. *IEEE Access*, 7:10045–10056, 2019.

Lau, N., Srinakarin, K., Aalfs, H., Zhao, X., and Palermo, T. M. Tiktok and teen mental health: an analysis of user-generated content and engagement. *Journal of pediatric psychology*, 50(1):63–75, 2025.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Meneses, R., Brito, C., Lopes, B., and Correia, R. Satisfaction and dissatisfaction in wine tourism: A user-generated content analysis. *Tourism and Hospitality Research*, 25 (1):120–134, 2025.

Nemkova, P., Ubani, S., Polat, S. O., Kim, N., and Nielsen, R. D. Detecting human rights violations on social media during russia-ukraine war. *arXiv preprint arXiv:2306.05370*, 2023.

O'Connell, T. and Young, S. No more hidden secrets: Human rights violations and remote sensing. *Genocide Studies and Prevention: An International Journal*, 8(3):5, 2014.

Pilankar, Y. *Human Rights Violation Detection on Social Media*. PhD thesis, Dublin, National College of Ireland, 2022.

Team, Q. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Twitter Trending Archive. Trending hashtags archive. https://archive.twitter-trending.com. Accessed: 2025-05-23.

Wang, W., Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Zhu, J., Zhu, X., Lu, L., Qiao, Y., and Dai, J. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.

Wang, X., Janssens, B., Bogaert, M., Vanderbauwhede, L., and Schetgen, L. Determining well-being during a crisis based on twitter data. *Annals of Operations Research*, pp. 1–38, 2025.

A. Hashtag Details

The table below lists the hashtags, along with the corresponding image counts collected for each category in our dataset.

Hashtag	Image Count
#RafahOnFire	4563
#getoutofrafah	4660
#GazaGenocide	5141
#PalestineGenocide	4785
#RafahOnFıre	4631
#Hamas	4805
#FreePalestine	5002
#ZionistLobbyAttack	4339
#StrikeForGaza	5092
#GazaStarving	4400
#ZionistCensorship	4393
#CeasefireNOW	5093
#Palestine	5013
#GazaUnderAttack	5605
#starlinkforgaza	5148
#AirDropAidForGaza	2552
#AllEyesOnRafah	2613
#AlShifaHospital	2463
#DontForgetGaza	2207
#Ebu_Ubeyde	2167
#ElonGoToGaza	2388
#FilistiniUnutma	2259
#ForPalestine	2978
#FreeDrHussamAbuSafiya	2582
#FreeGaza	2646
#FreePalestineNow	2534
#Gaza	2581
#Gaza_Genocide	2581
#GazaAttack	2501
#Gazagenocide	2462
#GazaHoloucast	2419
#GazaHospital	2604
#GazzedeKatliamVar	2360
#GazzedeSoykırımVar	2445
#GazzeninAhıDünyayıYakar	1599
#GazzeyeElUzat	1670
#GazzeyeHavadanYardım	2228
#GenocideSupporters	2530
#hamasattack	2398
#HandsOffRafah	2375
#Israel	2559
#IsraelTerorrist	2515
#IsraelTerrorism	2501
#IsraelTerrorists	2611
#KassamOnurumuzdur	2137
#KassamTugayları	2283
#LetGazaBabiesLive	2930
#MyAgendaGaza	2108
#OpenRafahBorder	2256

Hashtag	Image Count
#OpensTheDoorsToGaza	2440
#palastine	2514
#ProtectGazaNOW	1310
#Rafah	2691
#RefahSınırıAçılsın	2287
#RefahtaSoykırımYasanıyor	2206
#SaveTheBabiesOfGaza	1654
#StopGazaHolocaust	2170
#VisitGaza	2342
#ZionistTerror	2405
Total	**176731**

B. Classification Prompts

B.1 Media Authenticity Filtering Prompt (3 Categories)

We used the following prompt to classify each image into one of three broad media types: `"real_photo"`, `"drawing"`, or `"screenshot"`.

```
Classify the image into one of these categories:
"real_photo", "drawing", or "screenshot".
- real_photo: A natural photograph taken by a camera
- screenshot: A digital screen captured from a device
- drawing: A hand-drawn or digitally drawn artwork

Respond only with a valid JSON object using this format:

{ "category": "real_photo" }
```

This step served as a coarse-grained filtering layer to differentiate AI-generated visuals, synthetic scenes, and original photographs.

B.2 Conflict Relevance Classification Prompt (6 Categories)

For further semantic refinement of real-world photographs, we applied a second-stage classifier with the following prompt. It divides images into six thematic categories:

```
Classify the image into one of these categories:
- celebrity: An image of a well-known person such as a politician, actor, or
influencer
- protest: A public demonstration where people intentionally gather to express a
political or social message | often holding banners, placards, or flags. Do not
classify general crowds (e.g., refugees, funerals, aid lines) as protest.
- adultery: Images involving nudity or sexual content, including pornography or
suggestive themes
- textual: Images made up entirely or mostly of written text, slogans, brand logos,
captions, or signs without visible people
- neutral_portrait: A clear portrait of a person or people, with no visible wounds,
injuries, protest context, or signs of distress | often taken in a studio or neutral
setting
- other: Real photograph that doesn't fit the above categories

Respond only with a valid JSON object using this format:

{ "category": "neutral_portrait" }
```

This fine-grained categorization facilitated content-specific filtering for humanitarian and ethical auditing tasks within the dataset pipeline.

C. Model Architecture Details

C.1 InternViT-6B-448px-V2.5

This vision encoder served as the foundation for semantic embedding extraction, selected for its state-of-the-art performance and open weights availability. Key specifications include:

- **Parameters**: 5.5 billion (reduced from original 5.9B by removing final 3 layers)

- **Architecture**: Vision Transformer with 45 transformer blocks

- **Hidden Size**: 3200 dimensions

- **Innovations**:
  - QK-Norm for stable attention
  - RMSNorm for efficient layer normalization
  - Dynamic resolution training (448px base, up to 896px)

- **Training**: Contrastive pretraining followed by language alignment via MLP projector

- **Output**: 3200-dimensional semantic embeddings

C.2 InternVL2.5-78B-MPO
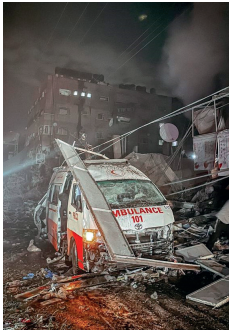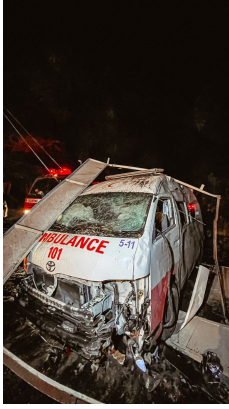
This integrated vision-language model enabled our zero-shot classification pipeline:

- **Vision Component**: InternViT-6B-448px-V2.5 encoder

- **Language Component**: Fine tuned Qwen2.5-72B-Instruct decoder

- **Integration**: Multi-layer perceptron (MLP) projector connecting vision and language modules

- **Training Objective**: Joint optimization using:
  - Contrastive loss for image-text matching
  - Generative loss (next-token prediction)

- **Optimizations**:
  - Enhanced high-resolution understanding
  - Improved local feature sensitivity
  - Multi-Precision Optimization (MPO) for efficient inference

- **Hardware**: Deployed on 4×NVIDIA A100 (80GB) GPUs

C.3 Implementation Details

- **Library**: vLLM

- **Precision**: FP16 mixed-precision training

- **Throughput**:
  - Vision encoder: 42 images/sec (single A100)
  - VLM: 0.4 images/sec (4×A100)

- **Memory**: Peak VRAM utilization of 72GB during VLM inference

- **Code Availability**: Full implementation at `github.com/OpenGVLab/InternVL`

D. Sample Images from Datasets

Cluster#	Example from Kant Dataset & Related Article	GazaVHR Example1	GazaVHR Example2	GazaVHR Example3
38	Article 8, 2-b (v) - Bombing cities, villages, settlements and buildings that are defenseless or do not constitute a military target, or attacking these places with any means			
95	Article 8, 2-b (iii) - Intentionally directing attacks against personnel, facilities, materials, units or vehicles assigned to peacekeeping forces or humanitarian aid			
126	Article 8, 2-a (i) - Wilful killing			