Beyond Visual Similarity: Rule-Guided Multimodal Clustering with explicit domain rules

Kishor Datta Gupta[1], Mohd Ariful Haque[1], Marufa Kamal[2], Ahmed Rafi Hasan[3], Md. Mahfuzur Rahman[1], Roy George[1]

[1]Clark Atlanta University, USA,
[2]BRAC University, Bangladesh,
[3] United International University, Bangladesh
Correspondence: mdmahfuzur.rahman@students.cau.edu

Abstract

Traditional clustering techniques often rely solely on similarity in the input data, limiting their ability to capture structural or semantic constraints that are critical in many domains. We introduce the Domain-Aware Rule-Triggered Variational Autoencoder (DART-VAE), a rule-guided multimodal clustering framework that incorporates domain-specific constraints directly into the representation learning process. DART-VAE extends the VAE architecture by embedding explicit rules, semantic representations, and data-driven features into a unified latent space, while enforcing constraint compliance through rule-consistency and violation penalties in the loss function. Unlike conventional clustering methods that rely only on visual similarity or apply rules as post-hoc filters, DART-VAE treats rules as first-class learning signals. The rules are generated by LLMs, structured into knowledge graphs, and enforced through a loss function combining reconstruction, KL divergence, consistency, and violation penalties. Experiments on aircraft and automotive datasets demonstrate that rule-guided clustering produces more operationally meaningful and interpretable clusters—for example, isolating UAVs, unifying stealth aircraft, or separating SUVs from sedans—while improving traditional clustering metrics. However, the framework faces challenges: LLM-generated rules may hallucinate or conflict, excessive rules risk overfitting, and scaling to complex domains increases computational and consistency difficulties. By combining rule encodings with learned representations, DART-VAE achieves more meaningful and consistent clustering outcomes than purely data-driven models, highlighting the utility of constraint-guided multimodal clustering for complex, knowledge-intensive settings.

1 Introduction

Many visual clustering methodologies presume that visual similarity reflects functional similarity; however, appearance and function can diverge in specialized domains. General image clustering methods perform well in natural image domains (Van Gansbeke et al., 2020), yet frequently falter in specialized datasets where operational semantics precede visual appearance. Public benchmark models, while achieving impressive results on large-scale datasets, often fail to transfer effectively when fine-tuned for domain-specific tasks. Vision Transformers (ViTs) have shown promise in this space due to their strong representational capacity, but their success typically relies on very large volumes of training data. Attempting to fine-tune ViTs with limited specialized datasets often results in overfitting(Liu et al., 2021; Zhao et al., 2025), whereas relying solely on public datasets without proper adaptation leads to underfitting and poor generalization in operational settings. Multimodal techniques (Radford et al., 2021; Jia et al., 2021) attempt to bridge this gap by inferring relationships through large-scale data-driven optimization; however, without embedding domain-specific constraints, they risk overlooking the expert knowledge essential for high-stakes applications.

As an illustrative case, stealth bombers and fighters often adopt angular geometries to minimize radar cross-section, yet aircraft with equivalent functions may present markedly distinct visual profiles due to differing manufacturer philosophies and generational design shifts. Similarly, crossover SUVs may appear visually similar, but subtle distinctions in structure and purpose reflect their alignment with separate market segments. These examples highlight that visual similarity alone is insufficient to capture functional

or operational equivalence.

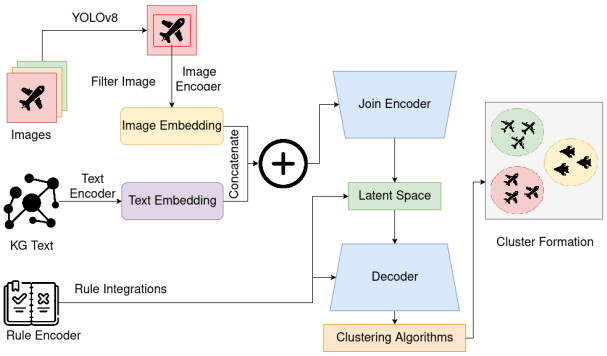


Figure 1: DART-VAE: Domain-Aware Rule-Guided Training for Variational Autoencoders. Our framework integrates visual features, semantic knowledge from structured ontologies, and explicit domain rules (Generated by LLM) through specialized encoders, ensuring learned representations inherently respect operational constraints.

Our primary contributions:

- A multimodal VAE architecture encodes visual features, semantic knowledge graphs, and explicit domain rules(Generated by LLM) via specialized pathways, ensuring that learned representations comply with operational constraints.
- A multifaceted objective function that equilibrates reconstruction fidelity with rule enforcement through the integration of consistency and violation losses, thereby maintaining generative properties while conforming to expert-defined relationships.

This study formulates a systematic framework for integrating expert knowledge into deep clustering, relevant to domains such as medical imaging, industrial inspection, and scientific analysis, where domain constraints are paramount.

2 Related Works

Clustering techniques for visual data have achieved notable success in grouping images based on surface-level similarities. Methods such as k-means, fuzzy c-means, Deep Embedded Clustering (DEC) (Xie et al., 2016), and Variational Deep Embedding (VaDE) (Jiang et al., 2016) optimize purely data-driven objectives but often fail in capturing fine-grained, domain-specific distinctions. In specialized datasets (e.g., vehicles), these approaches neglect semantic and structural knowledge required for meaningful subclass separation. While joint feature learning frameworks such as DAC (Chang et al., 2017) and its refinements with local structure preservation (Guo et al., 2017) and augmentation (Guo et al., 2018) improve robustness, they remain limited to visual similarity. More recent pipelines, including CPP (Chu et al., 2024), leverage pre-trained features for scalability yet still lack explicit integration of domain constraints. Constraint-based clustering has introduced must-link and cannot-link supervision (Ge et al., 2007), and generative models have been applied to encode expert priors (Andreeva et al., 2020). Multimodal approaches enhance interpretability by aligning visual and semantic features (Chen et al., 2021), and fine-grained methods have advanced through diffusion-based (Yang et al., 2024) and bipartite factorization (Peng et al., 2024). Despite these efforts, existing methods rarely embed domain rules as first-class constraints during representation learning, leaving a gap for frameworks that unify visual, semantic, and rule-based guidance.

3 Methodology

3.1 Problem Formulation and Motivation

The primary challenge in specialized domain clustering is the disparity between visual appearance and functional purpose. Conventional clustering techniques depend solely on pixel-level or derived visual features, operating under the premise that visually analogous objects possess functional similarities. This assumption fails profoundly in areas where form adheres to highly specialized function rather than aesthetic resemblance.

Human vs. Machine Perception Disparity: Humans possess inherent domain knowledge enabling them to instantly differentiate between a combat fighter and a transport aircraft, despite their similar visual characteristics. Machine learning models, however, lack contextual comprehension and rely solely on superficial visual patterns. For example, the F-16 Fighting Falcon and C-130 Hercules may seem analogous in aerial images due to their monoplane design; however, they fulfill distinctly different operational roles—one as an air superiority fighter and the other as a tactical transport aircraft.

Operational Reality: In military aviation, aircraft with nearly identical visual signatures can serve very different functions. The F-22 Raptor and F-117 Nighthawk both possess angular, stealth-optimized

Example	Without Rules	With Rules	Improvement
Aircraft Domain			
MQ-9 Reaper (UAV)	Mixed F-18 fighter	Isolated UAV cluster	✓ UAV separation
F-22, Su-57	Scatter multiple clusters	Unified stealth cluster	✓ Technology consistency
C-130, C-2	Mixed A-10 combat	Pure transport cluster	✓ Mission separation
Vehicle Domain			
BMW M3 (Performance)	Mixed Toyota Camry	Performance cluster	✓ Performance separation
Ferrari 488 (Luxury)	Grouped Honda Civic	Luxury sports cluster	✓ Market segmentation
Range Rover (SUV)	Mixed sedan vehicles	SUV cluster	✓ Body style coherence

Table 1: Rule-Guided Clustering Improvement Examples

designs for radar evasion; however, the F-22 serves as an air superiority fighter with supercruise capability, whereas the F-117 operated as a precision strike bomber. Conversely, functionally analogous aircraft may exhibit significant visual diversity owing to varying design epochs, manufacturers, and technological methodologies.

Clustering Inadequacy: Conventional clustering algorithms consistently categorize the MQ-9 Reaper (UAV), F-18 Hornet (fighter), and KC-135 Stratotanker (refueling aircraft) solely based on visual resemblance, resulting in operationally irrelevant clusters that contravene essential military doctrine principles.

Our DART-VAE framework addresses this limitation by embedding domain-specific physical rules (which generated by LLM and Domain specific Books) directly into the representation learning process, ensuring that learned embeddings respect both visual coherence and operational semantics. We define this as the acquisition of a latent representation $z \in \mathbb{R}^d$ that organizes data points based on visual similarity and domain constraints $R = \{r_j\}_{j=1}^{M}$.

3.2 Overall Architecture

The DART-VAE framework employs a three-stage pipeline that methodically converts raw multimodal data into constraint-aware latent representations appropriate for domain-informed clustering.

Stage 1: Multimodal Feature Extraction Raw images are subjected to object detection via YOLOv8 to delineate regions of interest (ROI) with adaptive padding, thereby removing background noise that results in erroneous groupings. Concurrently, structured domain knowledge from JSON-formatted knowledge graphs (aircraft) or CSV metadata (vehicles) is processed using Sentence-BERT to produce semantic embeddings. These Knowledge graphs and metadata are acquire by LLM fine-tuned by domain-specific contents). Binary rule features are extracted and encoded via specialized MLPs.

Stage 2: Constraint-Guided Representation Learning The DART-VAE encoder processes concatenated multimodal features (visual + semantic + rules) through a joint encoder network that learns the posterior distribution $q(z|x, t, r)$. The formation of latent space is directed by a multi-faceted loss function that equilibrates reconstruction accuracy, KL regularization, rule adherence, and penalty for violations.

Stage 3: Rule-Validated Clustering Acquired latent representations undergo hard (K-means) and soft (Fuzzy C-means) clustering, followed by rule-guided refinement, in which constraint violations lead to reassignment to the nearest compliant cluster, based on both latent distance and rule adherence.

3.3 Domain-Specific Physical Rules

The core innovation of DART-VAE lies in the explicit formalization of domain knowledge as enforceable constraints. We establish unique rule sets for each domain that encapsulate essential operational principles.

3.3.1 Aircraft Domain Rules

According to military aviation doctrine and aerospace engineering principles, a fine-tuned LLM generated four essential constraints:

Rule 1: Stealth Technology Consistency Stealth aircraft represent a highly specialized technological category requiring sophisticated systems integration. Aircraft with stealth capabilities must exhibit technological reliability via advanced avionics systems and possess either air superiority (fighter classification) or supersonic cruise capability.

Algorithm 1

DART-VAE: Domain-Aware Rule-Triggered Clustering

Require: Multimodal dataset $\mathcal{D} = \{(x_i, t_i, r_i)\}_{i=1}^{N}$, Domain rules $\mathcal{R}$

Ensure: Rule-compliant clusters $\mathcal{C} = \{C_k\}_{k=1}^{K}$

Stage 1: Multimodal Feature Extraction
$X_{\text{roi}} \leftarrow \text{ObjectDetection}(\{x_i\})$ ▷ ROI extraction
$F_v \leftarrow \text{VisualEncoder}(X_{\text{roi}})$ ▷ Visual features
$F_t \leftarrow \text{SemanticEncoder}(\{t_i\})$ ▷ Knowledge features
$F_r \leftarrow \text{RuleEncoder}(\{r_i\}, \mathcal{R})$ ▷ Rule features
$F_{\text{joint}} \leftarrow \text{Concatenate}(F_v, F_t, F_r)$
Stage 2: Constraint-Guided Representation Learning
for epoch $e = 1$ to T **do**
  $\mu, \sigma^2 \leftarrow \text{Encoder}(F_{\text{joint}})$
  $Z \leftarrow \text{Reparameterize}(\mu, \sigma^2)$ ▷ Latent sampling
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{recon}} + \beta\mathcal{L}_{\text{KL}} + \alpha_e(\mathcal{L}_{\text{consistency}} + \mathcal{L}_{\text{violation}})$
  Update θ via $\nabla_\theta \mathcal{L}$ ▷ Progressive rule integration
end for
Stage 3: Rule-Validated Clustering
$\mathcal{C}_{\text{init}} \leftarrow \text{Clustering}(Z)$ ▷ K-means or Fuzzy C-means
for each constraint $r \in \mathcal{R}$ **do**
  for each cluster $C_k \in \mathcal{C}_{\text{init}}$ **do**
    if $\text{ViolatesRule}(C_k, r)$ **then**
      Reassign violating samples to nearest compliant clusters
    end if
  end for
end for
return $\mathcal{C}$

Formal Definition: $\forall a \in \text{Aircraft} : \text{is\_stealth}(a) \rightarrow (\text{has\_advanced\_avionics}(a) \wedge (\text{is\_fighter}(a) \vee \text{has\_supercruise}(a)))$

Physical Rationale: Stealth technology necessitates advanced radar systems, electronic warfare capabilities, and intricate flight controls. The F-22 Raptor integrates stealth capabilities with supercruise, whereas the F-117 Nighthawk depended on sophisticated avionics for precision strike operations.

Rule 2: UAV Operational Separation Unmanned and manned aircraft function under distinct doctrines, certification criteria, and operational protocols. They must uphold distinct clustering boundaries to accurately represent operational reality.

Formal Definition: $\forall a_i, a_j \in C_k : \text{is\_uav}(a_i) \leftrightarrow \text{is\_uav}(a_j)$

Physical Rationale: UAVs such as the MQ-9 Reaper function with different risk profiles, endurance capacities, and mission specifications in contrast to manned aircraft like the F-16.

Rule 3: Mission-Type Doctrinal Enforcement Military aircraft are designed and optimized for specific mission profiles. Combat platforms must not be grouped with transport or logistics aircraft due to inherent disparities in operational requirements, threat environments, and deployment patterns.

Formal Definition: $\forall a_i, a_j \in C_k : \text{mission\_type}(a_i) = \text{combat} \rightarrow \text{mission\_type}(a_j) \neq \text{transport}$

Physical Rationale: Combat aircraft like the A-10 Thunderbolt II are armored for survivability in hostile environments, while transport aircraft like the C-130 prioritize cargo capacity and operational versatility.

Rule 4: Physical Attribute Coherence Aircraft within the same operational cluster must demonstrate comparable fundamental physical attributes, including propulsion systems and performance envelopes, indicative of analogous operational requirements.

Formal Definition:$\forall a_i, a_j \in C_k : \text{engine\_type}(a_i) = \text{engine\_type}(a_j) \wedge \text{speed\_class}(a_i) = \text{speed\_class}(a_j)$

Physical Rationale: Turbofan-powered aircraft function within individual performance parameters compared to turboprop aircraft, influencing range, altitude capabilities, and mission appropriateness.

3.3.2 Automotive Domain Rules

For vehicles, OPENAI GPT3.5 outlines four rules that encapsulate market segmentation and engineering principles for vehicles.

Rule 1: Body Style Coherence Vehicles with fundamentally diverse body structures cater to specific market niches and usage patterns. SUVs, sedans, and convertibles cater to their own market requirements and should maintain cluster differentiation.

Rule 2: Performance Tier Consistency Economy cars and high-performance cars have different engineering aims and target various market segments. Performance cars put handling and power-to-weight ratios first, whereas economy cars maximize cost and fuel efficiency.

Rule 3: Dimensional Proportionality Vehicles with notably distinct physical proportions (height-to-length ratios, ground clearance) fulfill diverse practical functions and should remain distinct in clustering.

Rule 4: Luxury Market Segmentation Luxury and standard market vehicles represent distinct value propositions with different feature sets, pricing strategies, and brand positioning.

Semantic Knowledge Encoding: Sentence-BERT (all-mpnet-base-v2) processes structured domain expertise. Aircraft use JSON-formatted knowl-

edge graph triples with technical specifications, operational roles, and performance attributes contained in 384D vectors and compressed to 256D via 2-layer MLP. Sentence-BERT to 768D embeddings procedure CSV brand hierarchies, technical specifications, body styles, and market categories for vehicles.

Rule Feature Engineering: Domain constraints are implemented as explicit feature vectors via specialized MLPs. Aircraft utilize 10 binary attributes (is_stealth, is_uav, has_crew, has_supercruise, has_advanced_avionics, mission_type indications) processed using MLP to 16-dimensional space. Vehicles utilize 18 derived attributes from the four automotive rules (body_style_category, performance_tier, size_ratios, luxury_indicators) compressed into a 32-dimensional space.

3.4 DART-VAE Architecture Details

Joint Encoding: The multimodal feature fusion creates domain-specific joint representations. Aircraft: $f_{\text{joint}} \in \mathbb{R}^{100,624}$ (visual: 100,352D + semantic: 256D + rules: 16D). Vehicles: $f_{\text{joint}} \in \mathbb{R}^{100,388}$ (visual: 100,352D + semantic: 768D + rules: 32D).

Encoder Network: A progressive compression architecture maps joint features to latent parameters: $h_1 = \text{ReLU}(\text{Linear}(f_{\text{joint}}, 512))$, $h_2 = \text{ReLU}(\text{Linear}(h_1, 256))$, $\mu, \log \sigma^2 = \text{Linear}(h_2, 64), \text{Linear}(h_2, 64)$.

Latent Sampling: The reparameterization trick enables gradient-based optimization: $z = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I_{64})$.

Decoder Network: Reconstructs multimodal features to ensure representation fidelity: $h_3 = \text{ReLU}(\text{Linear}(z, 256))$, $h_4 = \text{ReLU}(\text{Linear}(h_3, 512))$, $f_{\text{reconstructed}} = \text{Linear}(h_4, \dim(f_{\text{joint}}))$.

3.5 Multi-Component Loss Function

Our training objective combines traditional VAE losses with rule-specific penalties that address the fundamental challenge observed in our aircraft and automotive clustering experiments:

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \beta \mathcal{L}_{KL} + \alpha(\mathcal{L}_{consistency} + \mathcal{L}_{violation}) \quad (1)$$

Reconstruction Loss ($\mathcal{L}_{recon}$): Standard VAE reconstruction ensures that our multimodal features; visual ROI features, knowledge graph embeddings, and rule encodings, can be faithfully recovered from the latent space:

$$\mathcal{L}_{recon} = \text{MSE}(f_{joint}, \hat{f}_{joint}) \quad (2)$$

KL Divergence Loss ($\mathcal{L}_{KL}$): The standard VAE regularization prevents latent space collapse:

$$\mathcal{L}_{KL} = -\frac{1}{2} \sum_{i=1}^{d} (1 + \log \sigma_i^2 - \mu_i^2 - \sigma_i^2) \quad (3)$$

Rule Consistency Loss ($\mathcal{L}_{consistency}$): This component indicates a key finding from our experiments: aircraft with analogous operational profiles should exhibit comparable latent representations, despite visual differences.

$$\mathcal{L}_{consistency} = \sum_{i,j} \text{MSE}(\text{sim}(z_i, z_j), \text{sim}(r_i, r_j)) \quad (4)$$

The function $\text{sim}(r_i, r_j)$ estimates cosine similarity between 16-dimensional rule features generated by the rule encoder, rather than raw binary inputs. This groups the MQ-9 and TB2 UAVs despite their visual profiles and aligns BMW and Mercedes premium automobiles in latent space despite brand distinctions.

Rule Violation Loss ($\mathcal{L}_{violation}$): Direct constraint enforcement emerges from our domain analysis:

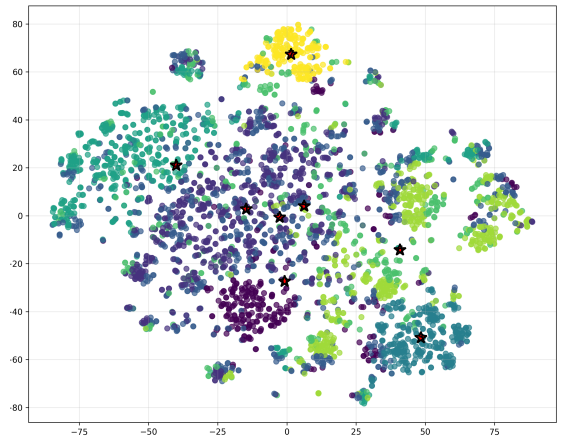
$$\mathcal{L}_{violation} = \text{MSE}(\sigma(v_{pred}), v_{target}) \quad (5)$$

where $v_{pred} \in \mathbb{R}^{N \times 4}$ are the raw violation predictions from the rule predictor network, σ is the sigmoid activation, and $v_{target} \in \{0, 1\}^{N \times 4}$ are the binary violation targets computed from the logical rules.

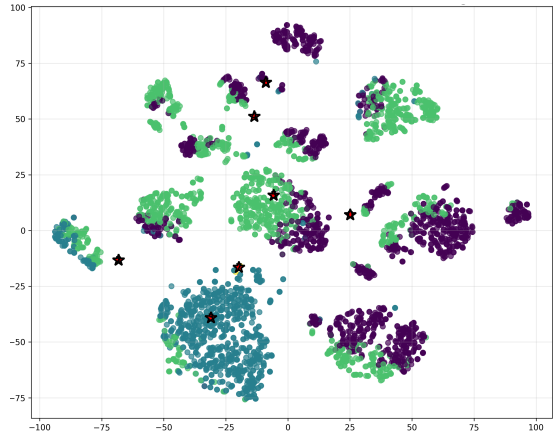
Rule Weight Configuration: Reconstruction fidelity and domain constraint enforcement are balanced using a 0.15 rule weight in airplane (40 epochs) and automobile (30 epochs) domain training.

4 Experimental Results and Analysis

We assess DART-VAE in the aircraft and automotive domains through extensive experimental setups to illustrate the incremental advantages of including domain constraints.



(a) Without physical rules



(b) With 4 physical rules

Figure 2: Car Hard Clustering t-sne visualization

4.1 Experiments

All experiments employ appropriate statistical validation and adhere to the implementation specifics outlined in Section 4.1 and appendix. We applied a two-stage filtering pipeline (Mask R-CNN + YOLOv8, confidence ≥ 0.97, coverage $\geq 60\%$) to curate high-quality images, selecting a representative subset (800 aircraft images across 76 classes) for efficient evaluation. Both datasets are enriched with structured metadata (JSON/CSV) capturing technical specifications and attributes for downstream clustering and analysis.

4.2 Aircraft Domain Clustering Results

4.2.1 Quantitative Performance Analysis

Table 2 breaks down comprehensive aircraft clustering outcomes across various rule configurations and techniques. The methodical advancement from baseline to rule-based grouping reveals distinct performance variation contingent upon the complexity of constraints. **Hard Clustering Performance**: The 2-rule hard configuration achieves remarkable quantitative measures, evidenced by a Silhouette Score of 0.7109 and a robust Calinski-Harabasz Score of 16,325.64, reflecting a 405% enhancement over the baseline hard clustering score of 0.1406. However, the 4-rule hard configuration exhibits diminished traditional metrics (Silhouette: 0.3325) while achieving improved Davies-Bouldin performance (0.9147), which means greater cluster compactness despite lowered separation scores. **Soft Clustering Performance**: Fuzzy C-means with 2-rule configuration achieves remarkable performance with Fuzzy Partition Coefficient of 0.9765 and minimal Fuzzy Partition Entropy of 0.0474, demonstrating optimal cluster separation. Table 2 shows the fuzzy-specific metrics where the 2-rule configuration excels with high certainty in cluster assignments and Average Membership Strength of 0.9850. The 4-rule soft configuration achieves Fuzzy Partition Coefficient of 0.4736 and Fuzzy Partition Entropy of 1.1456, with Average Membership Strength of 0.6184, trading traditional fuzzy metrics for comprehensive constraint coverage.

4.2.2 Rule Violation Analysis

Stealth Consistency Constraints: Violations of stealth technology persist at a consistent rate of 85 across all configurations, signifying structural issues within the stealth aircraft category that surpass clustering algorithms. This ongoing violation pattern indicates intrinsic data complexity, wherein stealth qualities do not fully correspond with other operational features. **UAV Operational Separation**: Hard clustering with rule refinement provides perfect UAV separation (0 violations) by post-processing optimization, while soft clustering keeps 90-214 violations but accepts boundary scenarios when UAV characteristics overlap with manned aircraft traits. With increasing rule complexity, constraint management improves from 214 UAV violations in the 2-rule soft setup to 90 in the 4-rule version. **Mission and Semantic Coherence**: The four-rule configurations impose constraints related to mission type and semantic coherence, resulting in 401 to 533 mission violations and 0 to 103 semantic violations. While

Table 2: Aircraft clustering performance comparison (K-means vs. Fuzzy C-means) across rule configurations

Configuration	Hard Clustering (K-means)				Rule Violations (Hard)			
	SS	DB	CH		Stealth	UAV	Mission	Semantic
Baseline	0.1406	1.5167	174.37	–	–	–	–	
2-Rule	0.7109	1.0666	16325.64		85	0	–	–
4-Rule	0.3325	0.9147	794.43		85	0	401	103
Configuration	Soft Clustering (Fuzzy C-means)				Rule Violations (Soft)			
	FPC	FPE	MS	FS	Stealth	UAV	Mission	Semantic
Baseline	0.5401	0.9775	0.396	–	–	–	–	–
2-Rule	0.9765	0.0474	0.9850	–	85	214	–	–
4-Rule	0.4736	1.1456	0.6184	0.1960	85	90	533	0

SS: Silhouette Score, DB: Davies-Bouldin Score, CH: Calinski-Harabasz Score,
FPC: Fuzzy Partition Coefficient, FPE: Fuzzy Partition Entropy, MS: Average Membership Strength, FS: Fuzzy Silhouette Index

Table 3: Automotive clustering performance comparison (K-means vs. Fuzzy C-means)

	FPC	FPE	MS	FS
Baseline (Soft)	0.125	2.07	0.125	0.167
Rule-Guided (Soft)	0.125	2.08	0.125	0.189
	SS	DB	CH	
Baseline (Hard)	0.054	3.27	173.8	–
Rule-Guided (Hard)	0.139	0.92	12234.9	–
Rule Violations (Soft vs. Hard, Rule-Guided)				
Metric	Soft	Hard		
Body	0.19	3.01		
Performance	3.60	4.83		
Size	2.00	3.92		
Luxury	2.72	4.53		
Total violations: Soft = 8.51; Hard = 16.29				

these increase overall constraint violations, they provide fine-grained operational classification essential for military applications.

4.2.3 Visual Clustering Analysis

Figure 3 illustrates t-SNE visualizations that depict the evolution of clustering across various rule configurations. The 2-rule guided clustering (Figure 3b and 3e) demonstrates effective cluster separation, resulting in well-defined operational groupings. The configurations with four rules (Figure 3c and 3f) demonstrate more intricate boundaries that indicate a higher level of constraint complexity, whereas Figure 3a presents the baseline performance in the absence of rule guidance.

Cluster Coherence: Rule-guided clustering efficiently creates operational coherence, as stealth platforms (F-22, Su-57, F-117) cluster according to technological uniformity, UAVs maintain clear operational distinctions from manned aircraft, and transport aircraft (C-130, C-2) are differentiated from combat platforms (A-10). This indicates a significant shift from visual similarity to functional effectiveness.

4.3 Automotive Domain Clustering Results

4.3.1 Quantitative Performance Analysis

Table 3 illustrates the performance of hard clustering in the automotive sector at various levels of rule integration. Rule-based hard clustering demonstrates significant enhancements: The Silhouette Score rises from 0.0543 (baseline) to 0.1393 (156% enhancement), while the Calinski-Harabasz Score escalates from 173.82 to 12,234.91 (6,941% enhancement), signifying significantly enhanced cluster separation and compactness.

4.3.2 Rule Violation Analysis

Body Style Coherence: The use of rule-guided clustering results in differing degrees of Body Style/Segment violations according on the clustering algorithm utilized. Tables 3 illustrate that fuzzy C-means clustering exhibits superior constraint management with merely 0.19 violations, whereas K-means clustering results in 3.01 violations. This notable disparity demonstrates fuzzy clustering's efficacy in managing the diverse classifications of body shapes in the automotive sector, including SUV, sedan, and hatchback, where automobiles may display ambiguous traits.

Performance and Engineering Constraints: Performance/Drivetrain Consistency is the hardest criteria for both clustering methods, with soft clustering (fuzzy C-means) obtaining 3.60 violations and hard clustering (K-means) 4.83. High violation rates reflect the car industry's sophisticated performance specs that don't match visual similarities—high-performance variations of ordinary models sometimes look identical despite having very different

powertrains. Size Proportion limitations yield mild violations, with fuzzy C-means achieving 2.00 and K-means 3.92. Luxury/Performance Feature consistency provides sufficient constraint adherence, with fuzzy C-means committing 2.72 violations and K-means 4.53.

Algorithm Comparison: The experimental results indicate complementary benefits among clustering methodologies. K-means demonstrates superior mathematical clustering effectiveness, as indicated by a higher Silhouette Score (0.1393 compared to 0.0100) and better cluster separation metrics, making it suitable for applications that prioritize geometric cluster quality. Fuzzy C-means shows improved adherence to domain constraints, with 48% fewer rule violations (8.51 versus 16.29), making it more suitable for applications that require semantic coherence over mathematical optimization. Figure 2 illustrates that rule-guided K-means clustering (Figure 2b) achieves remarkable separation with 8 distinct clusters, whereas baseline clustering (Figure 2a) exhibits considerable mixing.

4.4 Qualitative Clustering Improvements

Table 1 demonstrates specific clustering improvements through rule integration. Critical operational separations achieved include: **Military Applications**: MQ-9 Reaper UAVs isolated from F-18 fighters, stealth aircraft (F-22, Su-57) unified by technological consistency, and transport aircraft (C-130, C-2) separated from combat platforms. **Automotive Applications**: BMW M3 performance vehicles separated from economy cars, Ferrari luxury sports cars grouped appropriately, and Range Rover SUVs clustered by body style coherence. These improvements represent fundamental advances from appearance-based to function-based clustering, critical for domain expert applications where operational semantics transcend visual similarity.

5 Threats to Validity

Sveral threats to validity remain. First, the rules themselves are generated by LLMs and subsequently formatted as structured knowledge graphs. This process introduces the risk of hallucinations, where the LLM may produce inaccurate or spurious rules. Such negative rules can inadvertently bias the clustering process, leading to distortions rather than improvements in cluster quality. Second, the reliance on LLM-generated rules makes the framework sensitive to overfitting. When too many rules are imposed simultaneously, the latent space may become overly constrained, forcing the model to adhere to rigid relationships at the expense of generalizability. This effect was particularly evident when scaling from two-rule to four-rule configurations, where clustering metrics showed diminishing returns despite improved constraint enforcement. Third, scaling to complex domains presents a fundamental challenge. As the number of rules, modalities, and semantic categories increases, the computational overhead and difficulty of maintaining consistent enforcement escalate. In such cases, rule conflicts and inconsistencies may proliferate, complicating both training and interpretability. Without careful curation and validation of rule sets, domain expansion risks undermining the stability and reliability of the method. Finally, the evaluation relies on datasets where rule definitions are relatively well aligned with domain expertise (e.g., aircraft doctrine, automotive market segmentation). For less formalized or more ambiguous domains, the suitability of LLM-derived rules remains uncertain, limiting the external validity of the approach.

6 Conclusion

Proposed multimodal clustering framework that elevates LLM-generated rules and knowledge-graph constraints to first-class learning signals. By embedding rules directly into representation learning, the method balances visual coherence and operational semantics, producing clusters that are both interpretable and quantitatively robust. Results in aircraft and automotive domains show that rule-guided clustering achieves clearer functional separation than purely visual baselines. However, reliance on LLM-generated rules introduces risks of hallucination and inconsistency, and applying too many constraints can lead to overfitting. Scaling to complex domains with large rule sets remains challenging. Despite these limitations, DART-VAE demonstrates the potential of rule-informed clustering as a principled step toward interpretable and domain-aligned AI.

References

a2015003713. Military aircraft detection dataset. Kaggle dataset. Accessed: 2025-01-17.

Olga Andreeva, Wei Li, Wei Ding, Marieke Kuijjer, John Quackenbush, and Ping Chen. 2020. Catalysis clustering with gan by incorporating domain knowledge. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1344–1352.

Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. 2017. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 5879–5887.

Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, and 1 others. 2021. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8012–8021.

Tianzhe Chu Chu, Shengbang Tong, Tianjiao Ding, Xili Dai, Benjamin Haeffele, Rene Vidal, and Yi Ma. 2024. Image clustering via the principle of rate reduction in the age of pretrained models. International Conference on Learning Representations (ICLR).

Rong Ge, Martin Ester, Wen Jin, and Ian Davidson. 2007. Constraint-driven clustering. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 320–329.

Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. 2017. Improved deep embedded clustering with local structure preservation. In *Ijcai*, volume 17, pages 1753–1759.

Xifeng Guo, En Zhu, Xinwang Liu, and Jianping Yin. 2018. Deep embedded clustering with data augmentation. In *Asian conference on machine learning*, pages 550–565. PMLR.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. 2016. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. Ultralytics yolov8.

Zhenda Liu, Han Chen, Yujie Feng, Songtao Liu, Jizhou He, Jie Zhou, and James Zou. 2021. Efficient training of visual transformers with small datasets. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chong Peng, Pengfei Zhang, Yongyong Chen, Zhao Kang, Chenglizhao Chen, and Qiang Cheng. 2024. Fine-grained bipartite concept factorization for clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26264–26274.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Unit293. 2024. Car-models-3887: 3D meshes of 3887 car types. Hugging Face dataset. Accessed: 2024-09-17.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. 2020. Scan: Learning to classify images without labels. In *Computer Vision – ECCV 2020*, pages 268–285, Cham. Springer International Publishing.

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR.

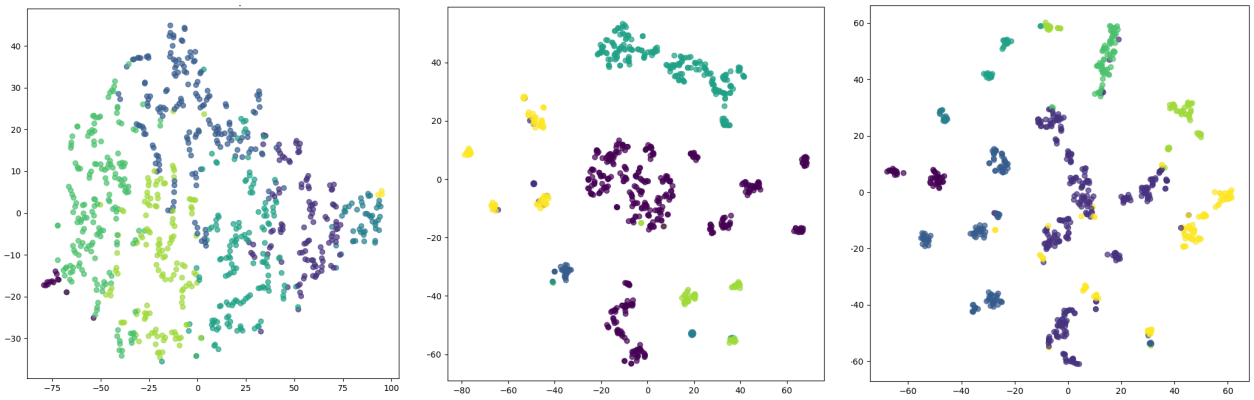
Ruohong Yang, Peng Hu, Xi Peng, Xiting Liu, and Yunfan Li. 2024. Dific: Your diffusion model holds the secret to fine-grained clustering. *arXiv preprint arXiv:2412.18838*.

Chen Zhao, Ziqian Chen, Li Zhang, and Yan Wang. 2025. The missing piece in vit fine-tuning for image aesthetic assessment. *arXiv preprint arXiv:2504.02522*.

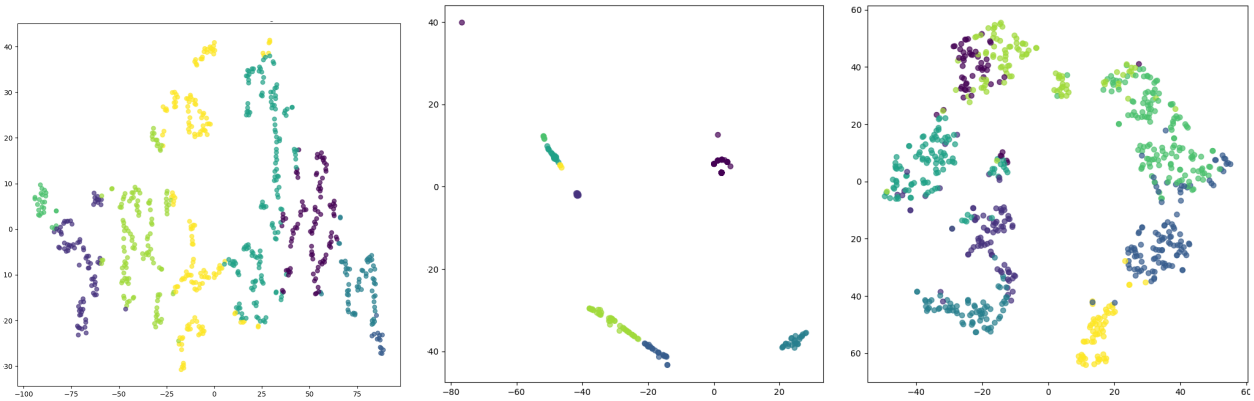
7 Appendix

7.1 Dataset Details

Military Aircraft: We conducted a two-stage data cleansing process on 8,311 raw photos from the military aircraft dataset (a2015003713). Preliminary filtering employing Mask R-CNN, succeeded by YOLOv8 (Jocher et al., 2023) with dual parameters (confidence ≥ 0.97 and aircraft coverage $\geq 60\%$), produced 3,103 high-quality images encompassing 77 aircraft types. To ensure computational efficiency and uniform assessment across rule configurations, we chose a representative subset of 800 photos (after YOLOv8 filtering, preserving 728 images with



(a) Hard clustering without rules (b) Hard clustering with 2 rules (c) Hard clustering with 4 rules



(d) Soft clustering without rules (e) Soft clustering with 2 rules (f) Soft clustering with 4 rules

Figure 3: Aircraft clustering performance demonstrated through t-SNE visualizations. Top row shows hard clustering (K-means), bottom row shows soft clustering (Fuzzy C-means), progressing from no rules (left) to 2 rules (center) to 4 rules (right).

aircraft detections) encompassing 76 aircraft classes for our clustering tests. Every aircraft is marked with operating characteristics and connected to a JSON knowledge graph that includes technical specifications as triples. **Automotive Dataset**: The vehicle dataset (Unit293, 2024) was subjected to quality control comparable to aircraft photos, using YOLOv8 to exclude non-vehicle images, notably interiors. A hierarchical brand/model/year directory structure and CSV metadata with technical specifications, body styles, segments, and dimensional attributes from manufacturer databases comprise the cleaned dataset.

7.2 Implementation Details

All tests utilize the PyTorch framework using NVIDIA RTX 6000 Ada Generation GPUs, each equipped with 49GB of VRAM. The visual encoder utilizes a custom CNN architecture with domain-specific configurations, comprising three convolutional layers for aircraft and four for cars. Text encoding using Sentence-BERT (all-mpnet-base-v2). In the preprocessing phase, YOLOv8 does object detection with a confidence threshold of 0.25 and adaptive padding for region of interest extraction. Optimization employs AdamW with a weight decay of $\lambda = 10^{-4}$. Training setups are specialized to their domains: airplane experiments utilize 40 epochs across all rule configurations, whereas car experiments employ 30 epochs with a 4-rule configuration.

7.3 Multimodal Feature Extraction Pipeline

Visual Feature Processing: We utilize YOLOv8 for object detection and ROI extraction, implementing adaptive padding to encompass entire aircraft and vehicle structures while minimizing background noise. Our experimental validation demonstrates that background contamination substantially affects clustering quality, rendering ROI extraction an essential preprocessing step.

Aircraft Domain: Three convolutional layers Conv2d(3,32) $\rightarrow$ Conv2d(32,64) $\rightarrow$ Conv2d(64,128), flattening yields 100,352D feature vectors ($128 \times 28 \times 28$), 20% adaptive padding applied during ROI extraction.

Automotive Domain: Four convolutional layers with adaptive average pooling, culminating in a 256-dimensional visual feature representation, with a confidence threshold of 0.25 for vehicle detection.