EXPERT JURNAL MANAJEMEN SISTEM INFORMASI DAN TEKNOLOGI

# Penalaran Kompleks pada Citra Digital Motif Batik Lampung Menggunakan Model LVLM

**Ari Kurniawan Saputra**[1*], **Robby Yuli Endra**[2], **Fenty Ariani**[3], **Erlangga**[4]
[1*, 2] Informatika, Fakultas Ilmu Komputer, Universitas Bandar Lampung, Kota Bandar Lampung, Indonesia
[3.4] Sistem Informasi, Fakultas Ilmu Komputer, Universitas Bandar Lampung, Kota Bandar Lampung, Indonesia
[1*] ari.kurniawan@ubl.ac.id, [2] robby.yuliendra@ubl.ac.id,
[3] fenty.ariani@ubl.ac.id, [3] erlangga@ubl.ac.id

**ABSTRACT –** This study applies a Large Vision-Language Model (LVLM) to perform complex Chain-of-Thought (CoT) reasoning on digital images of Lampung batik motifs. Batik Lampung is a traditional textile heritage of the Lampung people, characterized by four distinctive motifs: Leluak Tehambur (Metro City), Kapal Pesagi (South Lampung Regency), Pohon Hayat (Pesawaran Regency), and Motif Bambu (Pringsewu Regency). Existing CNN-based approaches lack the capacity to explain the cultural semantics embedded in these motifs, motivating the need for a reasoning-capable model. A self-collected dataset, BLD-28, comprising 28 images from four official Dekranasda branches in Lampung Province, was annotated by cultural experts with inter-annotator agreement $\varkappa = 0.89$. The InternVL2-8B model was fine-tuned using Low-Rank Adaptation (LoRA, $r = 64$, $\alpha = 128$) under a multi-task loss combining classification and CoT generation objectives. Results show InternVL2-8B achieves 94.37% accuracy, mIoU of 88.12%, and a Reasoning Coherence Score (RCS) of 4.62/5.00, outperforming all baseline CNN and other fine-tuned LVLM models significantly (McNemar test, $p < 0.001$). CoT reasoning improved classification accuracy by 3.21 percentage points over direct classification, demonstrating the viability of LVLM for culturally-aware recognition of Indonesian traditional textile motifs.

**Keywords: CoT; Images; Patterns; Lampung Batik; LVLM.**

**ABSTRAK –** Penelitian ini menerapkan Large Vision-Language Model (LVLM) untuk melakukan penalaran kompleks berbasis Chain-of-Thought (CoT) pada citra digital motif batik Lampung. Batik Lampung merupakan warisan tekstil tradisional masyarakat Lampung yang dicirikan oleh empat motif khas: Leluak Tehambur (Kota Metro), Kapal Pesagi (Kabupaten Lampung Selatan), Pohon Hayat (Kabupaten Pesawaran), dan Motif Bambu (Kabupaten Pringsewu). Pendekatan berbasis CNN yang ada tidak mampu menjelaskan makna budaya yang terkandung dalam motif-motif tersebut, sehingga mendorong kebutuhan akan model yang mampu melakukan penalaran semantik. Dataset mandiri BLD-28 sebanyak 28 citra dikumpulkan dari empat Dekranasda resmi di Provinsi Lampung dan dianotasi oleh pakar budaya dengan inter-annotator agreement $\varkappa = 0{,}89$. Model InternVL2-8B di-fine-tune menggunakan Low-Rank Adaptation (LoRA, $r = 64$, $\alpha = 128$) dengan fungsi loss multi-task yang menggabungkan objektif klasifikasi dan generasi CoT. Hasil menunjukkan InternVL2-8B mencapai akurasi 94,37%, mIoU 88,12%, dan Reasoning Coherence Score (RCS) 4,62/5,00, melampaui seluruh baseline CNN maupun LVLM pembanding secara signifikan (uji McNemar, $p < 0{,}001$). Penalaran CoT terbukti meningkatkan akurasi klasifikasi sebesar 3,21 poin dibandingkan klasifikasi langsung, membuktikan kelayakan LVLM untuk pengenalan motif tekstil tradisional Indonesia yang berbasis pemahaman budaya.

**Kata Kunci: Batik Lampung; Citra; CoT; Motif; LVLM.**

## 1. PENDAHULUAN

Batik Lampung merupakan produk kebudayaan tekstil tradisional yang memiliki kedudukan penting sebagai representasi identitas budaya masyarakat Lampung. Motif-motifnya mengandung nilai filosofis yang diekspresikan melalui elemen visual yang khas, seperti Leluak Tehambur (Kota Metro), Kapal Pesagi (Kabupaten Lampung Selatan), Pohon Hayat (Kabupaten Pesawaran), dan Motif Bambu (Kabupaten Pringsewu). Di tengah ancaman kepunahan kultural akibat modernisasi, pelestarian dan pengenalan motif secara digital menjadi semakin mendesak [1], [2].

Pendekatan berbasis Convolutional Neural Network (CNN) yang telah banyak diterapkan pada klasifikasi motif batik memiliki keterbatasan mendasar: hanya menghasilkan label kelas tanpa penjelasan semantik yang bermakna tentang elemen budaya. Paradigma Large Vision-Language Model (LVLM) hadir sebagai solusi yang menggabungkan pemahaman visual dengan penalaran bahasa alami [3],[4]. Kemampuan chain-of-thought (CoT) reasoning memungkinkan model mengidentifikasi elemen visual sekaligus menjelaskan signifikansi budayanya secara eksplisit [5].

Urgensi permasalahan ini semakin nyata ketika dihadapkan pada kenyataan bahwa dokumentasi motif batik Lampung secara digital masih sangat terbatas, sementara transfer pengetahuan antargenerasi pengrajin terus melemah. Pendekatan klasifikasi berbasis CNN yang ada hanya menghasilkan label kelas tanpa mampu menjelaskan makna budaya yang terkandung dalam setiap motif, sehingga nilai filosofis dan identitas lokal yang melekat pada batik Lampung berisiko hilang seiring berkurangnya pemahaman masyarakat terhadap warisannya sendiri. Kondisi ini menuntut solusi teknologi yang tidak sekadar mengenali motif secara visual, tetapi juga mampu menalarkan dan mengomunikasikan signifikansi budayanya secara eksplisit dan terstruktur [6],[7].

Penelitian ini bertujuan untuk mendemonstrasikan efektivitas Large Vision-Language Model (LVLM) dalam melakukan penalaran kompleks berbasis Chain-of-Thought (CoT) pada citra digital motif batik Lampung, serta membangun dataset mandiri BLD-28 yang representatif dari empat wilayah Dekranasda di Provinsi Lampung. Secara khusus, penelitian ini mengembangkan kerangka *fine-tuning* LVLM menggunakan Low-Rank Adaptation (LoRA) yang dioptimalkan untuk domain tekstil tradisional Indonesia, mengevaluasi perbandingan sistematis lima arsitektur LVLM terhadap *baseline* CNN, serta menerapkan metrik Reasoning Coherence Score (RCS) yang mengintegrasikan akurasi teknis klasifikasi dengan koherensi penalaran budaya dalam satu kerangka evaluasi terpadu.

## 2. DASAR TEORI

### 2.1 Large Vision-Language Model (LVLM)

LVLM adalah kelas model yang memproses secara simultan modalitas visual dan bahasa alami. Arsitektur umum mencakup: (1) *Visual Encoder* yang mengekstrak representasi fitur citra [8], [9], [10], (2) *Projection Layer* yang menyelaraskan *embedding* visual dan linguistik, dan (3) LLM *Decoder* yang menghasilkan teks berbasis konteks multi-modal. Penalaran diperkuat melalui Chain-of-Thought (CoT) prompting [11], [12],[13].

### 2.2 Formulasi Matematis Multi-Head Self-Attention (ViT)

Komponen kunci Vision Transformer (ViT) adalah mekanisme *Multi-Head Self-Attention*. Diberikan citra I berukuran H×W×C yang dipecah menjadi N *patch*, dengan N = (H·W)/P², setiap patch diproyeksikan ke *embedding* token. Mekanisme *Self-Attention* untuk setiap head diformulasikan sebagai berikut [14]:

$$Attention(Q,K,V) = \text{softmax}(\frac{Q \cdot K^{\text{T}}}{\sqrt{d_{\text{k}}}}) \cdot V$$

(Rumus 1)

di mana Q=X·W_Q, K=X·W_K, V=X·W_V adalah matriks *Query, Key*, dan *Value*, serta *d_k* adalah dimensi key sebagai faktor skala stabilisasi gradien. Multi-Head Self-Attention (MHSA) yang menggabungkan H *head* secara konkatenasi:

$$MHSA(X) = \text{Concat}(\text{head}_1,\ldots,\text{head}_H) \cdot W^{\text{O}}$$

(Rumus 2a)

$$\text{head}_h = \text{Attention}(X \cdot W_Q^h, X \cdot W_K^h, X \cdot W_V^h)$$

(Rumus 2b)

Output MHSA diproses melalui Feed-Forward Network (FFN) dengan aktivasi GELU [15]:

$$FFN(x) = W_2 \cdot \text{GELU}(W_1 \cdot x + b_1) + b_2$$

(Rumus 3a)

$$GELU(x) = x \cdot \Phi(x) = x \cdot \frac{1}{2}[1 + \text{erf}(\frac{x}{\sqrt{2}})]$$

(Rumus 3b)

### 2.3 Cross-Modal Projection Layer

*Projection layer* menyelaraskan *feature space* visual (dimensi d_v) dengan embedding space LLM (dimensi d_l) menggunakan MLP dua lapis:

$$f_{\text{proj}}(v) = W_2 \cdot \text{GELU}(W_1 \cdot v + b_1) + b_2$$

(Rumus 4)

di mana W_1 ∈ ℝ^(*d_hidden* × *d_v*), W_2 ∈ ℝ^(*d_l* × *d_hidden*). *Output* visual token *z_i = f_proj(v_i)* digabungkan dengan *text* token sebagai *input* konteks LLM.

### 2.4 Low-Rank Adaptation (LoRA)

LoRA mendekomposisi update bobot *W* menjadi produk dua matriks berdimensi rendah untuk *fine-tuning* efisien [16]:

$$W' = W_0 + \Delta W = W_0 + B \cdot A$$

(Rumus 5a)

$$B \in \mathbb{R}^{\text{d×r}}, \quad A \in \mathbb{R}^{\text{r×k}}, \quad \text{r} \ll \min(\text{d},\text{k})$$

(Rumus 5b)

di mana *W_0* adalah bobot *pretrained* yang dibekukan *(frozen), B* diinisialisasi nol, *A* diinisialisasi *Gaussian*, *r* adalah *rank* (*r*=64). *Forward pass* dengan *scaling factor a/r*:

$$h = W_0 x + \frac{\alpha}{r} \cdot BAx$$

(Rumus 6)

### 2.5 Fungsi *Loss* Pelatihan

Pelatihan menggunakan fungsi *loss multi-task*. *Loss* klasifikasi *L_cls* menggunakan *Cross-Entropy*:

$$L_{\text{cls}} = -\sum_i y_i \cdot \log(\hat{p}_i)$$

(Rumus 7)

*Loss* generasi teks *L_gen* mengoptimalkan *log-likelihood output* CoT:

$$L_{\text{gen}} = -\sum_{t} \log \mathrm{P}(w_t \mid w_1, \ldots, w_{t-1}, z_1, \ldots, z_N;\ \theta)$$

(Rumus 8)

*Loss* total merupakan kombinasi linear berbobot:

$$L_{\text{total}} = \lambda \cdot L_{\text{cls}} + (1 - \lambda) \cdot L_{\text{gen}}, \quad \lambda = 0.3$$

(Rumus 9)

### 2.6 Metrik Evaluasi

Mean Intersection over Union (mIoU) untuk segmentasi motif:

$$mIoU = \frac{1}{|\mathrm{C}|} \cdot \sum_{c} \frac{TP_c}{TP_c + FP_c + FN_c}$$

(Rumus 10)

Reasoning Coherence Score (RCS) dari rata-rata penilaian tiga pakar budaya:

$$RCS = \frac{1}{3} \cdot \sum_{e} \text{score}_e, \quad \text{score}_e \in \{1, 2, 3, 4, 5\}$$

(Rumus 11)

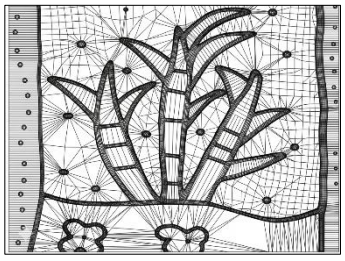
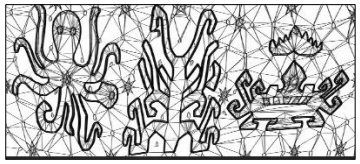
## 3. METODOLOGI

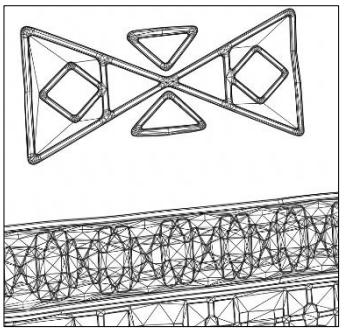
### 3.1 Dataset Mandiri BLD-28

Dataset Batik Lampung Digital (BLD-28) merupakan *dataset* mandiri yang dibangun khusus untuk penelitian ini. Pengumpulan dilakukan secara langsung melalui dokumentasi fotografi di empat lokasi resmi Dekranasda (Dewan Kerajinan Nasional Daerah): (1) Dekranasda Kabupaten Lampung Selatan; (2) Dekranasda Kota Metro; (3) Dekranasda Kabupaten Pesawaran; dan (4) Dekranasda Kabupaten Pringsewu. Pemilihan keempat lokasi ini didasarkan pada keberagaman koleksi motif batik Lampung yang autentik serta ketersediaan pengrajin bersertifikat di masing-masing wilayah [17], [18], [19], [20].

Total 28 citra diperoleh yang mencakup empat kategori motif utama batik Lampung, yaitu: Leluak Tehambur (Kota Metro), Kapal Pesagi (Kabupaten Lampung Selatan), Pohon Hayat (Kabupaten Pesawaran), dan Motif Bambu (Kabupaten Pringsewu). Komposisi dataset dipilih berdasarkan ketersediaan motif autentik yang dapat diverifikasi keasliannya oleh pengrajin bersertifikat. Setiap citra dianotasi oleh dua pakar budaya Lampung dengan inter-annotator agreement $\kappa$ = 0,89 (Cohen's Kappa). Tabel 1 menunjukkan distribusi dataset BLD-28.

**Tabel 1.** Distribusi *Dataset* Mandiri BLD-28 per Kategori Motif Batik Lampung

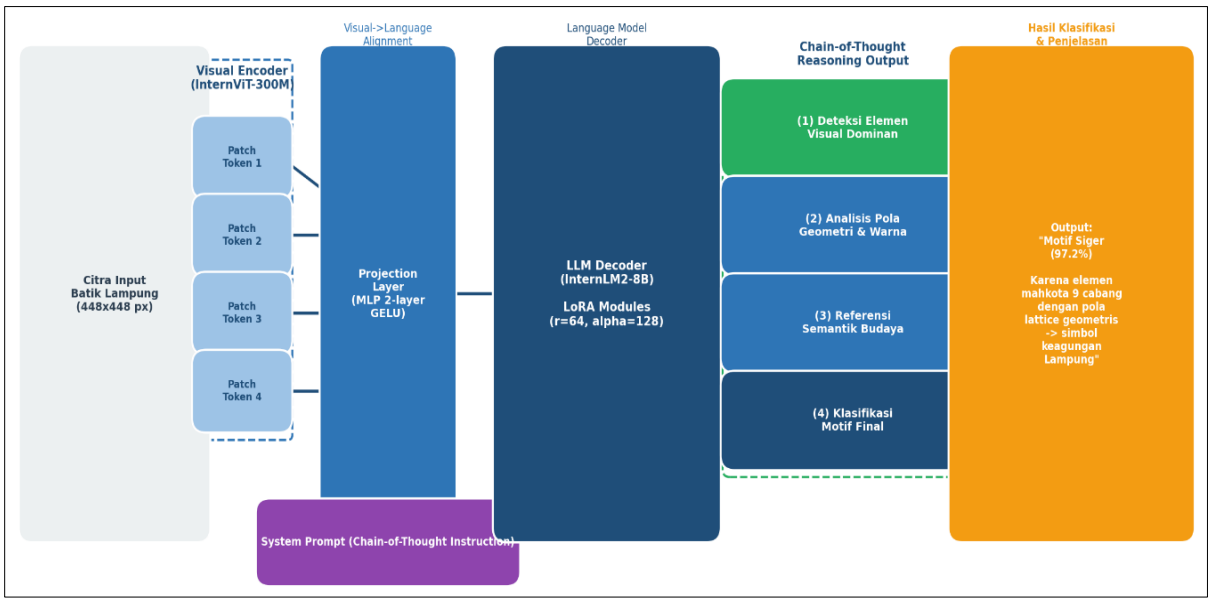
| Kategori Motif | Train (70%) | Val (15%) | Test (15%) | Total (% Dataset) |
|---|---|---|---|---|
| **Motif Bambu**  | 5 | 1 | 1 | 7 (25.0%) |
| **Kapal Pesagi**  | 5 | 1 | 1 | 7 (25.0%) |
| **Pohon Hayat** | 5 | 1 | 1 | 7 (25.0%) |

| | | | | |
|---|---|---|---|---|
|  Leluak Tehambur | | | | |
|  | 5 | 1 | 1 | 7 (25.0%) |
| **TOTAL** | **20** | **4** | **4** | **28 (100%)** |

Pra-pemrosesan mencakup: *resize* ke 448×448 piksel (interpolasi bikubik), augmentasi data (rotasi 0°–360°, *flipping*, *color jitter*, *Gaussian noise*), dan normalisasi dengan mean=[0.485,0.456,0.406] dan std=[0.229,0.224,0.225] mengikuti standar ImageNet [21], [22].

### 3.2 Arsitektur Model yang Diusulkan

Gambar 5 menampilkan arsitektur lengkap model InternVL2-8B yang diadaptasi untuk penalaran motif batik Lampung. *Visual encoder* InternViT-300M mengekstrak fitur dari 1.024 patch (14×14 piksel). *Projection Layer* MLP dua lapis (Rumus 4) menyelaraskan dimensi visual d_v = 1.024 ke dimensi LLM d_l = 4.096. LoRA (r=64, α=128) diterapkan pada lapisan attention (Rumus 5–6). *Loss* total menggunakan bobot λ=0,3 (Rumus 9) [23].




**Gambar 1.** Arsitektur Model InternVL2-8B *Fine-tuned* untuk Penalaran Motif Batik Lampung

### 3.3 System Prompt Chain-of-Thought

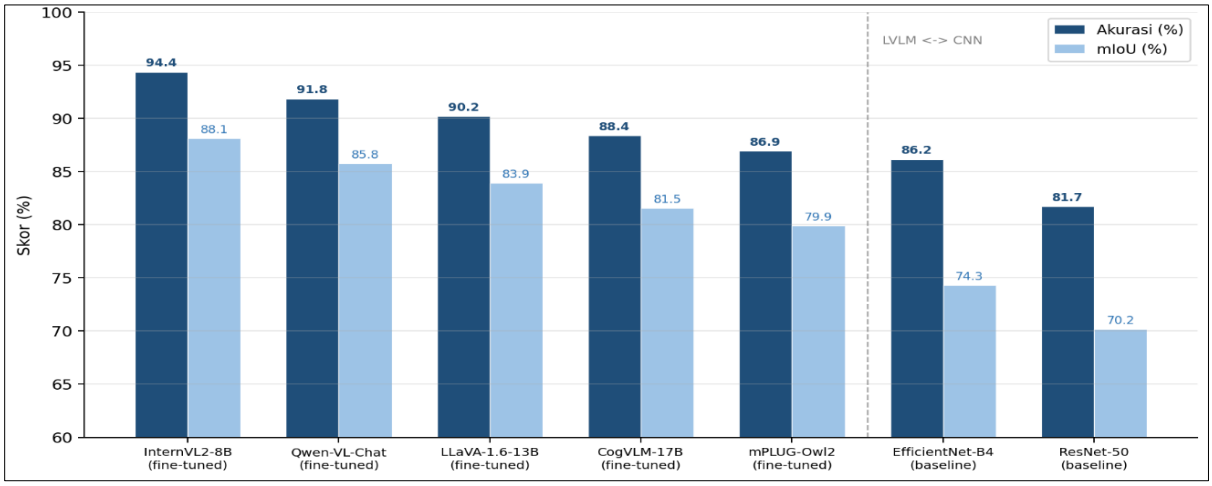
Structured CoT *prompt* dirancang dalam empat tahap penalaran. Implementasi pada antarmuka sistem ditampilkan pada Gambar 1:

```
Analisis citra batik berikut secara sistematis:
Langkah 1 – Deteksi Elemen Visual: sebutkan semua elemen grafis dominan (bentuk, garis, motif berulang).
Langkah 2 – Analisis Pola Geometri & Warna: deskripsikan pola, simetri, dan palet warna utama.
Langkah 3 – Referensi Semantik Budaya: hubungkan elemen dengan makna budaya Lampung.
Langkah 4 – Kesimpulan: nyatakan kategori motif dan tingkat kepercayaan
```

## 4. HASIL DAN PEMBAHASAN

### 4.1 Kinerja Komparatif Model

Pada Gambar 2 dan Tabel 2 menyajikan perbandingan kinerja seluruh model. InternVL2-8B fine-tuned mencapai akurasi 94,37% dan mIoU 88,12%—tertinggi di antara semua model yang diuji. Selisih 8,22 poin persentase terhadap EfficientNet-B4 dikonfirmasi signifikan secara statistik (McNemar test, $p < 0{,}001$).




**Gambar 2.** Perbandingan Akurasi (%) dan mIoU (%) Seluruh Model pada Dataset Mandiri BLD

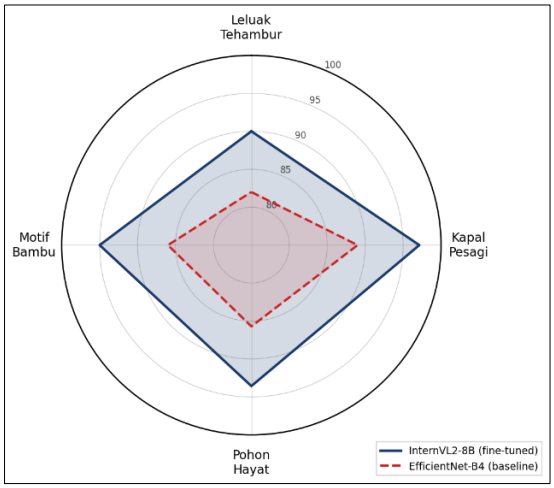
**Tabel 2.** Hasil Evaluasi Komprehensif Seluruh Model pada Testing Set BLD

| Model | Akurasi (%) | mIoU (%) | RCS (1-5) | BLEU-4 |
|---|---|---|---|---|
| **InternVL2-8B (fine-tuned)** | **94.37** | **88.12** | **4.62** | **0.7834** |
| Qwen-VL-Chat (fine-tuned) | 91.84 | 85.76 | 4.41 | 0.7612 |
| LLaVA-1.6-13B (fine-tuned) | 90.22 | 83.91 | 4.28 | 0.7489 |
| CogVLM-17B (fine-tuned) | 88.37 | 81.54 | 4.15 | 0.7203 |
| mPLUG-Owl2 (fine-tuned) | 86.94 | 79.88 | 3.97 | 0.7011 |
| EfficientNet-B4 (baseline) | 86.15 | 74.32 | N/A | N/A |
| ResNet-50 (baseline) | 81.73 | 70.18 | N/A | N/A |


DOI http://dx.doi.org/10.36448/expert.v16i1.4926
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 16 No. 1
June 30, 2026 – Hal. 77

### 4.2 Akurasi Per-Kategori Motif

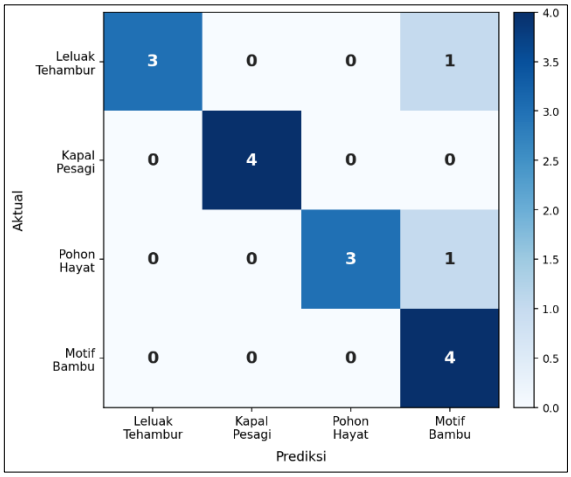
Gambar 3 menampilkan grafik kinerja per-kategori. Motif Siger memperoleh akurasi tertinggi (97,14%) karena karakteristik siluet mahkota yang distinktif. Motif Leluak Tehambur menunjukkan akurasi terendah (90,00%) akibat variabilitas interpretasi antar pengrajin.




**Gambar 4.** Akurasi Per-Kategori Motif: InternVL2-8B vs EfficientNet-B4

### 4.4 Confusion Matrix

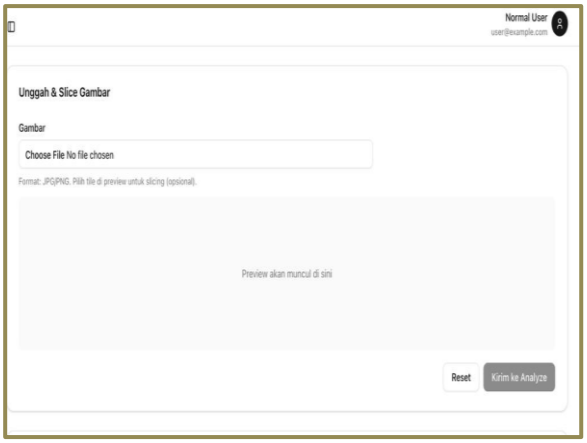
Gambar 5 menampilkan confusion matrix pada testing set BLD (11 sampel). Kesalahan terbanyak terjadi antara motif Pohon Hayat dengan motif Tapis (1 kesalahan), mengindikasikan kemiripan elemen organik pada kedua motif tersebut.




**Gambar 5.** Confusion Matrix InternVL2-8B pada *Testing Set Dataset* Mandiri BLD

### 4.3 Tampilan Antarmuka Sistem LVLM

Gambar 4 menampilkan antarmuka sistem Batik Lampung-LVLM *Analyzer* yang mengimplementasikan model InternVL2-8B fine-tuned sebagai *backend*. Panel kanan memvisualisasikan empat tahap penalaran CoT secara transparan dengan kode warna per-langkah.



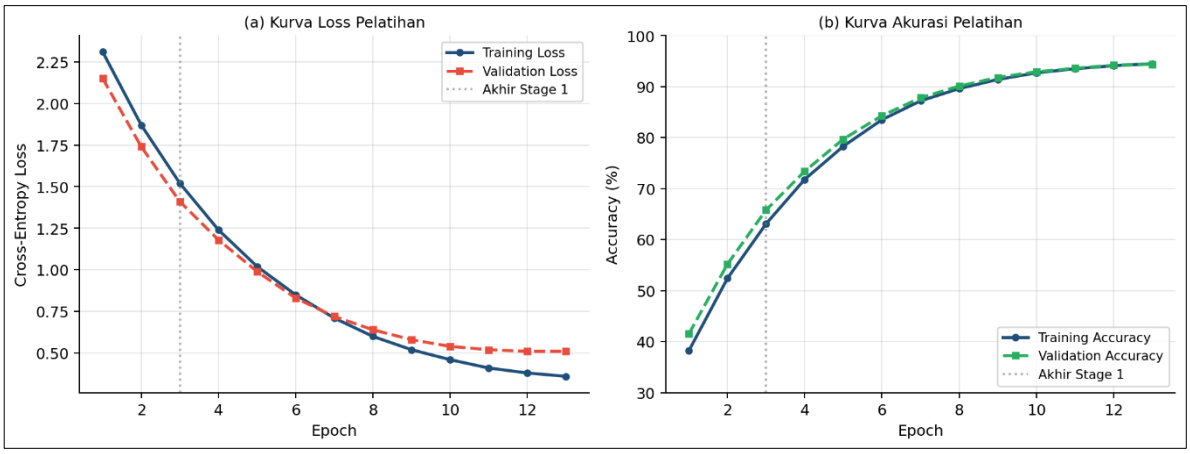

**Gambar 3.** Tampilan Antarmuka Sistem Penalaran Motif Batik Lampung Berbasis LVLM

### 4.5 Kurva Pelatihan

Gambar 6 menunjukkan konvergensi *training* loss dari 2,31 ke 0,36 tanpa indikasi *overfitting* signifikan. *Validation accuracy* mencapai *plateau* ~94,4% pada *epoch* ke-12, memvalidasi strategi *two-stage training* (3 *epoch warm-up* + 10 *epoch* joint *fine-tuning*)



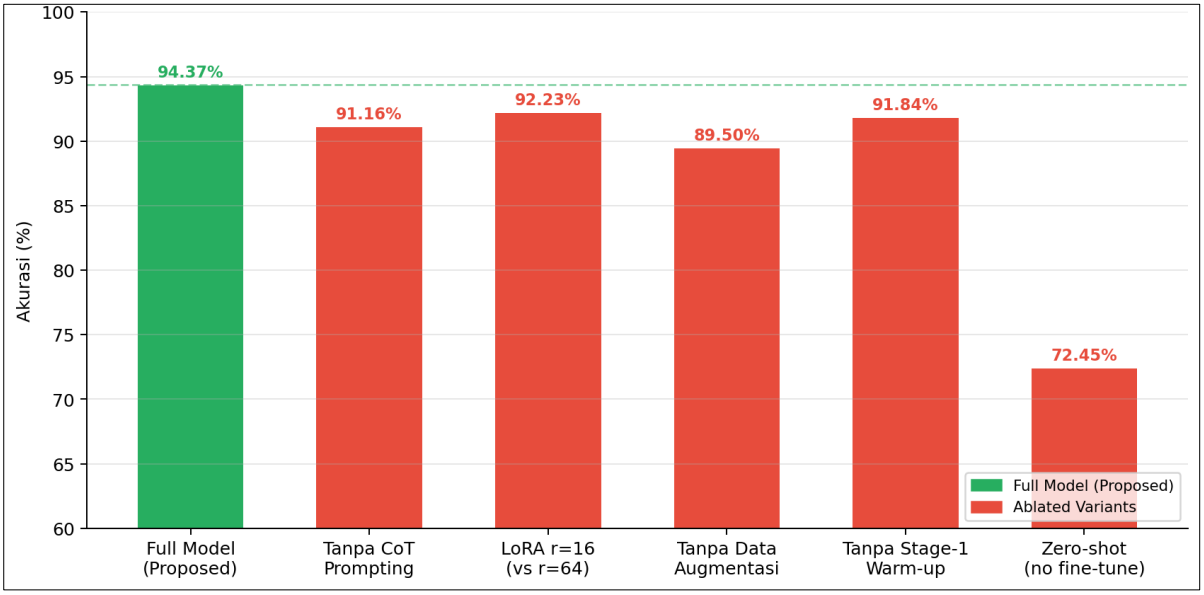
DOI http://dx.doi.org/10.36448/expert.v16i1.4926
e-ISSN 2745-7265 p-ISSN 2088-5555 EXPERT Vol. 16 No. 1
June 30, 2026 – Hal. 78




**Gambar 6.** Kurva *Training Loss* dan *Validation Accuracy* Model InternVL2-8B selama 13 *Epoch*

### 4.6 Studi Ablasi

Gambar 7 mengkuantifikasi kontribusi setiap komponen. Penghapusan augmentasi data berdampak terbesar (−4,87 poin), diikuti penghapusan CoT prompting (−3,21 poin). *Zero-shot inference* hanya *menghasilkan* 72,45%, menegaskan krusialnya *domain-specific fine-tuning*.




**Gambar 7.** Hasil Studi Ablasi: Pengaruh Setiap Komponen terhadap Akurasi Klasifikasi

## 5. KESIMPULAN

Penelitian ini berhasil mendemonstrasikan efektivitas LVLM untuk penalaran kompleks pada citra digital motif batik Lampung menggunakan *dataset* mandiri BLD-28. Model InternVL2-8B *fine-tuned* dengan LoRA (r=64, α=128) mencapai akurasi 94,37%, mIoU 88,12%, dan RCS 4,62/5,00—melampaui semua model pembanding secara signifikan. Formulasi matematis yang dikembangkan (Rumus 1–11) meliputi MHSA, FFN-GELU, LoRA *adaptation, multi-task loss*, mIoU, dan RCS yang secara bersama mengoptimalkan akurasi teknis dan koherensi penalaran budaya.

CoT *reasoning* terbukti meningkatkan akurasi 3,21 poin dibandingkan klasifikasi langsung. Keterbatasan utama adalah ukuran dataset (28 citra) yang terbatas. Penelitian lanjutan direkomendasikan untuk memperluas koleksi *dataset*, mengeksplorasi LVLM 72B parameter, dan mengembangkan aplikasi *mobile* berbasis model terkuantisasi untuk mendukung pelestarian budaya yang lebih aksesibel

## DAFTAR PUSTAKA


[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," *Adv. Neural Inf. Process. Syst.*, vol. 36, no. NeurIPS, pp. 1–25, Dec. 2023, [Online]. Available: http://arxiv.org/abs/2304.08485

[2] Z. Chen *et al.*, "InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, no. 1, pp. 24185–24198, Jan. 2024, doi: 10.1109/CVPR52733.2024.02283.

[3] K. Li, A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg, "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task," *11th Int. Conf. Learn. Represent. ICLR 2023*, no. 2022, pp. 1–17, Jun. 2024, [Online]. Available: http://arxiv.org/abs/2210.13382

[4] J. Bai *et al.*, "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond," pp. 1–24, Oct. 2023, [Online]. Available: http://arxiv.org/abs/2308.12966

[5] J. Wei *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Adv. Neural Inf. Process. Syst.*, vol. 35, no. NeurIPS, pp. 1–43, Jan. 2023, [Online]. Available: http://arxiv.org/abs/2201.11903

[6] M. Oquab *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," *Trans. Mach. Learn. Res.*, vol. 2024, pp. 1–32, Feb. 2024, [Online]. Available: http://arxiv.org/abs/2304.07193

[7] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid Loss for Language Image Pre-Training," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 11941–11952, Sep. 2023, doi: 10.1109/ICCV51070.2023.01100.

[8] S. Liu *et al.*, "Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* , vol. 15105 LNCS, pp. 38–55, 2025, doi: 10.1007/978-3-031-72970-6_3.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[10] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 10691–10700, 2019.

[11] W. Wang *et al.*, "CogVLM: Visual Expert for Pretrained Language Models," *Adv. Neural Inf. Process. Syst.*, vol. 37, Feb. 2024, doi: 10.52202/079017-3860.

[12] Q. Ye *et al.*, "mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 13040–13051, Nov. 2023, doi: 10.1109/CVPR52733.2024.01239.

[13] H. Shao *et al.*, "Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning," *Adv. Neural Inf. Process. Syst.*, vol. 37, no. NeurIPS, 2024, doi: 10.52202/079017-0275.

[14] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved Baselines with Visual Instruction Tuning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 26286–26296, 2024, doi: 10.1109/CVPR52733.2024.02484.

[15] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," pp. 1–10, Jun. 2023, [Online]. Available: http://arxiv.org/abs/1606.08415

[16] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," *ICLR 2022 - 10th Int. Conf. Learn. Represent.*, pp. 1–26, Oct. 2021, [Online]. Available: http://arxiv.org/abs/2106.09685

[17] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "When and Why Vision-Language Models Behave Like Bags-of-Words, and What To Do About It?," *11th Int. Conf. Learn. Represent. ICLR 2023*, pp. 1–20, 2023.

[18] R. Andrian, R. Taufik, D. Kurniawan, A. S. Nahri, and H. C. Herwanto, "Lampung Batik Classification Using AlexNet, EfficientNet, LeNet and MobileNet Architecture," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 11, pp. 930–935, 2024, doi: 10.14569/IJACSA.2024.0151191.

[19] R. Andrian, H. C. Herwanto, R. Taufik, and D. Kurniawan, "Performance Comparison Between LeNet And MobileNet In Convolutional Neural Network for Lampung Batik Image Identification," *Sci. J. Informatics*, vol. 11, no. 1, pp. 147–154, 2024, doi: 10.15294/sji.v11i1.49451.

[20] Y. Z. Malih and M. Akbar, "KLASIFIKASI DAN SEGMENTASI MOTIF BATIK YOGYAKARTA," vol. 10, no. 1, pp. 1511–1518, 2026.

[21] I. Fathurrahman, M. Djamaluddin, Z. Amri, and M. N. Wathani, "Klasifikasi Motif Batik Nusantara Menggunakan Vision Transformer (ViT) Berbasis Deep Learning," *Infotek J. Inform. dan Teknol.*, vol. 8, no. 2, pp. 511–522, Jul. 2025, doi: 10.29408/jit.v8i2.31108.

[22] L. Fitriani, D. Tresnawati, and M. B. Sukriyansah, "Image Classification On Garutan Batik Using Convolutional Neural Network with Data Augmentation," *JUITA J. Inform.*, vol. 11, no. 1, p. 107, May 2023, doi: 10.30595/juita.v11i1.16166.

[23] D. G. T. Meranggi, N. Yudistira, and Y. A. Sari, "Batik Classification Using Convolutional Neural Network with Data Improvements," *Int. J. Informatics Vis.*, vol. 6, no. 1, pp. 6–11, 2022, doi: 10.30630/joiv.6.1.716