Terbit online pada laman web jurnal: **http://journal.isas.or.id/index.php/JASENS**

JOURNAL OF APPLIED SMART ELECTRICAL NETWORK AND SYSTEMS (JASENS)

Vol. 7 No. 1 (2026) 40 - 45 ISSN Media Elektronik: 2723-5467

Pengaruh Prompt Teks pada Model CLIP terhadap Akurasi Estimasi Kepadatan Kerumunan

Mas Nurul Achmadiah[1], Anindya Dwi Risdhayanti[2]
[1,2] Jurusan Teknik Elektro, Politeknik Negeri Malang
[1]Electro-Optic Department, National Formosa University, Huwei Township, Yunlin County, Taiwan
[1]masnurul@polinema.ac.id , [2]risdhayanti@polinema.ac.id

Abstract

The text prompt in a CLIP-conditioned counting model is not merely an auxiliary input because its embedding is used as cross-attention context for the diffusion U-Net. This study evaluates prompt conditions as the main experimental variable and quantifies their effect on crowd-density estimation. The scientific contribution is an ablation on ShanghaiTech Part A that separates four conditions—no prompt, a standard crowd-description prompt, a visual-only token, and an unrelated random prompt—followed by benchmark comparison with language-based and non-language counting methods. Relative to the prompt-free baseline (MAE 58.21; RMSE 104.92), the standard prompt reduces MAE by 11.72% and RMSE by 10.06%, whereas the visual-only token reduces MAE by 6.08% and the random prompt by 2.37%. The 5.64-percentage-point gap between the standard prompt and visual-only token is treated as an observed difference rather than a causal estimate because repeated-run variability is not reported. Against CLIP-EBC on the three shared benchmarks, MAE improves from 52.5 to 51.39 on ShanghaiTech Part A, 6.6 to 5.79 on Part B, and 80.3 to 75.19 on UCF-QNRF. The largest absolute improvement is 5.11 MAE on UCF-QNRF, while the largest relative improvement is 12.27% on ShanghaiTech Part B. These results support a measurable prompt effect on ShanghaiTech Part A and show competitive cross-dataset performance without generalizing the prompt ablation to all datasets.

Keywords: text prompt; CLIP; crowd counting; density estimation; cross-dataset evaluation

Abstrak

Prompt teks pada model penghitungan berbasis CLIP bukan sekadar masukan pelengkap karena embedding teks digunakan sebagai konteks cross-attention pada U-Net difusi. Penelitian ini menempatkan kondisi prompt sebagai variabel eksperimen utama dan mengukur pengaruhnya terhadap estimasi kepadatan kerumunan. Kontribusi ilmiah penelitian adalah ablation pada ShanghaiTech Part A yang memisahkan empat kondisi, yaitu tanpa prompt, prompt baku berupa deskripsi kerumunan, token visual tanpa teks, dan prompt acak yang tidak terkait tugas, kemudian dilengkapi dengan perbandingan terhadap metode penghitungan berbasis bahasa dan non-bahasa. Terhadap baseline tanpa prompt (MAE 58,21; RMSE 104,92), prompt baku menurunkan MAE sebesar 11,72% dan RMSE 10,06%, sedangkan token visual tanpa teks menurunkan MAE 6,08% dan prompt acak 2,37%. Selisih 5,64 titik persentase antara prompt baku dan token visual diperlakukan sebagai perbedaan observasional, bukan estimasi kausal, karena variasi antar-pengulangan tidak dilaporkan. Terhadap CLIP-EBC pada tiga dataset yang sama-sama memiliki nilai pembanding, MAE berubah dari 52,5 menjadi 51,39 pada ShanghaiTech Part A, 6,6 menjadi 5,79 pada Part B, dan 80,3 menjadi 75,19 pada UCF-QNRF. Perbaikan absolut terbesar sebesar 5,11 MAE terjadi pada UCF-QNRF, sedangkan perbaikan relatif terbesar sebesar 12,27% terjadi pada ShanghaiTech Part B. Hasil ini mendukung adanya pengaruh prompt yang terukur pada ShanghaiTech Part A dan menunjukkan kinerja lintas dataset yang kompetitif tanpa menggeneralisasi hasil ablation prompt ke seluruh dataset.

Kata kunci: prompt teks; CLIP; penghitungan kerumunan; estimasi kepadatan; evaluasi lintas dataset

Diterima Redaksi : 06-05-2026 | Selesai Revisi : 23-06-2026 | Diterbitkan Online : 30-06-2026

1. Pendahuluan

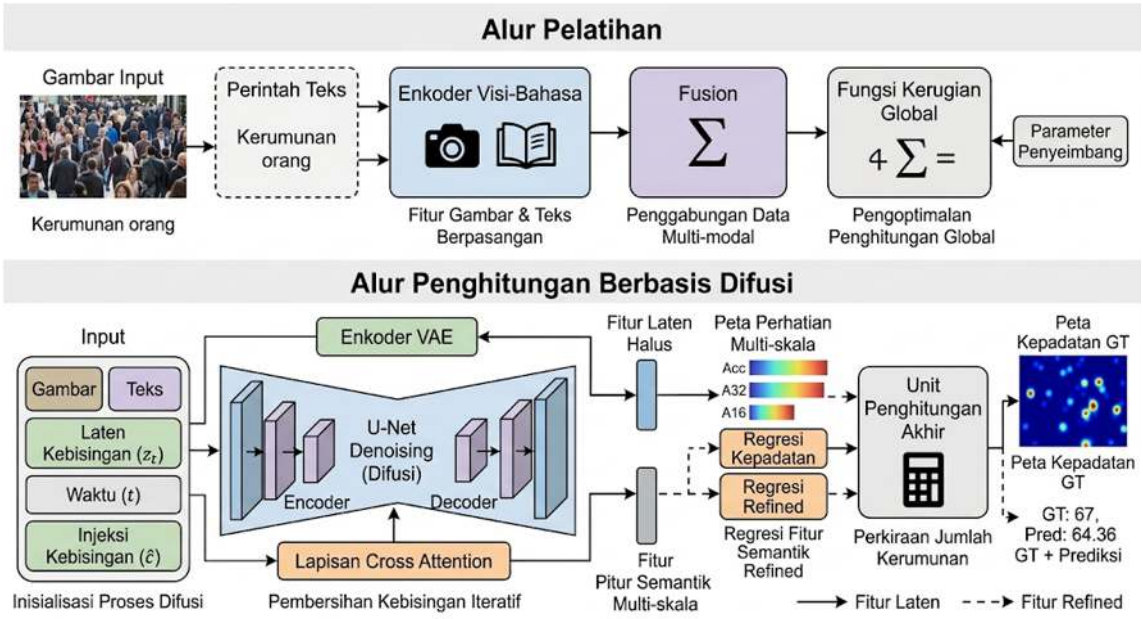
Estimasi kepadatan kerumunan merupakan bagian penting dari sistem pemantauan ruang publik dan keselamatan transportasi [1]. Pendekatan berbasis regresi peta kepadatan seperti MCNN [2] dan CSRNet [3] menjadi landasan bagi berbagai metode berikutnya, termasuk reparameterisasi struktural [4] dan

pencocokan distribusi [5]. Perkembangan vision-language model memperluas ruang desain dengan membawa informasi semantik ke proses penghitungan. CLIP [6] menyediakan representasi bersama antara citra dan bahasa; CLIP-EBC [7] memanfaatkan CLIP untuk penghitungan melalui enhanced blockwise classification, sedangkan CountGD [8] mengembangkan penghitungan multimodal dunia terbuka. Berbeda dari fokus kinerja model secara keseluruhan pada pendekatan tersebut, penelitian ini menempatkan kondisi prompt sebagai variabel evaluasi utama pada arsitektur penghitungan berbasis difusi yang menggunakan embedding teks sebagai konteks cross-attention.

Research gap penelitian terletak pada belum terpisahkannya secara eksplisit pengaruh kondisi bahasa dari konteks visual pada arsitektur yang menggunakan embedding teks sebagai konteks cross-attention. Untuk itu, penelitian ini membandingkan kondisi tanpa prompt, prompt baku, token visual tanpa teks, dan prompt acak pada ShanghaiTech Part A. Evaluasi tersebut kemudian ditempatkan dalam konteks kinerja model pada ShanghaiTech Part A, ShanghaiTech Part B, UCF-QNRF, dan JHU-Crowd++. Dengan rancangan ini, kontribusi penelitian bukan pada penggunaan CLIP semata, tetapi pada pengukuran ablation kondisi prompt serta interpretasi hubungan antara pengarahan semantik, supervisi spasial, dan galat penghitungan. Generalisasi pengaruh prompt dibatasi pada dataset tempat ablation dilakukan, sedangkan hasil lintas dataset digunakan sebagai pembanding kinerja model.

2. Metode Penelitian

2.1. Blok diagram



Gambar 1. Diagram Arsitektur untuk analisa Pengaruh Prompt Teks pada Model CLIP terhadap Akurasi Estimasi Kepadatan Kerumunan

Diagram arsitektur pada Gambar 1 membagi proses menjadi alur pelatihan dan alur penghitungan berbasis difusi. Pada pelatihan, citra dan prompt diproses oleh encoder visi-bahasa untuk membentuk fitur multimodal yang digunakan bersama fungsi kerugian penghitungan. Pada alur inferensi, fitur laten dari encoder VAE diproses oleh U-Net denoising dengan cross-attention, kemudian fitur laten dan fitur semantik multi-skala diteruskan ke regresi kepadatan dan unit penghitungan akhir. Posisi prompt pada jalur cross-attention menjadikan perubahan konteks teks berpotensi mengubah fokus semantik yang diterima kepala regresi, sehingga efek prompt dievaluasi secara terpisah pada rancangan pengujian.

Pada alur penghitungan berbasis difusi, sistem menerima citra, konteks teks, laten kebisingan, dan indeks waktu difusi. U-Net menghasilkan fitur laten yang selanjutnya digabungkan dengan fitur semantik multi-skala sebelum regresi peta kepadatan. Hasil akhir berupa peta kepadatan dan estimasi jumlah orang. Dalam eksperimen, encoder CLIP dan U-Net difusi dibekukan sehingga evaluasi difokuskan pada konteks prompt dan komponen ringan yang dipelajari.

2.2. Alur pemrosesan prompt

Prompt diproses oleh encoder teks CLIP yang dibekukan untuk menghasilkan embedding tekstual. Embedding tersebut digabungkan dengan embedding visual melalui proyeksi terpelajar ringan, lalu diproyeksikan ke dimensi konteks difusi dan digunakan sebagai token cross-attention pada lapisan U-Net. Prompt baku mengikuti deskripsi kerumunan yang

ditunjukkan pada diagram arsitektur sebagai "Kerumunan orang". Kondisi pembanding terdiri atas tanpa prompt, token visual tanpa teks, dan prompt acak yang tidak berkaitan dengan tugas. Karena encoder teks bawaan model difusi dilewati, perubahan konteks ini langsung memengaruhi informasi semantik yang masuk ke cross-attention.

Peta kepadatan diperoleh melalui regresi konvolusi ringan atas fitur laten yang telah disatukan dengan grid semantik, kemudian jumlah orang dihitung melalui integrasi spasial sebagaimana rumus (1).

$$C = \Sigma\, D(x,y), \text{ untuk seluruh koordinat } (x,y) \quad (1)$$

dengan D adalah peta kepadatan prediksi dan C adalah jumlah orang hasil estimasi. Supervisi pelatihan menggabungkan kerugian kepadatan tingkat piksel, kerugian penghitungan global, dan regularisasi struktural sebagaimana rumus (2).

$$L_{total} = L_{dens} + L_{global} + L_{struct} \quad (2)$$

dengan L_{dens} adalah kerugian regresi kepadatan, L_{global} adalah kerugian penghitungan global yang memuat komponen kerugian relatif, dan L_{struct} adalah gabungan regularisasi kehalusan lokal dan penyelarasan tingkat grid. Bobot masing-masing komponen disesuaikan menurut karakteristik dataset.

2.3. Rancangan pengujian

Pengujian pertama merupakan ablation prompt pada ShanghaiTech Part A. Kondisi tanpa prompt digunakan sebagai baseline dengan MAE 58,21 dan RMSE 104,92. Tiga kondisi pembanding adalah prompt baku berupa deskripsi kerumunan, token visual tanpa teks, dan prompt acak yang tidak terkait tugas. Seluruh hasil pada Tabel 1 dibaca terhadap baseline yang sama sehingga perubahan galat merepresentasikan perbedaan antar-kondisi bahasa pada kerangka evaluasi yang sama. Pengujian ini menghasilkan estimasi titik; tidak digunakan uji signifikansi karena variasi antar-pengulangan tidak tersedia.

Pengujian kedua membandingkan MAE model dengan metode berbasis bahasa dan non-bahasa pada ShanghaiTech Part A, ShanghaiTech Part B, UCF-QNRF, dan JHU-Crowd++. Perbandingan langsung dengan CLIP-EBC hanya dilakukan pada tiga dataset yang memiliki angka CLIP-EBC, yaitu ShanghaiTech Part A, Part B, dan UCF-QNRF. Pengujian ketiga mendeskripsikan konfigurasi bobot supervisi yang tersedia untuk beberapa dataset. Seluruh eksperimen dijalankan pada GPU NVIDIA RTX 5090 dengan encoder CLIP dan U-Net difusi dibekukan. MAE dan RMSE digunakan pada ablation prompt, sedangkan Tabel 2 menggunakan MAE karena sebagian metode pembanding tidak menyediakan RMSE pada seluruh benchmark.

Ruang lingkup konfigurasi yang dapat direproduksi dari data eksperimen meliputi perangkat komputasi (NVIDIA RTX 5090), pembekuan encoder CLIP dan U-Net difusi, empat kondisi bahasa pada ablation ShanghaiTech Part A, serta metrik MAE/RMSE. Dataset pembanding mencakup ShanghaiTech Part A/B, UCF-QNRF, dan JHU-Crowd++. Parameter rinci seperti seed, jumlah pengulangan, pembagian data yang digunakan pada setiap run, batch size, epoch, optimizer, dan string literal prompt acak tidak digunakan untuk menyusun klaim baru dalam penelitian ini; dampaknya terhadap reproduksibilitas dibahas pada bagian keterbatasan.

2.4. Metrik evaluasi

Ketelitian dinilai menggunakan MAE sebagaimana rumus (3) dan RMSE sebagaimana rumus (4), yang keduanya lazim dipakai sebagai tolok ukur pada penelitian penghitungan kerumunan.

$$MAE = (1/N)\, \Sigma\, |Ci - Ci,acuan| \quad (3)$$

$$RMSE = \sqrt{[(1/N)\, \Sigma\, (Ci - Ci,acuan)^2]} \quad (4)$$

dengan N adalah jumlah citra uji, C_i adalah jumlah orang hasil estimasi, dan $C_{i_{acuan}}$ adalah jumlah orang menurut anotasi. Karena penelitian ini menitikberatkan perbandingan antar kondisi prompt, hasil juga dinyatakan dalam bentuk penurunan galat relatif agar besaran manfaat tidak tersamar oleh perbedaan skala kepadatan antar dataset.

3. Hasil dan Pembahasan

3.1. Perolehan akurasi dari prompt teks

Tabel 1 menyajikan nilai galat dan penurunan relatif terhadap kondisi tanpa prompt. Nilai MAE/RMSE untuk tiga kondisi prompt dihitung dari baseline dan persentase penurunan yang tersedia, kemudian dibulatkan dua desimal; nilai MAE prompt baku konsisten dengan hasil 51,39 pada perbandingan benchmark.

Tabel 1. Hasil ablation kondisi prompt pada ShanghaiTech Part A

Kondisi bahasa	MAE	RMSE	Penurunan MAE (%)	Penurunan RMSE (%)
Tanpa prompt	58,21	104,92	0,00	0,00
Prompt baku	51,39	94,37	11,72	10,06
Token visual tanpa teks	54,67	98,13	6,08	6,47
Prompt acak	56,83	101,47	2,37	3,29

Prompt baku memberikan galat terendah pada ablation ShanghaiTech Part A, dengan penurunan MAE 11,72% dan RMSE 10,06% terhadap tanpa prompt. Token visual tanpa teks menghasilkan penurunan MAE 6,08%, sedangkan prompt acak 2,37%. Secara

arsitektural, urutan ini konsisten dengan fungsi embedding teks sebagai konteks cross-attention: prompt yang relevan memberi batasan semantik yang mengarahkan perhatian pada konsep kerumunan, token visual masih menyediakan konteks tanpa deskripsi eksplisit, sedangkan prompt acak memasukkan konteks yang tidak selaras dengan tugas. Selisih penurunan MAE antara prompt baku dan token visual adalah 5,64 titik persentase. Karena hasil yang tersedia berupa estimasi titik tanpa variasi antar-run, selisih tersebut ditafsirkan sebagai gap observasional, bukan sebagai kontribusi kausal murni informasi bahasa.

3.2. Perbandingan dengan metode berbasis bahasa dan non-bahasa

Tabel 2 membandingkan hasil model dengan beberapa metode penghitungan pada empat dataset. Kelompok pertama merupakan metode tanpa modalitas bahasa, sedangkan kelompok kedua merupakan metode yang memanfaatkan model vision-language.

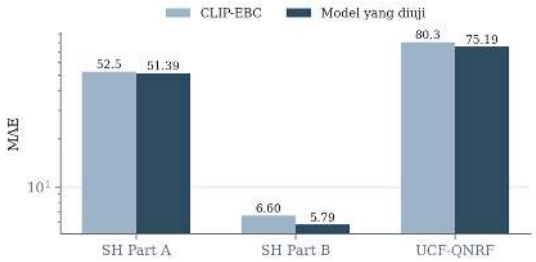
Tabel 2. Perbandingan galat pada empat dataset tolok ukur

Metode	SH Part A	SH Part B	UCF-QNRF	JHU-Crowd++
MCNN (tanpa bahasa) [2]	110,2	26,4	243,5	-
CSRNet (tanpa bahasa) [3]	68,2	10,6	120,3	-
DM-Count (tanpa bahasa) [5]	59,7	7,4	85,6	-
DSGC-Net (tanpa bahasa) [9]	48,9	5,9	79,3	-
LIMM (tanpa bahasa) [13]	50,8	6,5	76,4	53,0
CLIP-EBC (berbasis bahasa) [7]	52,5	6,6	80,3	-
Model yang diuji (berbasis bahasa)	51,39	5,79	75,19	56,21

Perbandingan langsung dengan CLIP-EBC tersedia pada tiga benchmark bersama. Pada ShanghaiTech Part A, MAE turun 1,11 poin dari 52,5 menjadi 51,39 atau 2,11% relatif. Pada ShanghaiTech Part B, penurunan absolutnya 0,81 poin dari 6,6 menjadi 5,79, tetapi secara relatif merupakan perbaikan terbesar, yaitu 12,27%. Pada UCF-QNRF, penurunan absolut terbesar mencapai 5,11 poin dari 80,3 menjadi 75,19 atau 6,36% relatif. Dengan demikian, dasar interpretasi harus dibedakan: UCF-QNRF memberikan perbaikan absolut terbesar, sedangkan ShanghaiTech Part B memberikan perbaikan relatif terbesar. CLIP-EBC tidak memiliki nilai pada JHU-Crowd++ di Tabel 2, sehingga tidak dilakukan klaim keunggulan terhadap CLIP-EBC pada dataset tersebut. Pada JHU-Crowd++, MAE model 56,21 masih lebih tinggi daripada LIMM 53,0 [13].

Tabel 3. Perbaikan MAE terhadap CLIP-EBC pada benchmark bersama

Dataset	CLIP-EBC	Model diuji	Perbaikan absolut	Perbaikan relatif (%)
SH Part A	52,5	51,39	1,11	2,11
SH Part B	6,6	5,79	0,81	12,27
UCF-QNRF	80,3	75,19	5,11	6,36



Gambar 2. Perbandingan MAE terhadap metode pembanding berbasis bahasa CLIP-EBC pada tiga dataset (skala logaritmik).

3.3. Konfigurasi bobot supervisi yang digunakan

Konfigurasi supervisi berbeda antar-dataset. Data yang tersedia menunjukkan dua pola konfigurasi: UCF-CC-50 menggunakan bobot regularisasi struktural 0,10 untuk kehalusan lokal dan penyelarasan grid, sedangkan pada JHU-Crowd++ komponen kerugian relatif dan penyelarasan grid dinonaktifkan serta digunakan laju pembelajaran paling kecil di antara konfigurasi yang dicatat. Karena tidak tersedia sweep bobot terkontrol dengan metrik variasi untuk setiap nilai, konfigurasi ini diperlakukan sebagai deskripsi setting eksperimen, bukan bukti kausal bahwa satu jenis anotasi secara pasti membutuhkan bobot tertentu.

UCF-CC-50 menggunakan bobot 0,10 untuk kehalusan lokal dan penyelarasan grid pada konfigurasi yang dicatat, sedangkan JHU-Crowd++ menggunakan kerugian relatif = 0 dan penyelarasan grid = 0. Pola tersebut menunjukkan bahwa performa akhir tidak hanya berkaitan dengan prompt, tetapi juga dengan bentuk supervisi spasial yang digunakan. Namun, tanpa ablation bobot yang terkontrol, hubungan antara noise anotasi dan penonaktifan komponen loss tidak ditafsirkan sebagai hubungan sebab-akibat.

Tabel 4. Konfigurasi supervisi yang tersedia pada data eksperimen

Dataset	Kerugian relatif	Kehalusan lokal	Penyelarasan grid	Laju pembelajaran
UCF-CC-50	—	0,10	0,10	—
JHU-Crowd++	0	—	0	Paling kecil pada konfigurasi yang dicatat

3.4. Implikasi dan keterbatasan

Prompt dapat diubah pada tingkat masukan tanpa mengubah bobot model, sehingga secara konseptual memberikan fleksibilitas konfigurasi ketika konteks semantik berubah. Namun, penelitian ini tidak mengukur waktu inferensi, penggunaan memori, konsumsi energi, throughput, atau biaya komputasi. Oleh karena itu, hasil tidak digunakan untuk menyimpulkan efisiensi pada sistem tertanam. Kajian mengenai kompleksitas model dan efisiensi edge [10], [11] ditempatkan sebagai konteks bahwa evaluasi deployment memerlukan metrik komputasi tersendiri.

Keterbatasan penelitian mencakup empat hal. Pertama, ablation kontribusi prompt hanya dilakukan pada ShanghaiTech Part A, sehingga persentase penurunan galat tidak digeneralisasi ke dataset lain. Kedua, hasil ablation dilaporkan sebagai estimasi titik tanpa jumlah pengulangan, simpangan baku antar-run, atau confidence interval, sehingga perbedaan 5,64 titik persentase diperlakukan sebagai gap observasional. Ketiga, detail reproduksibilitas seperti string literal prompt acak, seed, pembagian data setiap run, batch size, epoch, optimizer, dan parameter pelatihan lengkap tidak tersedia pada data eksperimen yang digunakan untuk artikel ini. Keempat, analisis bobot supervisi belum berupa sweep terkontrol. Penelitian lanjutan perlu menguji beberapa formulasi prompt pada setiap dataset, melaporkan variasi antar-run, dan melakukan ablation bobot supervisi dengan konfigurasi lain tetap sama. Penghitungan agnostik kelas [12] dan pendekatan yang menekankan informasi lokal [13] dapat menjadi pembanding untuk pengembangan berikutnya.

3.5. Ringkasan temuan

Hasil dapat diringkas menjadi tiga temuan. Pertama, pada ShanghaiTech Part A, prompt baku menghasilkan galat lebih rendah daripada tanpa prompt, token visual tanpa teks, dan prompt acak; pola ini konsisten dengan peran konteks semantik pada cross-attention. Kedua, pada tiga benchmark yang memiliki pembanding CLIP-EBC, peningkatan bersifat berbeda menurut skala: terbesar secara absolut pada UCF-QNRF dan terbesar secara relatif pada ShanghaiTech Part B. Ketiga, konfigurasi supervisi berbeda antar-dataset, tetapi data yang tersedia belum mendukung atribusi kausal terhadap bobot loss tertentu. Dengan demikian, kontribusi utama penelitian adalah bukti ablation prompt pada ShanghaiTech Part A dan analisis perbandingan benchmark yang lebih terukur, bukan klaim universal bahwa satu prompt atau konfigurasi supervisi optimal untuk seluruh dataset.

4. Kesimpulan

Penelitian ini mengevaluasi pengaruh kondisi prompt pada model penghitungan berbasis CLIP. Pada ShanghaiTech Part A, prompt baku menurunkan MAE sebesar 11,72% dan RMSE sebesar 10,06% terhadap kondisi tanpa prompt; token visual tanpa teks menurunkan MAE 6,08% dan prompt acak 2,37%. Selisih 5,64 titik persentase antara prompt baku dan token visual merupakan gap observasional karena variasi antar-pengulangan tidak tersedia. Pada perbandingan dengan CLIP-EBC, peningkatan hanya dinyatakan pada tiga dataset yang memiliki nilai pembanding: 2,11% relatif pada ShanghaiTech Part A, 12,27% pada Part B, dan 6,36% pada UCF-QNRF. Perbaikan absolut terbesar terjadi pada UCF-QNRF sebesar 5,11 MAE, sedangkan perbaikan relatif terbesar terjadi pada ShanghaiTech Part B. Nilai CLIP-EBC untuk JHU-Crowd++ tidak tersedia pada tabel pembanding, sehingga tidak dibuat klaim keunggulan terhadap metode tersebut pada JHU-Crowd++. Konfigurasi bobot supervisi diperlakukan sebagai setting eksperimen, bukan hubungan kausal, karena belum tersedia ablation bobot terkontrol. Kesimpulan mengenai kontribusi langsung prompt dibatasi pada ShanghaiTech Part A, dan generalisasi lintas dataset memerlukan pengujian prompt berulang dengan konfigurasi serta pelaporan variasi yang setara.

Daftar Rujukan

[1] M. N. Achmadiah, N. Setyawan, A. A. Bryantono, C.-C. Sun, and W.-K. Kuo, "Fast Person Detection Using YOLOX With AI Accelerator For Train Station Safety," in *2024 International Electronics Symposium (IES)*, IEEE, Aug. 2024, pp. 504–509. doi: 10.1109/IES63037.2024.10665874.

[2] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 589–597. doi: 10.1109/CVPR.2016.70.

[3] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 1091–1100. doi: 10.1109/CVPR.2018.00120.

[4] M. N. Achmadiah, C.-C. Sun, W.-K. Kuo, and J.-W. Hsieh, "RepSFNet : A Single Fusion Network with Structural Reparameterization for Crowd Counting," in *2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS)*, IEEE, Aug. 2025, pp. 1–6. doi: 10.1109/AVSS65446.2025.11149882.

[5] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai, "Distribution Matching for Crowd Counting," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.*, 2020.

[6] Alec Radford * 1 Jong Wook Kim * 1 Chris Hallacy 1 Aditya Ramesh 1 Gabriel Goh 1 Sandhini Agarwal 1 Girish Sastry 1 Amanda Askell 1 Pamela Mishkin 1 Jack Clark 1 Gretchen Krueger 1 Ilya Sutskever 1, "Learning Transferable Visual

Models From Natural Language Supervision," in *38 th International Conference on Machine Learning*, 2021.

[7] Y. Ma, V. Sanchez, and T. Guha, "CLIP-EBC: CLIP Can Count Accurately through Enhanced Blockwise Classification," in *2025 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Jun. 2025, pp. 1–6. doi: 10.1109/ICME59968.2025.11209839.

[8] N. Amini-Naieni, T. Han, and A. Zisserman, "CountGD: Multi-Modal Open-World Counting," in *Advances in Neural Information Processing Systems 37*, San Diego, California, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024, pp. 48810–48837. doi: 10.52202/079017-1547.

[9] Yihong wu, jinqiau wei, xionghui zhao, and yidi li, "DSGC-Net: A Dual-Stream Graph Convolutional Network for Crowd Counting via Feature Correlation Mining,"

[10] N. Setyawan, C.-C. Sun, M.-H. Hsu, W.-K. Kuo, and J.-W. Hsieh, "MicroViT: A Vision Transformer with Low Complexity Self Attention for Edge Device," in *2025 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, May 2025, pp. 1–5. doi: 10.1109/ISCAS56072.2025.11043206.

[11] M. Nurul Achmadiah, A. Ahamad, C.-C. Sun, and W.-K. Kuo, "Energy-Efficient Fast Object Detection on Edge Devices for IoT Systems," *IEEE Internet Things J.*, vol. 12, no. 11, pp. 16681–16694, Jun. 2025, doi: 10.1109/JIOT.2025.3536526.

[12] Q. W. ⋆, H. R. and J. L. Xiaofei Hui, "Class-Agnostic Object Counting with Text-to-Image Diffusion Model," in *European Computer Vision Association (ECVA)*,

[13] Tianhang Pan and Xiuyi Jia, "Local Information Matters: A Rethink of Crowd Counting,"