Multimodal Prompt Insertion Strategies for Vision-Language Foundation Models

Sanil Bheatia
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
hellosunil@ucf.edu

Kreaishna Shetty
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
krishnas@uab.edu

Mahesh Mistry
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
hellomahesh@oregonstate.edu

Abstract

Vision-language foundation models have ushered in a new era of multimodal intelligence, enabling systems to generate coherent textual descriptions from images, reason across modalities, and perform complex comprehension tasks. The integration of prompt engineering into these models has emerged as a pivotal mechanism for guiding behaviour without full retraining. Yet the design space for inserting prompts, both textual and visual, across the layered architectures of large-scale vision-language models remains fragmented and underexplored from a systems perspective. This paper presents a comprehensive analysis of multimodal prompt insertion strategies, examining the structural, infrastructural, and socio-technical dimensions that shape their effectiveness, efficiency, and trustworthiness. We develop a taxonomy of insertion architectures ranging from early fusion and cross-modal prefix tuning to dynamic, learnable insertion policies. System-level trade-offs involving computational cost, memory footprint, latency, and modular governance are dissected. The discussion extends to robustness under distribution shift and adversarial interference, fairness across demographic groups embedded in multimodal data, and the policy implications of prompt-based model steering in high-stakes deployments. Through cross-domain comparisons spanning medical imaging, autonomous perception, and content moderation, we reveal how insertion strategies mediate between flexibility and control, between adaptation and foundational knowledge preservation. The paper further addresses sustainability concerns tied to prompt-tuning overhead and the risk of prompt injection vulnerabilities. We argue that advancing prompt insertion requires not only architectural innovation but also holistic infrastructure design, monitoring frameworks, and governance protocols attuned to the evolving landscape of foundation model ecosystems.

Keywords

Multimodal prompt insertion, vision-language models, prompt tuning, foundation models, system architecture, robustness, governance.

1. Introduction

The landscape of artificial intelligence has been reconfigured by the rise of foundation models that jointly process vision and language, enabling capabilities that span image captioning, visual question answering, cross-modal retrieval, and instruction-following across visual contexts [1, 2, 3]. These models, pretrained on massive corpora of paired image-text data, capture rich associations between perceptual features and linguistic concepts. Despite the considerable representational power locked within their pretrained parameters, domain-specific adaptation and task specialisation demand flexible interface mechanisms that avoid full fine-tuning, which is often computationally prohibitive and risks catastrophic forgetting. Prompting has accordingly emerged as a lightweight paradigm for steering model behaviour, originally popularised in the language domain and rapidly extended to multimodal settings through architectures that augment input sequences with learnable vectors, textual instructions, or hybrid tokens [4, 5, 6]. In vision-language systems, the concept of prompt insertion transcends simple textual prefixing, encompassing the strategic placement of adaptive signals at different depths of the encoder-decoder backbone, across modality-specific streams, and within cross-attention modules [7, 8].

The multifaceted nature of multimodal prompt insertion invites a systems-oriented inquiry that goes beyond the accuracy metrics reported in benchmark evaluations. The decisions regarding where, when, and how to insert prompts are not purely model-centric but are entangled with infrastructure constraints, deployment environments, and societal expectations. Insertion strategies mediate critical trade-offs: early visual prompt tuning can provide rich perceptual grounding but demands higher forward-pass memory, while late-stage linguistic prompts may enjoy lower latency at the cost of visual grounding fidelity [9, 10, 11]. Moreover, learnable insertion policies that select layers dynamically can improve parameter efficiency and robustness but introduce additional training overhead and complicate explainability [12, 13]. From a governance standpoint, the ability to modify model behaviour through prompt insertion holds implications for accountability, as the boundary between the fixed foundation model and the inserted signal becomes a locus of regulatory interest.

In this paper, we systematically examine multimodal prompt insertion strategies for vision-language foundation models, adopting a comprehensive, interdisciplinary lens that integrates architectural design, system-level trade-offs, robustness, fairness, security, and policy. We do not propose a new insertion algorithm; rather, we synthesise emerging trends, structural taxonomies, and deployment considerations to provide a conceptual map for researchers and engineers constructing reliable prompt-based systems at scale. The analysis draws on cross-domain case illustrations in medical diagnostic support, autonomous driving, and content moderation to highlight how different insertion choices reverberate through the entire socio-technical stack. By framing prompt insertion as a systems problem, we aim to bridge the gap between empirical tuning efforts and the broader imperatives of sustainable, trustworthy, and governable AI infrastructures.

2. Foundational Concepts and Problem Framing

Vision-language foundation models such as CLIP, BLIP-2, and LLaVA operationalise multimodal interaction through architectures that combine a visual encoder, often a Vision Transformer or convolutional network, with a pretrained language model or a cross-modal encoder [1, 2, 3]. The typical inference pathway converts an image into a sequence of visual embeddings, which are then fused with textual tokens via cross-attention, concatenation, or gated channels. Prompting in this context refers to the deliberate augmentation of the input or hidden representations with signals that influence downstream generation or classification.

While discrete text prompts have been widely explored, the frontier has shifted toward continuous soft prompts that are trained in a parameter-efficient manner, adapting the model to new tasks while keeping the vast majority of pretrained weights frozen [4, 5, 6].

A central framing challenge is that prompt insertion is inherently multimodal and multi-scale. A prompt can be injected as a prefix to the text input, as a learnable visual prefix concatenated to image patches, as gated cross-attention biases within fusion layers, or as residual adapters placed after specific attention blocks [8, 14, 9]. Each insertion point operates at a different representational granularity and influences what aspects of the multimodal signal are modulated. Insertions early in the visual pipeline encourage the model to reinterpret low-level perceptual features, which is advantageous for fine-grained recognition tasks but may overfit to spurious correlations in small datasets [10]. In contrast, prompts inserted near the language decoding head exert influence on lexical choice and syntactic structure with minimal impact on perceptual encoding, offering a form of controlled stylistic or task-oriented steering.

The systems problem is compounded by the frozen nature of the foundation backbone. In typical parameter-efficient tuning scenarios, the vast base model is treated as an immutable infrastructure component that must be served across many tenants and applications. In such multi-tenant deployments, each downstream task may require its own set of inserted prompt parameters, and the infrastructure must efficiently manage, load, and compose these prompts while meeting latency and throughput requirements. This paradigm shifts prompt insertion from a mere training-time design choice to a runtime orchestration challenge. The insertion strategy determines how much additional compute is consumed, how easily prompts can be swapped or updated, and how interference between concurrent prompt configurations is mitigated. Therefore, framing prompt insertion as a system-level design variable is essential for building scalable and maintainable vision-language services.

3. Taxonomies of Prompt Insertion Architectures

A structural taxonomy of multimodal prompt insertion architectures can be constructed along several axes: the modality of insertion, the depth of insertion within the network topology, the granularity of the prompt parameters, and the degree of dynamic selection. At the coarsest level, insertion strategies fall into early fusion, late fusion, and hierarchical insertion categories. Early fusion strategies inject learnable prompt tokens directly into the visual token sequence before it enters the transformer encoder, as exemplified by Visual Prompt Tuning (VPT) [6]. This approach treats prompts as additional patch embeddings, fully participating in self-attention across all layers. It achieves strong task adaptation by reshaping the model's visual representations from the outset, yet it increases the sequence length processed by every subsequent layer, thereby amplifying computational cost multiplicatively. In latency-sensitive deployments, this cumulative overhead can become a bottleneck, especially when multiple visual prompts must be batched together.

Late fusion insertion strategies, by comparison, target the linguistic side of the model, often appending learnable prefix tokens to the text input or injecting adapters after cross-modal fusion layers [5, 11]. Flamingo's gated cross-attention mechanism represents a sophisticated late fusion variant where visual information is fed into the language model through interleaved cross-attention layers, and the gates are themselves inserted as learnable parameters that modulate the influence of visual signals [7]. By restricting prompt insertion to a few layers near the top of the model, late fusion strategies preserve the integrity of the pretrained visual hierarchy and minimise the total number of operations, achieving favourable memory and latency profiles. However, this choice can limit the model's ability to learn

detailed perceptual distinctions that require reshaping low-level visual features, thereby constraining performance on tasks like fine-grained object localisation or texture classification.

Hierarchical insertion strategies bridge these extremes by distributing prompt parameters across multiple depths of the architecture. MaPLe, for instance, couples visual and language prompts through a shared coupling mechanism that enforces cross-modal alignment, inserting learnable vectors at multiple transformer blocks in both the visual and language branches [17]. Such hierarchical designs exploit the fact that vision-language tasks often require coordination between low-level visual detail and high-level semantic reasoning. The design challenge lies in balancing the number of inserted parameters with the risk of overfitting and in ensuring that the coupled insertion does not introduce unwanted interference between modalities. Further complexity arises when insertion policies are made learnable, allowing the model to decide at inference time which layers should receive prompts based on input characteristics. Work on selective prompt tuning has demonstrated that learning to choose insertion positions can improve both efficiency and generalisation by concentrating adaptation capacity where it matters most [18]. In these architectures, a lightweight gating network or policy head is trained jointly with the prompts, and its decisions are conditioned on intermediate representations. The dynamic insertion approach introduces an additional optimisation problem and makes the resulting system more difficult to analyse, as the computational graph becomes input-dependent.

The taxonomy reveals a fundamental tension between the expressive power of dense, pervasive insertion and the operational desiderata of modularity, interpretability, and cost. In multi-task serving systems, managing a library of insertion configurations whose behaviour may vary per input instance necessitates sophisticated orchestration logic. Infrastructure providers must track prompt versions, monitor their resource consumption per query, and implement back-off mechanisms that fall back to simpler insertion strategies under high load. The choice of insertion architecture thus propagates into the design of inference engines, caches, and pricing models, cementing the status of prompt insertion as a systems concern.

4. System-Level Trade-offs and Infrastructure Challenges

Deploying vision-language foundation models with multimodal prompt insertion at scale surfaces a range of system-level trade-offs that extend beyond traditional accuracy metrics. One of the most salient trade-offs involves the interplay between compute cost, memory footprint, and task adaptability. Early visual prompt strategies that increase effective sequence length impose a quadratic expansion of self-attention operations, which can become prohibitive in high-resolution visual inputs or in real-time applications such as autonomous driving perception, where inference must complete within strict latency budgets [12, 14]. Even when the per-query increase appears modest, the aggregated effect across millions of daily queries can substantially elevate energy consumption and cloud expenditure, undermining sustainability objectives. Conversely, late fusion approaches minimise the sequence length penalty but require the foundation model to implicitly carry all visual grounding through the bottleneck of a few cross-attention layers, which may become a performance limiter in complex scenes with multiple interacting objects.

Memory pressure emerges as another critical dimension. In many server environments, a single GPU must serve numerous fine-tuned variants by swapping in prompt parameters while keeping the backbone model resident in memory. Prompt insertion configurations that scatter parameters across many layers, as hierarchical strategies tend to do, increase the total

parameter count per task variant, potentially exhausting GPU memory when many tasks are multiplexed [16, 17]. To counteract this, system designers employ techniques such as prompt distillation, weight sharing across tasks, or gradient-free insertion methods that approximate full prompt effects. Yet each mitigation introduces additional engineering complexity and can degrade the fidelity of the inserted signals. Furthermore, the interaction between prompt insertion and model compression techniques, such as quantisation or pruning, remains poorly understood. Inserted prompts that have been trained in full precision may lose their steering efficacy when the backbone is quantised to int8, raising concerns about deployment consistency across hardware platforms.

Latency variability also emerges from the input-dependent nature of dynamic insertion strategies. If a learnable policy determines which layers receive prompts on a per-example basis, the computational depth becomes non-uniform, complicating batch scheduling and pipeline parallelism. Inference serving systems optimised for static computation graphs may experience load imbalance and tail latency spikes when handling dynamically routed prompts. Addressing these challenges calls for the co-design of insertion policies and serving infrastructure, potentially through profiling-driven insertion compilation that precomputes insertion plans for clusters of similar inputs. Such solutions, however, require extensive telemetry and runtime systems that can gracefully adapt to model updates.

The economic dimensions of prompt insertion are equally significant. In platform business models where multiple clients fine-tune a shared foundation model using their own proprietary prompt data, the infrastructure must guarantee isolation between clients to prevent cross-task interference and model inversion attacks. Prompt isolation can be enforced at the software level, but the shared nature of the backbone raises the spectre of side-channel leakage through the attention mechanism, requiring careful security auditing. As organisations increasingly adopt foundation models as service backbones, the governance of prompt insertion becomes inextricably linked with service-level agreements, data provenance, and compliance with auditing regimes.

5. Robustness, Fairness, and Security Implications

The robustness of vision-language foundation models under prompt insertion is a multifaceted concern. Inserted prompts that are optimised on clean training distributions may fail catastrophically when the model is exposed to corrupted images, adversarial patches, or out-of-distribution visual concepts. Because prompts operate by modulating internal representations, small perturbations in the visual input can interact with the inserted signal to produce large, unpredictable shifts in model output, a phenomenon amplified by the nonlinearities of transformer architectures [13, 15]. Adversarial robustness studies indicate that inserted soft prompts can themselves become attack vectors; an adversary who gains knowledge of the prompt structure can craft inputs that induce targeted misbehaviour, a threat model known as adversarial prompt reprogramming. Defending against such attacks requires robust prompt regularisation, adversarial training of insertion policies, and real-time monitoring of model consistency, all of which impose additional system overhead.

Fairness represents another pressing dimension. Vision-language models have been shown to exhibit biases that correlate with protected attributes such as gender, skin tone, and age, often because the pretraining data embeds societal stereotypes [13, 22]. Prompt insertion can both mitigate and exacerbate these biases depending on how the adaptation data are curated and how the insertion strategy interacts with the frozen backbone. An insertion that emphasises late-stage textual prompts may inadvertently amplify biased linguistic associations while

leaving visual representations unchanged, leading to outputs that are fluent but discriminatory. In contrast, inserting counter-stereotypical prompts deep within the visual encoder may reduce bias in visual recognition but risks distorting foundational knowledge in unintended ways. Fairness-aware prompt insertion is therefore not simply a matter of curating unbiased adaptation data but also requires careful architectural calibration to ensure that inserted signals do not override ethical constraints embedded in the base model or fail to propagate them across modalities.

Security concerns around prompt injection take on new dimensions in multimodal settings. The concatenation of user-supplied images and text with inserted system prompts creates multiple channels through which malicious content can influence model behaviour. Visual prompt injection, where an attacker embeds imperceptible patterns in an image to hijack the inserted prompt's effect, is an evolving threat that has received limited attention compared to textual injection in language models [19]. Because visual signals enter the model at the earliest layers, a compromised image can preempt the influence of any subsequent prompt insertion, effectively neutralising safety filters. System designs must therefore implement defensive insertion strategies that verify the integrity of visual inputs before they interact with prompt parameters, possibly through cryptographic hashing of image fingerprints or through anomaly detection modules that run in parallel. The insertion architecture itself must be hardened, with prompts that are robust to visual noise and that are continuously updated based on threat intelligence, mirroring practices from cybersecurity operations.

6. Governance, Policy, and Sustainability

The increasing reliance on prompt insertion to steer vision-language foundation models raises substantive governance and policy questions. When a healthcare provider uses a prompt-adapted model to assist in radiology report generation, the boundaries of liability become blurred: the foundation model provider may claim that the inserted prompt, developed by the hospital's AI team, is primarily responsible for the model's behaviour, while the hospital may argue that the underlying model's biases and limitations are the root cause of errors [21, 22]. This diffusion of responsibility demands new governance frameworks that explicitly attribute accountability for each component in the prompt-insertion pipeline. Regulatory approaches such as mandatory model cards and prompt documentation can increase transparency, but they must evolve to capture the interplay between frozen base models and inserted adaptations, as well as the lineage of prompt data used during fine-tuning.

The sustainability implications of prompt insertion are twofold. On the one hand, parameter-efficient prompt tuning dramatically reduces the carbon footprint of adaptation compared with full fine-tuning, potentially enabling more frequent model updates without proportional energy costs [14]. On the other hand, the trend toward dynamic, input-dependent insertion with complex routing policies reintroduces computational variability and can lead to higher per-query energy consumption than a simple static insertion. When multiplied across millions of queries, these marginal increases may negate the initial savings. Sustainable prompt insertion therefore requires a holistic view that optimises not just training efficiency but also inference energy, encompassing hardware-aware insertion design, sparsity-inducing regularisation, and the use of carbon-aware schedulers that shift non-latency-critical prompt-adapted workloads to times of abundant renewable energy.

Policy interventions must also consider the risk of prompt monocultures. If the ecosystem converges on a small number of insertion architectures or pretrained backbones, the diversity of adaptation pathways narrows, making the entire system fragile to shared vulnerabilities. A

single adversarial technique that compromises a popular insertion pattern could cascade across numerous applications. Encouraging architectural diversity through open standards for prompt interfaces and interoperable prompt libraries can mitigate such systemic risk. Furthermore, regulatory bodies should foster transparency by requiring that publicly deployed vision-language systems disclose whether and how prompts are inserted, including the granularity and depth of insertion, akin to nutritional labels for AI components. Such disclosures would empower third-party auditors to assess fairness, robustness, and security systematically, building public trust in multimodal AI systems that increasingly mediate information access, education, and creative expression.

7. Conclusion

Multimodal prompt insertion has evolved into a pivotal systems-design problem that bridges algorithmic innovation and real-world deployment exigencies. This paper has traced the structural taxonomies, system-level trade-offs, and broader societal implications of how prompts are placed within vision-language foundation models. The analysis demonstrates that no single insertion strategy universally dominates; instead, the choice of architecture must be guided by the specific operational constraints, task requirements, and risk profiles of the deployment context. Early visual insertion offers perceptual richness but strains computational resources, late linguistic insertion provides efficiency but constrains grounding, and hierarchical or dynamic strategies balance flexibility with increased orchestration complexity. Beyond performance, robustness against adversarial manipulation, fairness across demographic groups, and the complexity of governing hybrid human-AI systems demand that prompt insertion be treated as a socio-technical design variable. Future progress will depend on the co-evolution of insertion algorithms, serving infrastructure, fairness auditing protocols, and policy instruments that together can ensure vision-language foundation models are not only powerful but also trustworthy, sustainable, and accountable. The research community, infrastructure engineers, and policymakers must collaborate to establish shared benchmarks, transparency standards, and threat models that reflect the full stack of concerns from the hardware to the societal layer, thereby enabling a responsible trajectory for the next generation of multimodal AI systems.

References

1. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748–8763). PMLR.

2. Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning (pp. 19730–19742). PMLR.

3. Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023). LLaVA: Large language and vision assistant. arXiv preprint arXiv:2304.08485.

4. Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 3045–3059). Association for Computational Linguistics.

5. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations.

6. Jia, M., Tang, L., Chen, B. C., Cardie, C., Belongie, S., Hariharan, B., & Lim, S. N. (2022). Visual prompt tuning. In European Conference on Computer Vision (pp. 709–727). Springer.

7. Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., … Simonyan, K. (2022). Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35, 23716–23736.

8. Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S. M. A., Vinyals, O., & Hill, F. (2021). Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems, 34, 200–212.

9. Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16816–16825). IEEE.

10. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., & Qiao, Y. (2023). CLIP-Adapter: Better vision-language models with feature adapters. International Journal of Computer Vision, 131(8), 1986–2002.

11. Sung, Y. L., Cho, J., & Bansal, M. (2022). VL-Adapter: Parameter-efficient transfer learning for vision-and-language tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5227–5237). IEEE.

12. Liang, P. P., Zadeh, A., & Morency, L. P. (2022). Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. arXiv preprint arXiv:2209.03430.

13. Agarwal, S., Rekabsaz, N., & Vlachos, A. (2022). Measuring and reducing gendered correlations in pre-trained vision-language models. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 3859–3869). Association for Computational Linguistics.

14. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., Texeira, A., Urbach, M., Bhowmick, A., Li, F., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

15. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In International Conference on Machine Learning (pp. 2790–2799). PMLR.

16. Zeng, A., Attarian, M., Ichter, B., Choromanski, K., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M. S., Sindelar, V., Lee, J., Vanhoucke, V., Florence, P., & Hausman, K. (2022). Socratic models: Composing zero-shot multimodal reasoning with language. arXiv preprint arXiv:2204.00598.

17. Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., & Khan, F. S. (2023). MaPLe: Multi-modal prompt learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19113–19122). IEEE.

18. Zhu, W., & Tan, M. (2023, December). SPT: Learning to selectively insert prompts for better prompt tuning. In Proceedings of the 2023 conference on empirical methods in natural language processing (pp. 11862-11878).

19. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). More than you've asked for: A comprehensive analysis of novel prompt injection threats to application-integrated large language models. In 32nd USENIX Security Symposium (pp. 3675–3692). USENIX Association.

20. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

21. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220–229). ACM.

22. Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: misogyny, objectification, and exploitation in visual language models. In ACM Conference on Fairness, Accountability, and Transparency (pp. 672–686). ACM.