RESEARCH AS AN IMITATION GAME: WHEN GOOD RESEARCHERS COPY AND GREAT RESEARCHERS STEAL

Hanhui Wang
Northeastern University
wang.hanh@northeastern.edu

THE MYTH OF INNOVATION

There has long been a pervasive myth in computer science research, particularly among early-stage doctoral students (such as myself), that "novelty" implies creation *ex nihilo* (out of nothing). Under this framework, a valid contribution must be a brand-new mechanism that no one has seen before. However, this view might fundamentally misunderstand the nature of breakthrough innovation.

In the rapidly evolving research fields, especially Large Language Models and Video Generation, we face the paradox that the components are often decades old (neural networks, attention, *etc*.), yet the resulting systems represent a genuine leap in capacity. So it seems reasonable to argue that high-impact research doesn't only involve *building new bricks*, but is also about recognizing that *a cathedral can be built using the same bricks once used for a warehouse.* In this context, the distinction between "copying" and "stealing" is not ethical, but philosophical. Good researchers copy by optimizing within a given domain, while great researchers steal by identifying structural equivalence between disparate domains and transplanting solutions across the boundaries of time and fields.

FROM SURFACE MIMICRY TO DEEP ABSTRACTION

To understand why "stealing" is a legitimate and advanced research strategy, we must define it strictly against "copying".

Copying is a surface-level form of adaptation. It means taking a solution from problem A and applying it to problem B with minimal reflection (or even within the same domain) and mostly without questioning whether the solution is even appropriate. In other words, copying preserves the form of an idea while ignoring the context that originally made it work. For example, a surprisingly common practice is to copy a fancy cross-attention module Vaswani et al. (2017) from the latest state-of-the-art paper and drop it into a new model "just to see if the curve goes up", without any clear hypothesis for why it should help in the first place. Overall, **copying starts with the old solution and tries to find a new problem to apply it**.

Stealing, on the other hand, **begins with the new problem, not the old solution**. A researcher must strip their current challenge of its domain-specific semantics to reveal its abstract structure. This may allow them to recognize a structural similarity with a solved problem in a different field. And building upon this similarity between abstractions of the problems, they are able to adapt "old" ideas to "new" settings while understanding why they should work. The core innovation, therefore, lies not in the act of copying the solution, but in the insight of understanding why an established solution worked in its original context and verifying whether the underlying logic remains valid when transplanted into a new environment with different constraints.

A CASE STUDY OF VLLM

We can see this dynamic clearly in the development of **vLLM** and **PagedAttention** Kwon et al. (2023). When LLMs moved from research prototypes to production serving, a primary bottleneck shifted from compute to memory, specifically due to the fragmentation of the Key-Value (KV) cache.

A "copying" mindset might have tried to prune the cache or compress the model. However, the researchers behind vLLM adopted a "stealing" mindset, and recognized the non-continuous nature of token generation is similar to the non-continuous nature of process memory allocation in operating systems. By abstracting the KV cache blocks as virtual memory pages, and tokens as bytes, they effectively "stole" the paging mechanism from 1960s OS textbooks.

This was not merely copying and rewriting the code; it was the insight that the serving bottleneck is fundamentally a GPU memory management problem. The innovation wasn't the paging algorithm itself, but the audacity to apply it to an attention mechanism, thereby solving a modern GPU bottleneck with an "ancient" CPU solution.

A Case Study of Video Generation

A similar, perhaps even more profound "imitation" is occurring in my own field of Computer Vision. The true "theft" here is not about copying code or specific architectures; it is about adopting a belief system.

Historically, the primary goal of video generation was visual fidelity—creating clips that looked realistic and temporally smooth. The focus was largely on the signal itself: texture, flow, and resolution. However, a shift in perspective is signaled by recent works such as "**Video Models are Zero-shot Learners and Reasoners**." Wiedemer et al. (2025) The title itself is a deliberate echo of the seminal GPT-3 paper, "**Language Models are Few-Shot Learners**." Brown et al. (2020) By explicitly mirroring this naming convention, researchers are not merely paying homage; they are signaling a deeper theft. They are "stealing" the core intuition that defined the LLM breakthrough: the relationship between generation and reasoning. In the LLM era, we learned that training a model to predict the next text segment, if scaled sufficiently Kaplan et al. (2020), could lead to the emergence of reasoning capabilities. CV researchers are now testing this same hypothesis in the visual domain. The "imitation" lies in the belief that accurate generation serves as a forcing function for understanding and reasoning. We are moving beyond merely trying to make pretty images and videos to exploring whether large-scale video generation models can implicitly learn physics and causality as a byproduct of learning to predict visual dynamics. This represents a transition from treating video as a graphics task to treating it as a potential path toward world models.

Embracing the Imitation Game

Ultimately, we return to the famous observation of Pablo Picasso: "Good artists copy; great artists steal." In the context of computer science research, this is not a confession of unoriginality, but a hierarchy of innovation. The **Imitation Game** we play is not about avoiding the hard work of invention or reaching for low-hanging fruit. Instead, it is about the evolution from a *good* researcher into a *great* one. The lesson is clear: to achieve this kind of greatness, one must possess the vision to look sideways to other disciplines and backward into history. This does not imply that "stealing" is the *best* and *only* path to novelty. It *complements*, rather than *replaces*, the pursuit of pure invention. It echoes that, in an era where AI systems are becoming increasingly complex and convergent, the ability to bridge isolated islands of knowledge is becoming a critical differentiator. The next breakthrough may not come from creating a new mechanism from scratch, but from the spark that occurs when we realize that the key to a locked door in the future might have been forged long ago in the past.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.