ChatGPT

Tổng hợp các bài báo nổi bật (2021–2025) giao thoa Computer Vision và Large Language Models

1. Vision-Language Pretraining (VLP)

Tên bài báo	Tác giả chính	Hội nghị/ Năm	Đóng góp chính	Đường dẫn
Learning Transferable Visual Models From Natural Language Supervision	A. Radford **et al.**	ArXiv (2021)	Tiền huấn luyện theo phương pháp đối chiếu trên 400 triệu cặp ảnh-chữ, học được biểu diễn hình ảnh mạnh mẽ. Mô hình CLIP cho phép sử dụng ngôn ngữ tự nhiên để tham chiếu khái niệm thị giác, giúp chuyển giao zero-shot sang nhiều tác vụ thị giác [1].	ArXiv
Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision	C. Jia **et al.**	ICML (2021)	Mô hình ALIGN sử dụng 1 tỷ cặp ảnh-alt-text, huấn luyện đối chiếu (contrastive). Giúp biểu diễn hình ảnh và ngôn ngữ tách biệt, đạt kết quả cao cho phân loại zero-shot và thiết lập SOTA cho truy xuất ảnh-chữ (COCO, Flickr30K) [2].	ArXiv
ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision	W. Kim **et al.**	ICML (2021)	Đề xuất ViLT, mô hình VLP tối giản không dùng CNN hay supervision khu vực. ViLT xử lý ảnh dưới dạng patch giống văn bản, nhanh hơn gấp nhiều lần và vẫn đạt hiệu năng cạnh tranh trên các nhiệm vụ thị giác-ngôn ngữ [3].	ArXiv
Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation (ALBEF)	J. Li **et al.**	NeurIPS (2021)	ALBEF sử dụng loss đối chiếu để đồng bộ hóa biểu diễn ảnh-chữ trước khi kết hợp. Kết hợp momentum distillation và làm sạch dữ liệu, đạt SOTA trên nhiều task VL: ví dụ VQA tăng +2.37%, NLVR2 +3.84%, và truy xuất ảnh-chữ vượt các mô hình lớn hơn nhiều [4].	ArXiv

Tên bài báo	Tác giả chính	Hội nghị/ Năm	Đóng góp chính	Đường dẫn
SimVLM: Simple Visual Language Model Pretraining with Weak Supervision	Z. Wang **et al.**	ICLR (2022)	SimVLM chỉ dùng một hàm mục tiêu LM prefix và dữ liệu giám sát lỏng lẻo quy mô lớn. Mô hình đạt SOTA trên nhiều task V+L: VQA tăng +3.74%, NLVR2 +1.17%, SNLI-VE +1.37%, và captioning tăng +10.1% (CIDEr) so với tiền nhiệm [5].	ArXiv
CoCa: Contrastive Captioners are Image-Text Foundation Models	J. Yu **et al.**	NeurIPS (2022)	CoCa kết hợp mô hình encoder-decoder với hai loss: đối chiếu và chú thích. Kết quả zero-shot SOTA trên nhiều nhiệm vụ: ví dụ 86.3% top-1 trên ImageNet (zero-shot), 91.0% khi fine-tune, cùng SOTA với retrieval, VQA và captioning [6].	ArXiv
FLAVA: A Foundational Language And Vision Alignment Model	A. Singh **et al.**	CVPR (2022)	FLAVA là mô hình nền tảng cho cả thị giác và ngôn ngữ: thực thi tốt trên cả các tác vụ chỉ vision, chỉ language và cross-modal. Mô hình này đạt hiệu quả ấn tượng trên 35 tác vụ khác nhau thuộc 3 loại (vision, language, multi-modal) [7].	ArXiv
An Empirical Study of Training End-to-End Vision-and-Language Transformers (METER)	Z. Dou **et al.**	CVPR (2022)	Mô hình METER end-to-end kết hợp encoder hình ảnh và ngôn ngữ. Đạt 77.64% VQA-v2 (test-std) với 4M ảnh (tăng 1.04% so với SOTA trước) và 80.54% khi mở rộng mô hình; cải thiện đáng kể hiệu năng trên VQA, NLVR2, và các nhiệm vụ ngôn ngữ khác [8].	ArXiv
BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation	J. Li **et al.**	ArXiv (2022)	BLIP sử dụng mô hình tạo caption và lọc dữ liệu để tận dụng tốt nguồn dữ liệu web nhiều nhiễu. Đạt SOTA trên nhiều tác vụ: ví dụ +2.7% recall@1 (retrieval), +2.8% CIDEr (captioning), +1.6% VQA score so với trước [9], và tổng quát tốt sang các nhiệm vụ video đa phương thức.	ArXiv

Tên bài báo	Tác giả chính	Hội nghị/ Năm	Đóng góp chính	Đường dẫn
BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models	J. Li **et al.**	ArXiv (2023)	BLIP-2 đề xuất huấn luyện hiệu quả bằng encoder ảnh đóng băng và LLM đóng băng, sử dụng bộ biến đổi truy vấn nhỏ. Kết quả SOTA: ví dụ hơn Flamingo-80B 8.7% trên VQA-zero-shot, với số param huấn luyện ít hơn gấp 54 lần; thể hiện khả năng tạo văn bản zero-shot theo chỉ dẫn.	ArXiv
OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework	P. Wang **et al.**	ICML (2022)	OFA là framework Seq2Seq tích hợp đa nhiệm và đa phương thức (image gen, caption, detection, v.v.). Tiền huấn luyện trên 20M cặp ảnh-chữ, đạt SOTA trên nhiều task cross-modal dù dữ liệu nhỏ; thể hiện khả năng chuyển giao sang các nhiệm vụ đa phương thức khác [10].	ArXiv
Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks (BEiT-3)	W. Wang **et al.**	ArXiv (2022)	BEiT-3 sử dụng Multiway Transformer với tiền huấn luyện thống nhất: masked modeling cho ảnh ("Imglish"), văn bản, và cặp ảnh-chữ. Đạt SOTA trên nhiều nhiệm vụ: phát hiện, phân đoạn, phân loại, reasoning (NLVR2), VQA, caption và cross-modal retrieval [11].	ArXiv

2. Image/Video Captioning và Text-to-Image Generation

Tên bài báo	Tác giả chính	Hội nghị/ Năm	Đóng góp chính	Đường dẫn
ClipCap: CLIP Prefix for Image Captioning	R. Mokady **et al.**	ArXiv (2021)	Đề xuất ClipCap: sử dụng biểu diễn ảnh từ CLIP làm tiền tố cho GPT-2 tạo caption. Mô hình giữ CLIP và GPT-2 đóng băng, chỉ huấn luyện lớp mapping. Vẫn đạt kết quả tương đương SOTA trên Conceptual Captions và NoCaps, với mô hình gọn nhẹ, huấn luyện nhanh [12].	ArXiv

Tên bài báo	Tác giả chính	Hội nghị/ Năm	Đóng góp chính	Đường dẫn
GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models	A. Nichol **et al.**	ICML (2022)	Đề xuất GLIDE: mô hình diffusion có hướng dẫn ngôn ngữ. So sánh hai kỹ thuật hướng dẫn: sử dụng CLIP hoặc classifier-free guidance. Phát hiện rằng classifier-free guidance cho hình ảnh chân thực hơn và có tính đồng bộ với văn bản cao hơn. Kết quả GLIDE được đánh giá cao hơn DALL·E (phát sinh cỡ 300M) trong thử nghiệm người dùng [13].	ArXiv
Hierarchical Text-Conditional Image Generation with CLIP Latents (DALL·E 2)	A. Ramesh **et al.**	NeurIPS (2022)	Mô hình hai giai đoạn: bộ phát (prior) sinh embedding ảnh CLIP từ văn bản, sau đó diffusion decoder sinh ảnh từ embedding đó. Cải thiện sự đa dạng mẫu đồng thời giữ ngữ nghĩa chặt chẽ. Đây chính là phương pháp của DALL·E 2, thể hiện tạo ảnh độ phân giải cao từ văn bản với chất lượng top đầu [14].	ArXiv
Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (Imagen)	C. Saharia **et al.**	NeurIPS (2022)	Imagen kết hợp LLM (T5) để mã hóa văn bản cho diffusion hình ảnh. Nghiên cứu cho thấy tăng cỡ LLM cải thiện chất lượng hơn là tăng mạng tạo ảnh. Imagen đạt FID 7.27 trên COCO mà không dùng COCO trong training, ảnh photorealistic, người dùng đánh giá ảnh của Imagen trung thực và đúng nghĩa hơn các model khác [15].	ArXiv
High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion)	R. Rombach **et al.**	CVPR (2022)	Đề xuất Latent Diffusion (LDM): áp dụng diffusion trong không gian mã hóa (latent) thay vì pixel, sử dụng autoencoder mạnh. Thêm cross-attention để nhận đầu vào văn bản. Kỹ thuật này giảm chi phí tính toán đáng kể mà vẫn giữ chất lượng cao, cho phép tạo ảnh phân giải cao. Ứng dụng nổi bật là mô hình Stable Diffusion đạt SOTA về inpainting và nhiều task khác [16].	ArXiv
Scaling Autoregressive Models for Content-Rich Text-to-Image Generation (Parti)	J. Yu **et al.**	ArXiv (2022)	Parti: mô hình seq2seq autoregressive dùng tokenizer ViT-VQGAN và Transformer lên đến 20 tỷ tham số. Đạt FID 7.23 (zero-shot) và 3.22 (fine-tuned) trên COCO, dẫn đầu SOTA. Hỗ trợ tạo ảnh nội dung phức tạp và phát triển benchmark PartiPrompts ~1600 prompt đa dạng [17].	ArXiv

Tên bài báo	Tác giả chính	Hội nghị/ Năm	Đóng góp chính	Đường dẫn
CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers	M. Ding **et al.**	ArXiv (2022)	CogView2 cải thiện T2I của CogView bằng transformer phân cấp và song song hóa quá trình tạo ảnh. Huấn luyện model 6B trên nhiều cặp ảnh-chữ, sau đó fine-tune cho siêu phân giải. Kết quả tạo ảnh so sánh được với DALL·E 2, đồng thời hỗ trợ chỉnh sửa ảnh theo chỉ dẫn (text-guided editing) một cách tự nhiên [18] .	ArXiv

3. Visual Question Answering (VQA)

Tên bài báo	Tác giả chính	Hội nghị/ Năm	Đóng góp chính	Đường dẫn
Flamingo: a Visual Language Model for Few-Shot Learning	J.-B. Alayrac **et al.**	NeurIPS (2022)	Giới thiệu Flamingo: một VLM kết hợp pretrained vision và language, có khả năng few-shot học trực tiếp trên các task mở như VQA và captioning. Flamingo đạt SOTA trên nhiều benchmark VQA/VL với chỉ vài ví dụ, vượt các model fine-tuned nhiều dữ liệu hơn [19] .	ArXiv
Visual Instruction Tuning (LLaVA)	H. Liu **et al.**	NeurIPS (2023)	LLaVA fine-tune tích hợp encoder ảnh với LLM (Vicuna/GPT-4) bằng bộ dữ liệu instruction do GPT-4 tạo. Mô hình đạt khả năng hội thoại đa phương thức ấn tượng: được 85.1% điểm của GPT-4 trên bộ dữ liệu nhân tạo và 92.53% trên ScienceQA khi fine-tune [20] .	ArXiv
MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models	D. Zhu **et al.**	ArXiv (2023)	MiniGPT-4 nối encoder ảnh (CLIP) với LLM nhỏ (Vicuna) qua một lớp chiếu. Mô hình này thể hiện nhiều khả năng giống GPT-4: mô tả chi tiết ảnh, tạo website từ bản vẽ tay, viết truyện/bài thơ dựa trên ảnh, và các nhiệm vụ sáng tạo khác. Tăng cường bằng tập chú thích ảnh chi tiết để giảm lỗi sinh ngôn ngữ [21] .	ArXiv

4. Vision Grounding và Referring Expression Comprehension

Tên bài báo	Tác giả chính	Hội nghị/ Năm	Đóng góp chính	Đường dẫn
MDETR: Modulated Detection for End-to-End Multi-Modal Understanding	A. Kamath **et al.**	ICCV (2021)	MDETR tích hợp trực tiếp ảnh và ngôn ngữ bằng DETR. Sau tiền huấn luyện trên 200k ảnh-câu có chú thích box, MDETR đạt SOTA trên Phrase Grounding (Flickr30k) và Referring Expression Comprehension (RefCOCO/+/g) [22]. Mô hình cũng cho khả năng tốt cho VQA và các bài toán suy luận đơn giản.	ArXiv
Grounded Language-Image Pre-training (GLIP)	H. Li **et al.**	CVPR (2022)	GLIP kết hợp pretraining cho detection và phrase grounding. Học trên 27M ảnh-chữ (kết hợp giám sát yếu), cải thiện mạnh cho object detection và grounding. Zero-shot GLIP đạt 49.8 AP trên COCO (chưa thấy COCO) và fine-tune 60.8 AP (val) vượt SOTA trước đó [23].	ArXiv
GLIPv2: Unifying Localization and Vision-Language Understanding	H. Zhang **et al.**	NeurIPS (2022)	GLIPv2 là mô hình một cho cả phát hiện và hiểu VL. Kết hợp phrase grounding (như detection), region-word contrastive, và MLM. Một model duy nhất đạt hiệu năng gần SOTA cho detection và các task VQA/caption trong VLP, đồng thời thể hiện khả năng zero-shot object detection và grounding vượt trội [24].	ArXiv

5. Multimodal Retrieval

Tên bài báo	Tác giả chính	Hội nghị/ Năm	Đóng góp chính	Đường dẫn
Learning Transferable Visual Models From Natural Language Supervision (CLIP)	A. Radford **et al.**	ICLR (2021)	CLIP thiết lập phương pháp đối chiếu lớn cho học biểu diễn ảnh và văn bản từ 400M cặp. Kết quả mạnh cho retrieval và zero-shot. CLIP trở thành nền tảng cho nhiều hệ thống tìm kiếm ảnh-chữ sau này [1].	ArXiv

Tên bài báo	Tác giả chính	Hội nghị/ Năm	Đóng góp chính	Đường dẫn
Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision (ALIGN)	C. Jia **et al.**	ICML (2021)	ALIGN tiền huấn luyện trên 1 tỷ cặp ảnh-alt-text. Đạt SOTA trên benchmark retrieval (MS-COCO, Flickr30K) và rất tốt cho image-text retrieval zero-shot [2].	ArXiv
FILIP: Fine-grained Interactive Language-Image Pre-Training	L. Yao **et al.**	ICLR (2022)	FILIP sử dụng tương tác muộn giữa patch ảnh và token từ để đối chiếu chi tiết. Tiền huấn luyện trên 300M cặp, đạt SOTA trên nhiều tác vụ, bao gồm retrieval và zero-shot classification, nhờ khả năng dò tìm sự tương đồng cục bộ giữa từ và vùng ảnh [25].	ArXiv
CoCa: Contrastive Captioners are Image-Text Foundation Models	J. Yu **et al.**	NeurIPS (2022)	CoCa kết hợp mô hình chú thích và đối chiếu, tạo embedding tốt cho cả retrieval và nhận dạng. Đạt hiệu suất rất cao trên retrieval ảnh-chữ; đạt 86.3% zero-shot ImageNet, 91.0% khi fine-tune, cùng nhiều SOTA khác trên VQA và caption [6].	ArXiv
BLIP: Bootstrapping Language-Image Pre-training for VL Understanding	J. Li **et al.**	ArXiv (2022)	BLIP dùng captioner để làm sạch dữ liệu huấn luyện. Model đạt recall@1 tăng ~2.7% cho retrieval và CIDEr tăng ~2.8% cho captioning, cho thấy khả năng khớp ảnh-ngôn ngữ chặt chẽ, cải thiện độ chính xác tìm kiếm ảnh theo văn bản [9].	ArXiv

6. Embodied AI & Robotics Perception

Tên bài báo	Tác giả chính	Hội nghị/ Năm	Đóng góp chính	Đường dẫn
PaLM-E: An Embodied Multimodal Language Model	D. Driess **et al.**	ArXiv (2023)	PaLM-E tích hợp đầu vào cảm biến (thị giác, trạng thái robot) vào mô hình PaLM lớn. Mô hình đa phương thức này giải quyết các nhiệm vụ lập kế hoạch và kiểm soát robot tuần tự lẫn trả lời hình ảnh (VQA), đồng thời đạt SOTA trên OK-VQA và nhiều nhiệm vụ robot khác, nhờ đào tạo chung trên dữ liệu hình ảnh-ngôn ngữ phong phú [26].	ArXiv

7. Multimodal Reasoning và Tool Use

Tên bài báo	Tác giả chính	Hội nghị/ Năm	Đóng góp chính	Đường dẫn
Multimodal Chain-of-Thought Reasoning in Language Models	Z. Zhang **et al.**	TMLR (2024)	Đề xuất MCoT: mở rộng Chain-of-Thought để kết hợp cả ngôn ngữ và hình ảnh. Mô hình tách thành hai bước (sinh luận cứ rồi sinh đáp án) giúp cải thiện khả năng suy luận. MCoT đạt SOTA trên ScienceQA và A-OKVQA với mô hình nhỏ (<1B), giảm hiện tượng ảo giác và tăng tốc độ hội tụ [27].	ArXiv

Ghi chú: Các đóng góp chính được trích dẫn từ nguồn gốc của bài báo [1] [2] [7] [4] [8] [5] [9] [28] [11] [6] [14] [13] [15] [16] [17] [18] [19] [20] [21] [29] [22] [24] [25] [26] [27]. Các liên kết kèm theo dẫn tới bản arXiv hoặc trang chính thức của hội nghị.

---

1. **[2103.00020] Learning Transferable Visual Models From Natural Language Supervision**
   https://arxiv.org/abs/2103.00020
2. **[2102.05918] Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision**
   https://arxiv.org/abs/2102.05918
3. **[2102.03334] ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision**
   https://arxiv.org/abs/2102.03334
4. **[2107.07651] Align before Fuse: Vision and Language Representation Learning with Momentum Distillation**
   https://arxiv.org/abs/2107.07651
5. **[2108.10904] SimVLM: Simple Visual Language Model Pretraining with Weak Supervision**
   https://arxiv.org/abs/2108.10904
6. **[2205.01917] CoCa: Contrastive Captioners are Image-Text Foundation Models**
   https://arxiv.org/abs/2205.01917
7. **[2112.04482] FLAVA: A Foundational Language And Vision Alignment Model**
   https://arxiv.org/abs/2112.04482
8. **[2111.02387] An Empirical Study of Training End-to-End Vision-and-Language Transformers**
   https://arxiv.org/abs/2111.02387
9. **[2201.12086] BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation**
   https://arxiv.org/abs/2201.12086
10. **[2202.03052] OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework**
    https://arxiv.org/abs/2202.03052

11 **[2208.10442] Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks**
https://arxiv.org/abs/2208.10442

12 **[2111.09734] ClipCap: CLIP Prefix for Image Captioning**
https://arxiv.org/abs/2111.09734

13 **[2112.10741] GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models**
https://arxiv.org/abs/2112.10741

14 **[2204.06125] Hierarchical Text-Conditional Image Generation with CLIP Latents**
https://arxiv.org/abs/2204.06125

15 **[2205.11487] Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding**
https://arxiv.org/abs/2205.11487

16 **[2112.10752] High-Resolution Image Synthesis with Latent Diffusion Models**
https://arxiv.org/abs/2112.10752

17 **[2206.10789] Scaling Autoregressive Models for Content-Rich Text-to-Image Generation**
https://arxiv.org/abs/2206.10789

18 **[2204.14217] CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers**
https://arxiv.org/abs/2204.14217

19 **[2204.14198] Flamingo: a Visual Language Model for Few-Shot Learning**
https://arxiv.org/abs/2204.14198

20 **Visual Instruction Tuning**
https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html

21 **[2304.10592] MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models**
https://arxiv.org/abs/2304.10592

22 **MDETR - Modulated Detection for End-to-End Multi-Modal Understanding**
https://openaccess.thecvf.com/content/ICCV2021/papers/Kamath_MDETR_-_Modulated_Detection_for_End-to-End_Multi-Modal_Understanding_ICCV_2021_paper.pdf

23 29 **[2112.03857] Grounded Language-Image Pre-training**
https://arxiv.org/abs/2112.03857

24 **[2206.05836] GLIPv2: Unifying Localization and Vision-Language Understanding**
https://arxiv.org/abs/2206.05836

25 **FILIP: Fine-grained Interactive Language-Image Pre-Training | OpenReview**
https://openreview.net/forum?id=cpDhcsEDC2

26 **[2303.03378] PaLM-E: An Embodied Multimodal Language Model**
https://arxiv.org/abs/2303.03378

27 **[2302.00923] Multimodal Chain-of-Thought Reasoning in Language Models**
https://arxiv.org/abs/2302.00923

28 **[2301.12597] BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models**
https://arxiv.org/abs/2301.12597