



南京航空航天大学
Nanjing University of Aeronautics and Astronautics

Enhancing Visual Document Understanding with Contrastive Learning in Large Visual-Language Models

Xin Li^{†*} Yunfei Wu^{*} Xinghua Jiang Zhihao Guo Mingming Gong Haoyu Cao
Yinsong Liu Deqiang Jiang Xing Sun
Tencent YouTu Lab

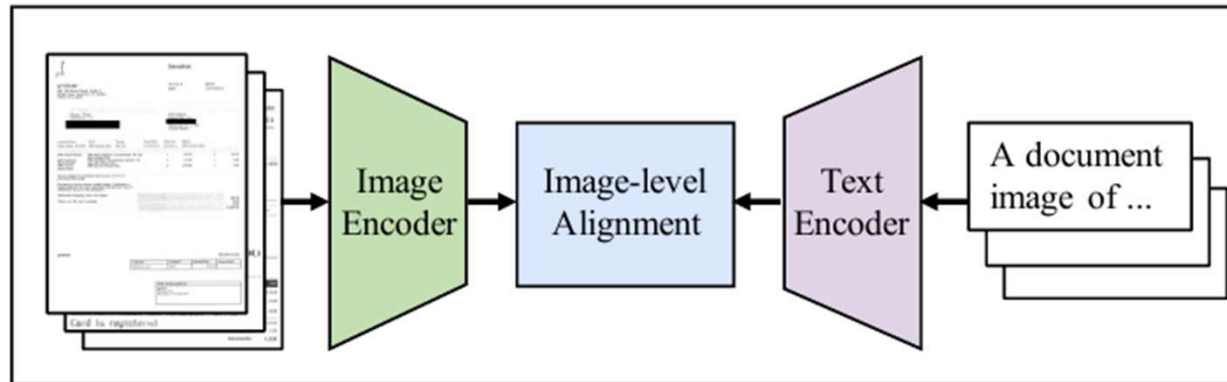
{fujikoli, marcowu, clarkjiang, nicholasguo, riemanngong, rechyciao,
jasonysliu, dqiangjiang, winfredsun}@tencent.com

CVPR2024

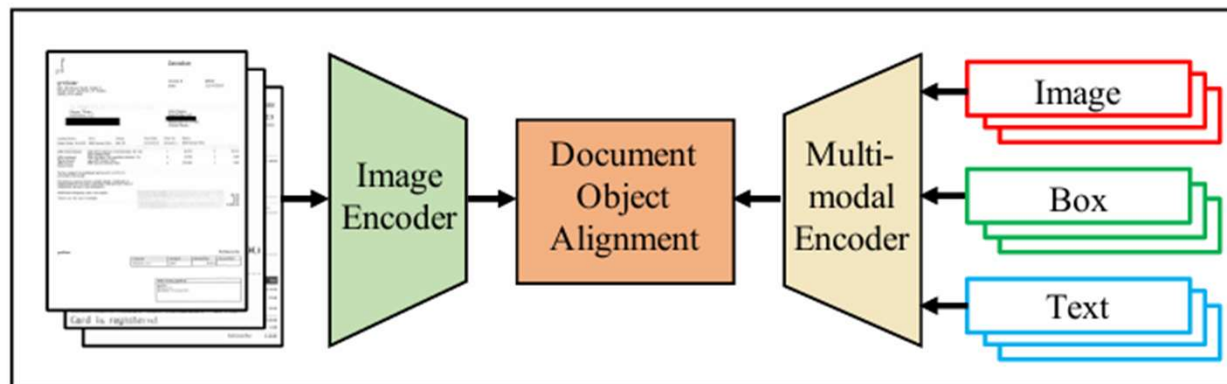
Background



南京航空航天大学
Nanjing University of Aeronautics and Astronautics



(a) Image-level instance discrimination (CLIP)



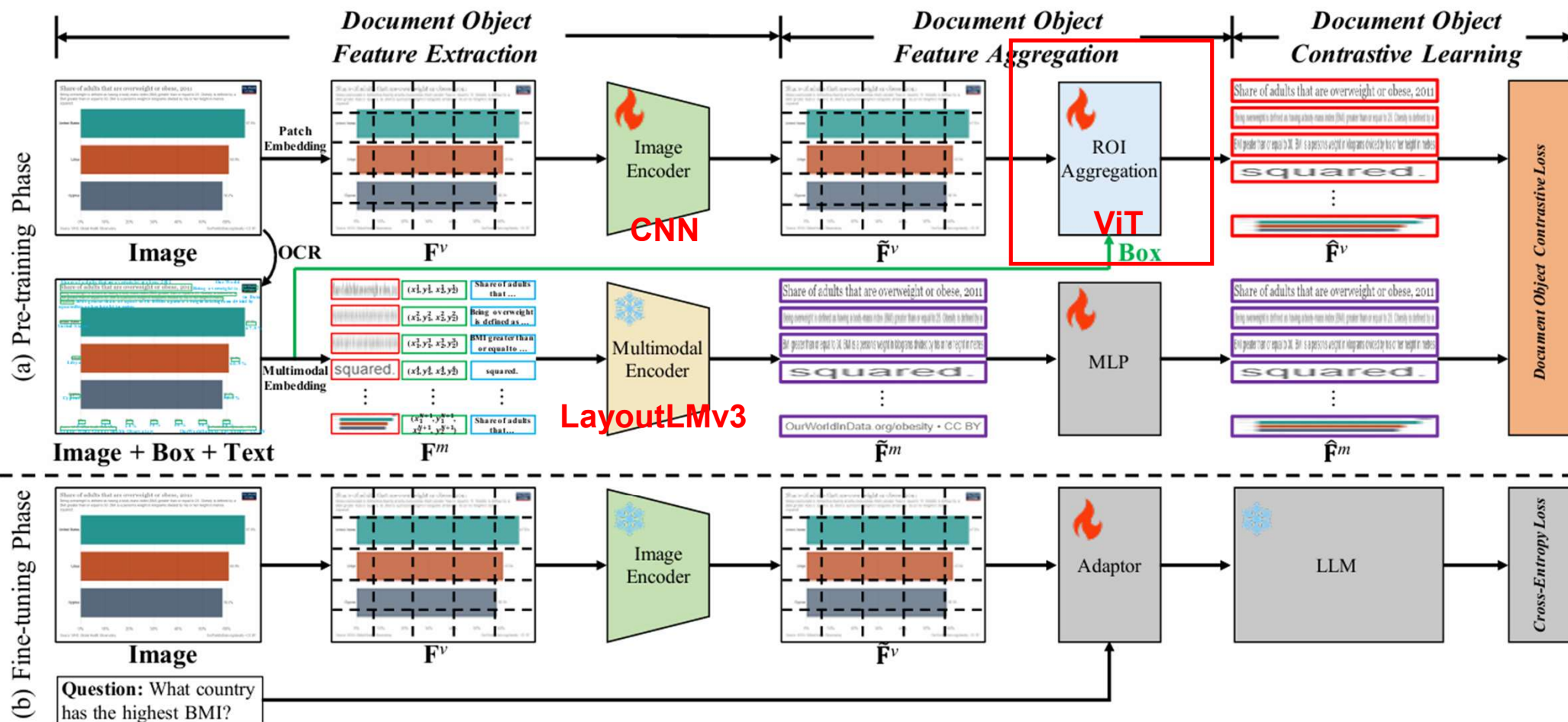
(b) Document object discrimination (Our DoCo)

Schematic Overview



南京航空航天大学

Nanjing University of Aeronautics and Astronautics

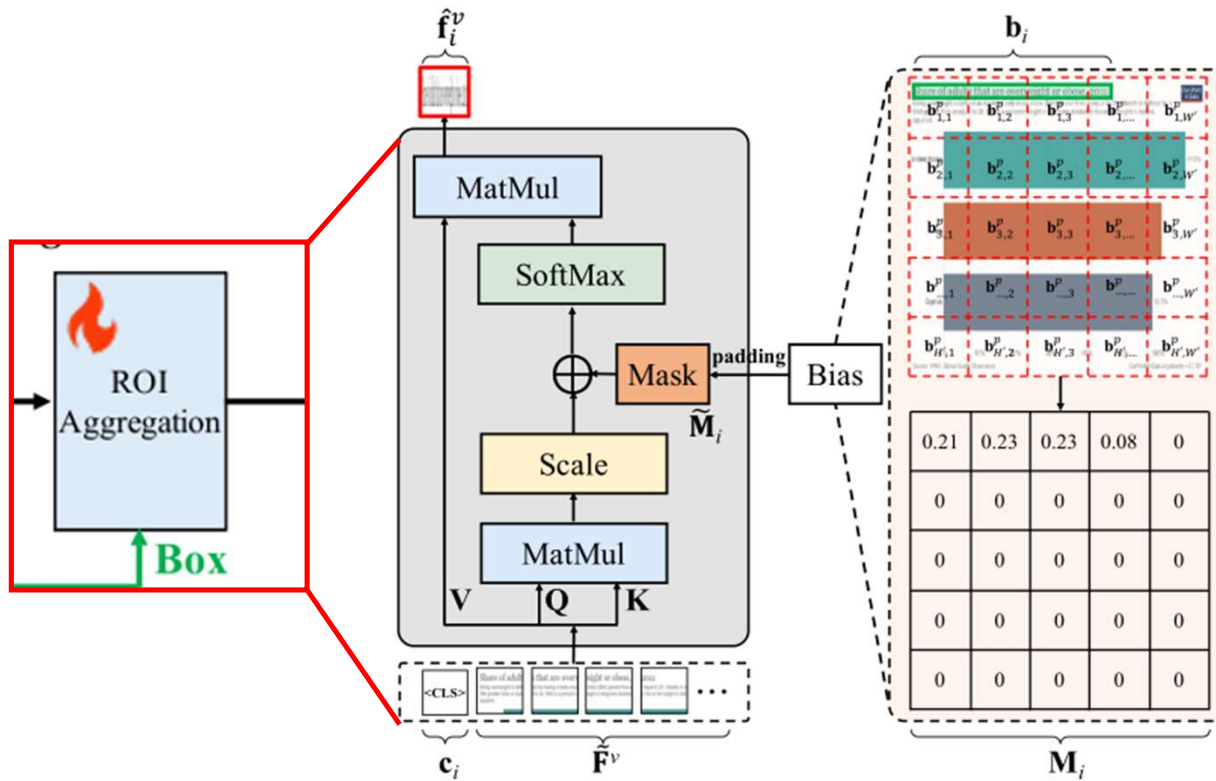


ROI Aggregation



南京航空航天大学

Nanjing University of Aeronautics and Astronautics



$$\mathbf{M}_i = \text{Overlap}(\mathbf{b}_i, \mathbf{b}^p) / \text{Area}(\mathbf{b}^p),$$

$$\widetilde{\mathbf{M}}_i \in \mathbb{R}^{(H'W'+1) \times (H'W'+1)}$$

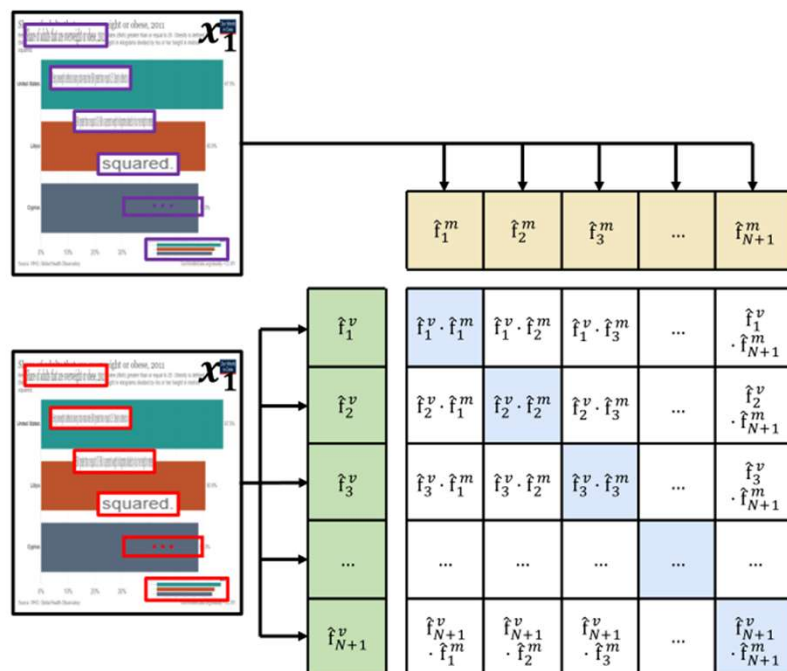
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^\top}{\sqrt{d_v}} + \widetilde{\mathbf{M}}_i\right) \cdot \mathbf{V},$$

Illustrative View of DoCo

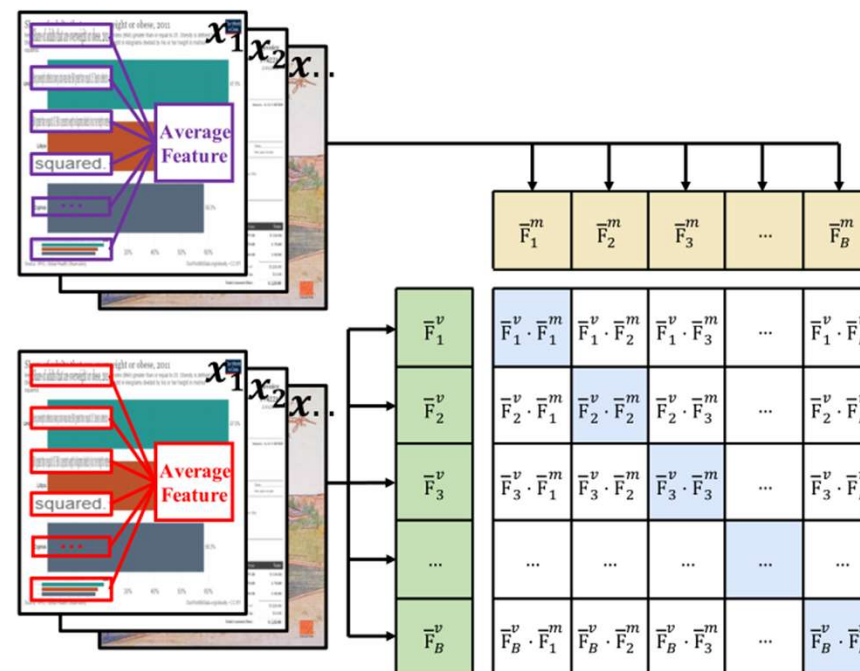


南京航空航天大學

Nanjing University of Aeronautics and Astronautics



(a) Intra-DoCo



(b) Inter-DoCo

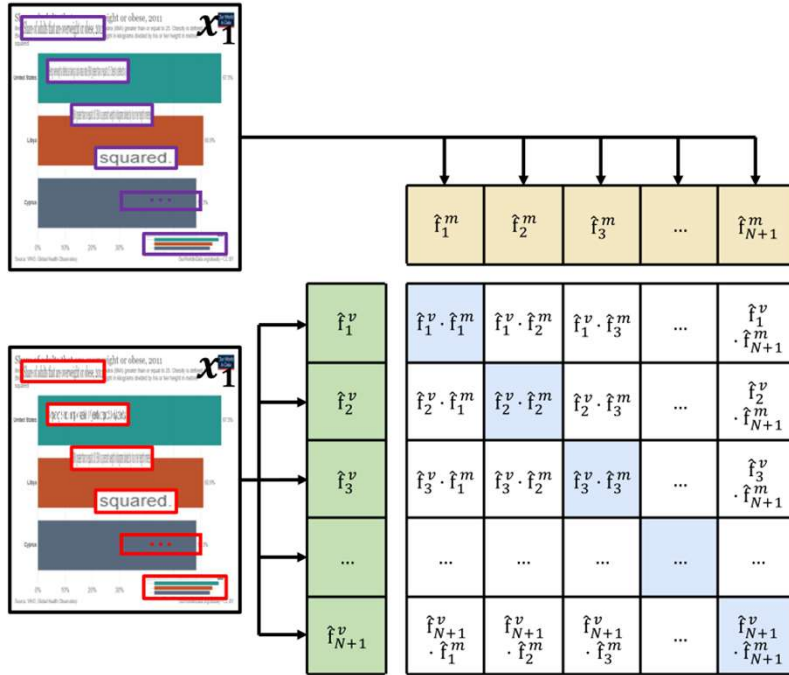
$$\mathcal{L}_{\text{DoCo}} = \mathcal{L}_{\text{Intra-DoCo}} + \mathcal{L}_{\text{Inter-DoCo}}.$$

Intra-DoCo



南京航空航天大学

Nanjing University of Aeronautics and Astronautics



(a) Intra-DoCo

$$\mathcal{L}_{\text{Intra-DoCo}}^x = -\frac{1}{N+1} \sum_{i=1}^{N+1} \log \left(\frac{e^{\text{sim}(\hat{\mathbf{f}}_i^v, \hat{\mathbf{f}}_i^m)}}{\sum_{j=1}^{N+1} e^{\text{sim}(\hat{\mathbf{f}}_i^v, \hat{\mathbf{f}}_j^m)}} \right)$$

$$\mathcal{L}_{\text{Intra-DoCo}}^y$$

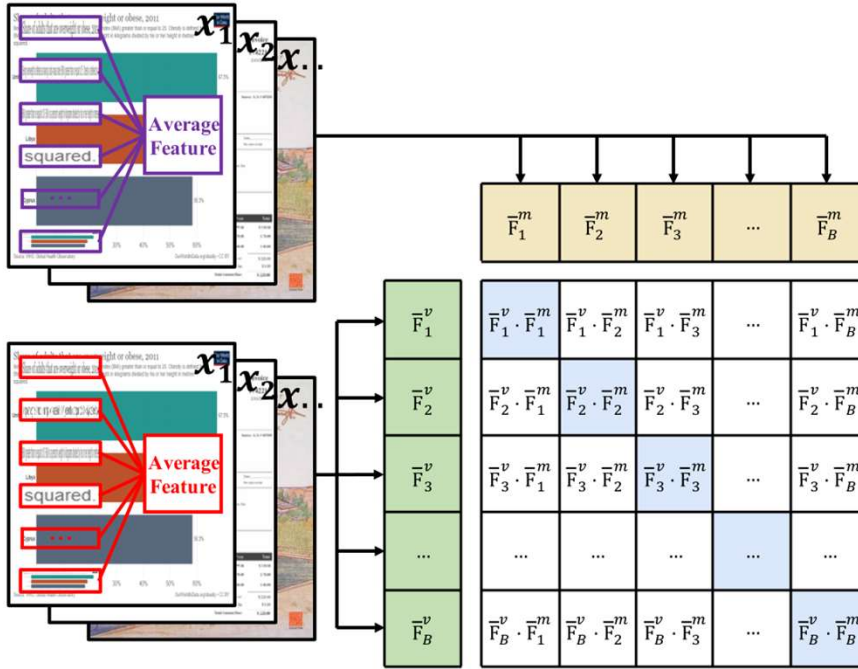
$$\mathcal{L}_{\text{Intra-DoCo}} = (\mathcal{L}_{\text{Intra-DoCo}}^x + \mathcal{L}_{\text{Intra-DoCo}}^y) / 2.$$

Inter-DoCo



南京航空航天大学

Nanjing University of Aeronautics and Astronautics



(b) Inter-DoCo

$$\hat{\mathbf{F}}^v = \{\hat{\mathbf{F}}_1^v, \hat{\mathbf{F}}_2^v, \dots, \hat{\mathbf{F}}_B^v\} \in \mathbb{R}^{B \times (N+1) \times d_v}$$

$$\hat{\mathbf{F}}^m = \{\hat{\mathbf{F}}_1^m, \hat{\mathbf{F}}_2^m, \dots, \hat{\mathbf{F}}_B^m\} \in \mathbb{R}^{B \times (N+1) \times d_v}$$

$$\mathcal{F}_{inter}^v = \{\bar{\mathbf{F}}_1^v, \bar{\mathbf{F}}_2^v, \dots, \bar{\mathbf{F}}_B^v\} \in \mathbb{R}^{B \times d_v}$$

$$\mathcal{F}_{inter}^m = \{\bar{\mathbf{F}}_1^m, \bar{\mathbf{F}}_2^m, \dots, \bar{\mathbf{F}}_B^m\} \in \mathbb{R}^{B \times d_v}$$

$$\mathcal{L}_{Inter-DoCo}^x = -\frac{1}{B} \sum_{i=1}^B \log \left(\frac{e^{sim(\bar{\mathbf{F}}_i^v, \bar{\mathbf{F}}_i^m)}}{\sum_{j=1}^B e^{sim(\bar{\mathbf{F}}_i^v, \bar{\mathbf{F}}_j^m)}} \right),$$

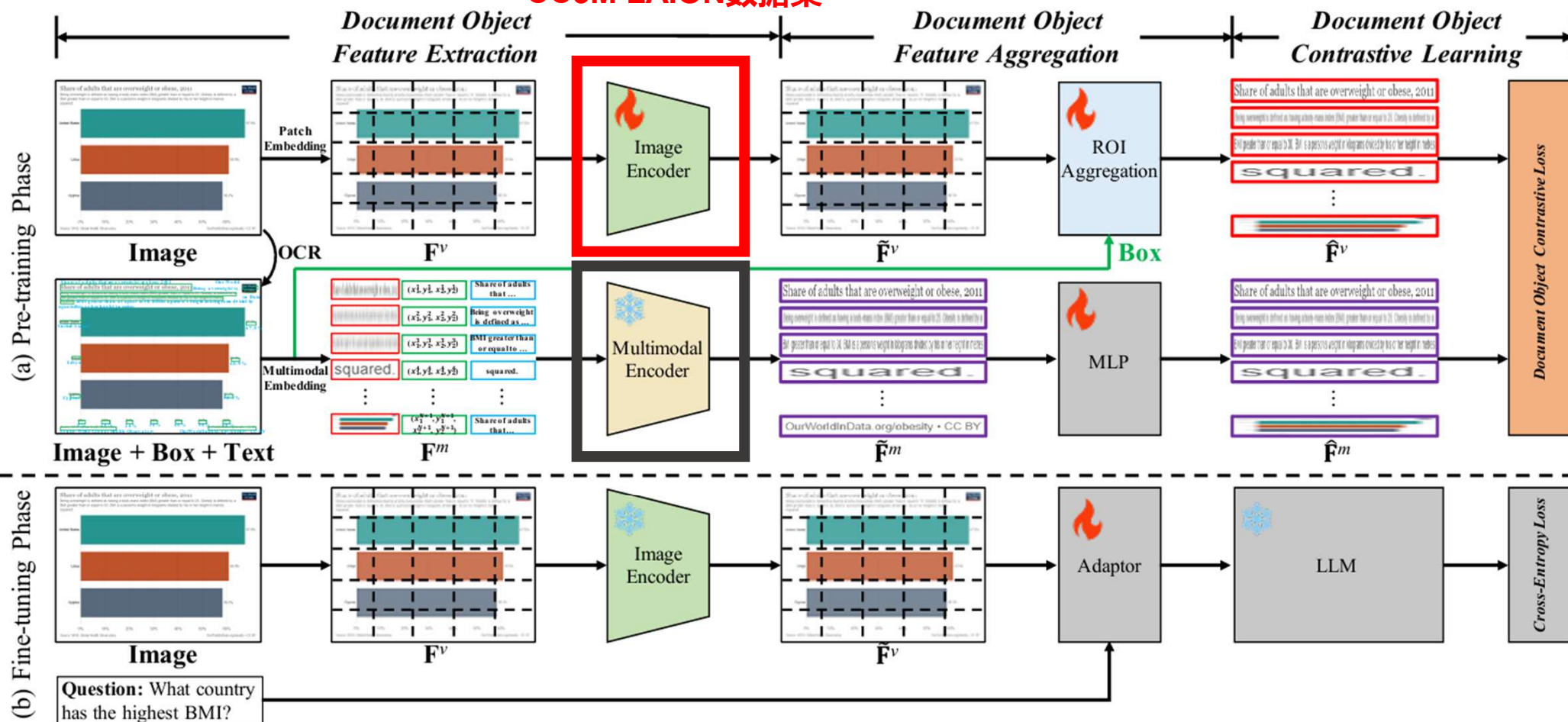
$$\mathcal{L}_{Inter-DoCo} = (\mathcal{L}_{Inter-DoCo}^x + \mathcal{L}_{Inter-DoCo}^y) / 2.$$

Training Strategies: Pre-training



南京航空航天大学
Nanjing University of Aeronautics and Astronautics

CC3M LAION数据集

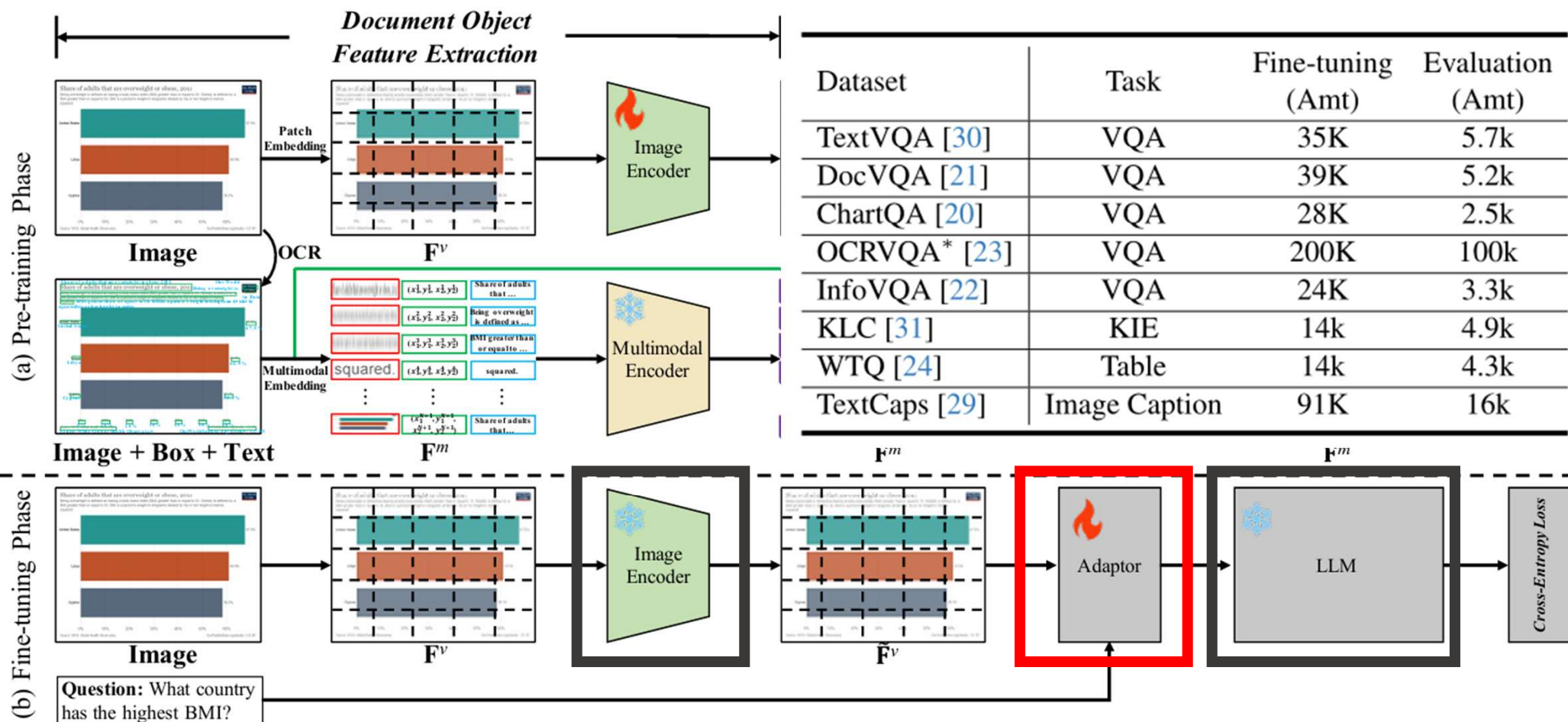


Training Strategies: Fine-tuning datasets



南京航空航天大学

Nanjing University of Aeronautics and Astronautics



Outcomes



南京航空航天大学
Nanjing University of Aeronautics and Astronautics

Method	Resolution	OCRVQA	TextVQA	DocVQA	InfoVQA	ChartQA	KLC	WTQ	TextCaps
MiniGPT-4 [‡] [43]	224 ²	11.5	18.7	3.0	13.3	4.3	-	-	-
mPLUG-Owl [‡] [40]	224 ²	28.6	40.2	6.9	16.5	9.5	-	-	-
Qwen-VL [2]	448 ²	75.7	63.8	65.1	-	65.7	-	-	-
Qwen-VL-Chat [2]	448 ²	70.5	61.5	62.6	-	66.3	-	-	-
mPLUG-DocOwl [38]	-	-	52.6	62.2	38.2	57.4	30.3	26.9	111.9
LLaVAR(336) [41]	336 ²	23.8	48.5	11.6	-	-	-	-	-
UReader [39]	224 ²	-	57.6	65.4	42.2	59.3	32.8	29.4	118.4
Qwen-VL-Chat [†]	448 ²	71.1	61.7	62.2	33.1	67.3	31.5	24.8	112.3
Qwen-VL-Chat^{††}	448 ²	73.2	63.6	64.8	34.9	68.9	33.8	26.9	114.5
mPLUG-Owl [†]	448 ²	70.3	53.5	61.8	32.5	58.3	31.2	25.2	113.4
mPLUG-Owl^{††}	448 ²	72.1	55.7	63.6	34.1	60.1	32.9	26.4	115.9

The models with “†” and “††” denote pre-training with **CLIP** and **DoCo** respectively, which are optimized with the same datasets and experimental settings for a fair comparison.

Accuracy----准确率

Outcomes



南京航空航天大学
Nanjing University of Aeronautics and Astronautics

Method	Resolution	OCRVQA	TextVQA	DocVQA	InfoVQA	ChartQA	KLC	WTQ	TextCaps
MiniGPT-4 [‡] [43]	224 ²	11.5	18.7	3.0	13.3	4.3	-	-	-
mPLUG-Owl [‡] [40]	224 ²	28.6	40.2	6.9	16.5	9.5	-	-	-
Qwen-VL [2]	448 ²	75.7	63.8	65.1	-	65.7	-	-	-
Qwen-VL-Chat [2]	448 ²	70.5	61.5	62.6	-	66.3	-	-	-
mPLUG-DocOwl [38]	-	-	52.6	62.2	38.2	57.4	30.3	26.9	111.9
LLaVAR(336) [41]	336 ²	23.8	48.5	11.6	-	-	-	-	-
UReader [39]	224 ²	-	57.6	65.4	42.2	59.3	32.8	29.4	118.4
Qwen-VL-Chat [†]	448 ²	71.1	61.7	62.2	33.1	67.3	31.5	24.8	112.3
Qwen-VL-Chat^{††}	448 ²	73.2	63.6	64.8	34.9	68.9	33.8	26.9	114.5
mPLUG-Owl [†]	448 ²	70.3	53.5	61.8	32.5	58.3	31.2	25.2	113.4
mPLUG-Owl^{††}	448 ²	72.1	55.7	63.6	34.1	60.1	32.9	26.4	115.9

The models with “†” and “††” denote pre-training with **CLIP** and **DoCo** respectively, which are optimized with the same datasets and experimental settings for a fair comparison.

Average Normalized
Levenshtein Similarity
(ANLS)-----平均归一化
Levenshtein相似度

Outcomes



南京航空航天大学
Nanjing University of Aeronautics and Astronautics

Method	Resolution	OCRVQA	TextVQA	DocVQA	InfoVQA	ChartQA	KLC	WTQ	TextCaps
MiniGPT-4 [‡] [43]	224 ²	11.5	18.7	3.0	13.3	4.3	-	-	-
mPLUG-Owl [‡] [40]	224 ²	28.6	40.2	6.9	16.5	9.5	-	-	-
Qwen-VL [2]	448 ²	75.7	63.8	65.1	-	65.7	-	-	-
Qwen-VL-Chat [2]	448 ²	70.5	61.5	62.6	-	66.3	-	-	-
mPLUG-DocOwl [38]	-	-	52.6	62.2	38.2	57.4	30.3	26.9	111.9
LLaVAR(336) [41]	336 ²	23.8	48.5	11.6	-	-	-	-	-
UReader [39]	224 ²	-	57.6	65.4	42.2	59.3	32.8	29.4	118.4
Qwen-VL-Chat [†]	448 ²	71.1	61.7	62.2	33.1	67.3	31.5	24.8	112.3
Qwen-VL-Chat^{††}	448 ²	73.2	63.6	64.8	34.9	68.9	33.8	26.9	114.5
mPLUG-Owl [†]	448 ²	70.3	53.5	61.8	32.5	58.3	31.2	25.2	113.4
mPLUG-Owl^{††}	448 ²	72.1	55.7	63.6	34.1	60.1	32.9	26.4	115.9

The models with “†” and “††” denote pre-training with **CLIP** and **DoCo** respectively, which are optimized with the same datasets and experimental settings for a fair comparison.

Relaxed Accuracy-
---放宽标准度

Outcomes



南京航空航天大学
Nanjing University of Aeronautics and Astronautics

Method	Resolution	OCRVQA	TextVQA	DocVQA	InfoVQA	ChartQA	KLC	WTQ	TextCaps
MiniGPT-4 [‡] [43]	224 ²	11.5	18.7	3.0	13.3	4.3	-	-	-
mPLUG-Owl [‡] [40]	224 ²	28.6	40.2	6.9	16.5	9.5	-	-	-
Qwen-VL [2]	448 ²	75.7	63.8	65.1	-	65.7	-	-	-
Qwen-VL-Chat [2]	448 ²	70.5	61.5	62.6	-	66.3	-	-	-
mPLUG-DocOwl [38]	-	-	52.6	62.2	38.2	57.4	30.3	26.9	111.9
LLaVAR(336) [41]	336 ²	23.8	48.5	11.6	-	-	-	-	-
UReader [39]	224 ²	-	57.6	65.4	42.2	59.3	32.8	29.4	118.4
Qwen-VL-Chat [†]	448 ²	71.1	61.7	62.2	33.1	67.3	31.5	24.8	112.3
Qwen-VL-Chat^{††}	448 ²	73.2	63.6	64.8	34.9	68.9	33.8	26.9	114.5
mPLUG-Owl [†]	448 ²	70.3	53.5	61.8	32.5	58.3	31.2	25.2	113.4
mPLUG-Owl^{††}	448 ²	72.1	55.7	63.6	34.1	60.1	32.9	26.4	115.9

The models with “†” and “††” denote pre-training with **CLIP** and **DoCo** respectively, which are optimized with the same datasets and experimental settings for a fair comparison.

F1-score----精确率 (Precision) 和召回率 (Recall) 的调和平均值

Outcomes



南京航空航天大学
Nanjing University of Aeronautics and Astronautics

Method	Resolution	OCRVQA	TextVQA	DocVQA	InfoVQA	ChartQA	KLC	WTQ	TextCaps
MiniGPT-4 [‡] [43]	224 ²	11.5	18.7	3.0	13.3	4.3	-	-	-
mPLUG-Owl [‡] [40]	224 ²	28.6	40.2	6.9	16.5	9.5	-	-	-
Qwen-VL [2]	448 ²	75.7	63.8	65.1	-	65.7	-	-	-
Qwen-VL-Chat [2]	448 ²	70.5	61.5	62.6	-	66.3	-	-	-
mPLUG-DocOwl [38]	-	-	52.6	62.2	38.2	57.4	30.3	26.9	111.9
LLaVAR(336) [41]	336 ²	23.8	48.5	11.6	-	-	-	-	-
UReader [39]	224 ²	-	57.6	65.4	42.2	59.3	32.8	29.4	118.4
Qwen-VL-Chat [†]	448 ²	71.1	61.7	62.2	33.1	67.3	31.5	24.8	112.3
Qwen-VL-Chat^{††}	448 ²	73.2	63.6	64.8	34.9	68.9	33.8	26.9	114.5
mPLUG-Owl [†]	448 ²	70.3	53.5	61.8	32.5	58.3	31.2	25.2	113.4
mPLUG-Owl^{††}	448 ²	72.1	55.7	63.6	34.1	60.1	32.9	26.4	115.9

The models with “†” and “††” denote pre-training with **CLIP** and **DoCo** respectively, which are optimized with the same datasets and experimental settings for a fair comparison.

Consensus-based
Image Description
Evaluation (CIDER)---
基于共识的图像描述
评估

Outcomes



南京航空航天大学

Nanjing University of Aeronautics and Astronautics

ROUTING AND TRANSMITTAL SLIP Date April 10, 1979

To: (Name, office symbol, room number, building, Agency/Post) Initials Date

1. William J. Darby, M.D., Ph.D.

2.

3.

4.

Action Approval File For Clearance Note and Return

As Requested Per Conversation Per Conversation

Complete 3/22 For Your Information Please Reply

Comment Investigate See Me

Coordination Justify

REMARKS

Norman Kretschmer, M.D., Ph.D.

DO NOT use this form as a RECORD of approvals, enclosures, disposals, clearances, and similar actions

FROM: (Name, org. symbol, Agency/Post) Page 1 of 1

Office of the Director, SECDD

Phone No. 496-3454

OPTIONAL FORM 41 (Rev. 7-75)

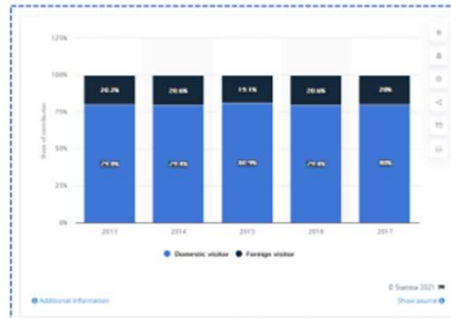
Printed on 10/10/78

Source: <https://www.industrydocuments.ucsf.edu/docs/667029>

Question: What is the "phone number" given on the slip?

Qwen-VL-Chat†: 5041-102

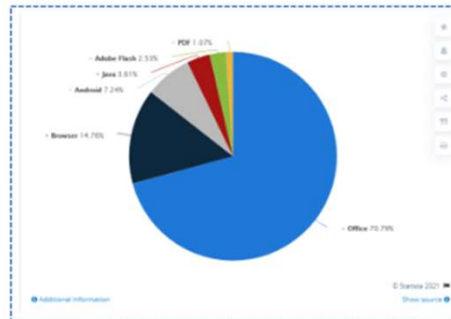
Qwen-VL-Chat††: 496-3454



Question: Can you identify what is the percentage value of foreign visitor in 2017?

Qwen-VL-Chat†: 80

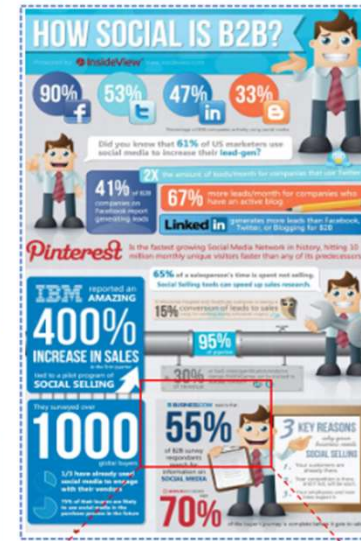
Qwen-VL-Chat††: 20



Question: Find which category is shown in yellow color?

Qwen-VL-Chat†: Adobe Flash 2.53

Qwen-VL-Chat††: PDF



Question: What percentage of B2B survey respondents search for information on social media?

Qwen-VL-Chat†: 41%

Qwen-VL-Chat††: 55%

Monograph Health Publication Management

June 13, 2001

10:00 - 11:30 AM, Conference Room 4555-4B

page 6 of 6

III. Upcoming Presentations/Publications

IV. Women's BOP Publications

V. Upcoming Data

VI. PB Opportunities

VII. Upcoming Meetings and Deadlines

MEETING	DATE	DEADLINE
ISGE - World Congress of Gastroenterology (Hong Kong)	December 2-5, 2001	July 15, 2001
ACC - American College of Cardiology	March 2002	September 5, 2001
ACOG - American College of Obstetricians and Gynecologists	May 2002	September 26, 2001
AANP - American Academy of Nurse Practitioners	June 2002	October 2001
ASAC - American Association of Cancer Research	March 2002	September 26, 2001
AAN - American Academy of Neurology	May 2002	November 2001
ASCO - American Society for Clinical Oncology	May 2002	November 2001

Source: <https://www.industrydocuments.ucsf.edu/docs/667029>

Question: What is the deadline given for AANP?

Qwen-VL-Chat†: June 2002

Qwen-VL-Chat††: October 2001

Qualitative results between **CLIP** ("†") and **DoCo** ("††"). Crucial regions are enlarged for clearer visualization.

Ablation Study



南京航空航天大学
Nanjing University of Aeronautics and Astronautics

Method	Intra	Inter	R	A	I	B	T	DocVQA
Qwen-VL-Chat [*] _{w/o DoCo}	✗	✗	✗	✗	✗	✗	✗	62.2
Qwen-VL-Chat [*] _{w/ Intra&R&I&B&T}	✓	✗	✓	✗	✓	✓	✓	63.9
Qwen-VL-Chat [*] _{w/ DoCo&A&I&B&T}	✓	✓	✗	✓	✓	✓	✓	64.2
Qwen-VL-Chat [*] _{w/ DoCo&R&T}	✓	✓	✓	✗	✗	✗	✓	63.7
Qwen-VL-Chat [*] _{w/ DoCo&R&B&T}	✓	✓	✓	✗	✗	✓	✓	64.4
Qwen-VL-Chat^{††}	✓	✓	✓	✗	✓	✓	✓	64.8



南京航空航天大学
Nanjing University of Aeronautics and Astronautics

恳请各位老师批评指正

汇报人：赵翔鸽
