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

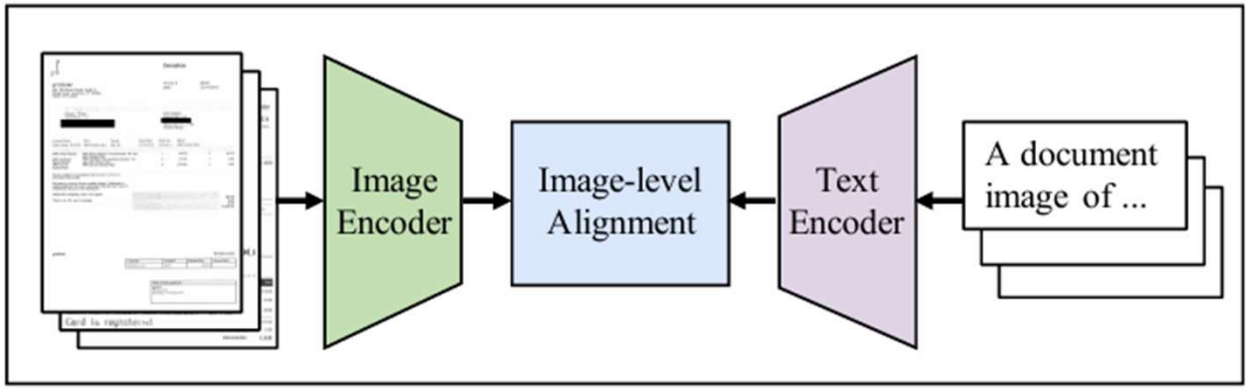
Enhancing Visual Document Understanding with Contrastive Learning in Large Visual-Language Models

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, rechycao,
jasonysliu, dqiangjiang, winfredsun}@tencent.com

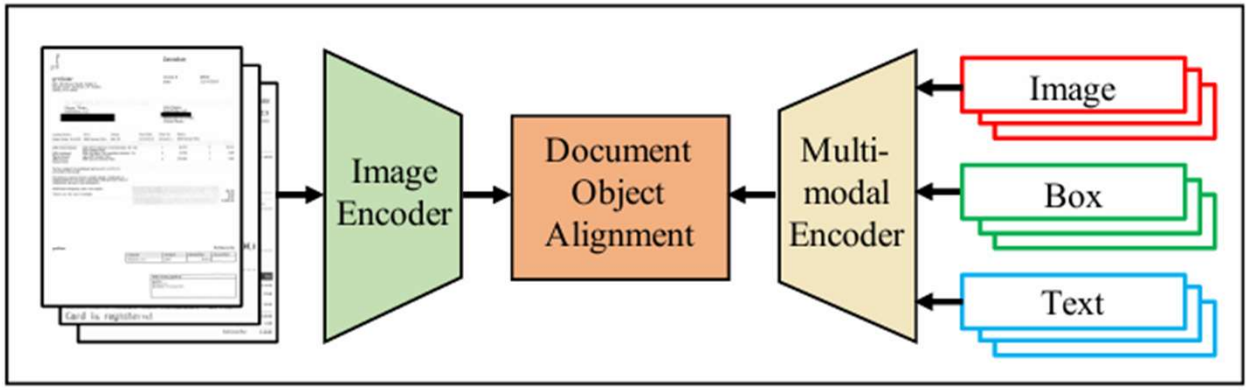
CVPR2024

Background

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



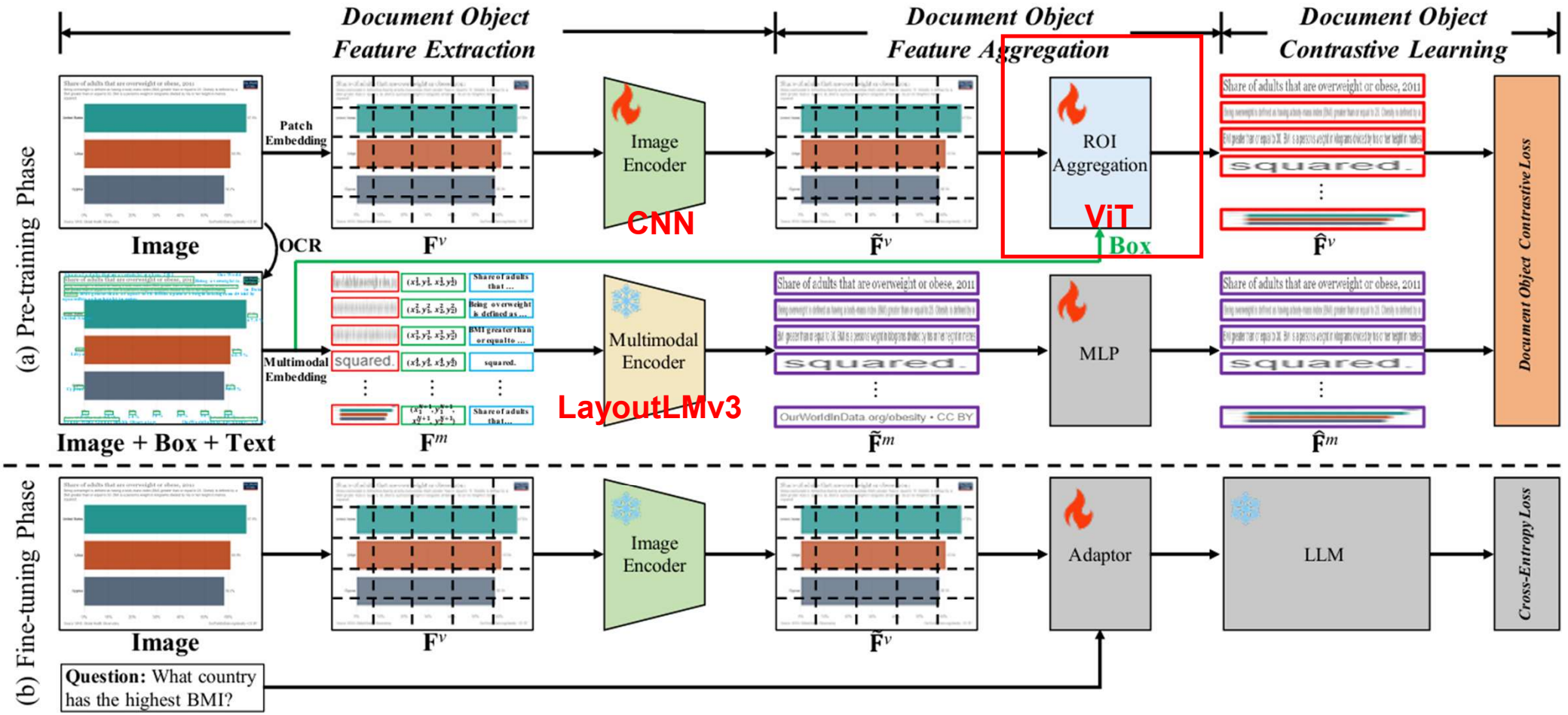
(a) Image-level instance discrimination (CLIP)



(b) Document object discrimination (Our DoCo)

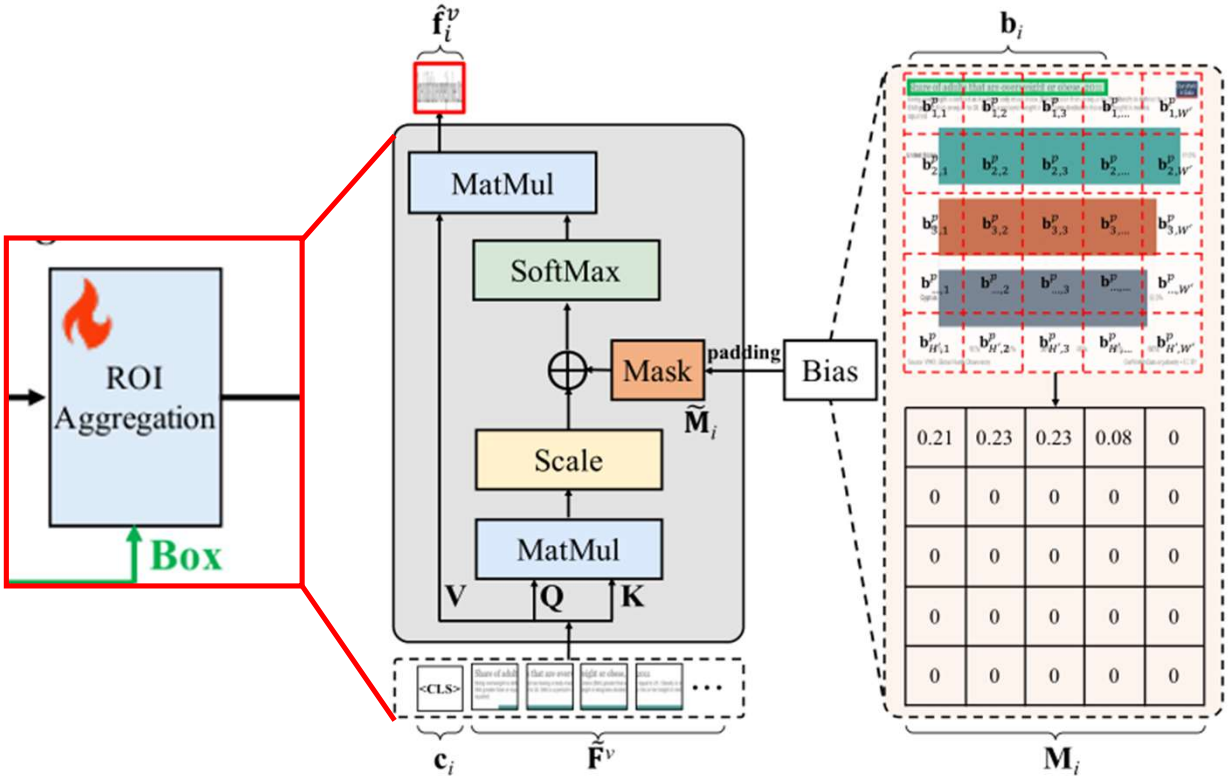
Schematic Overview

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



ROI Aggregation

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



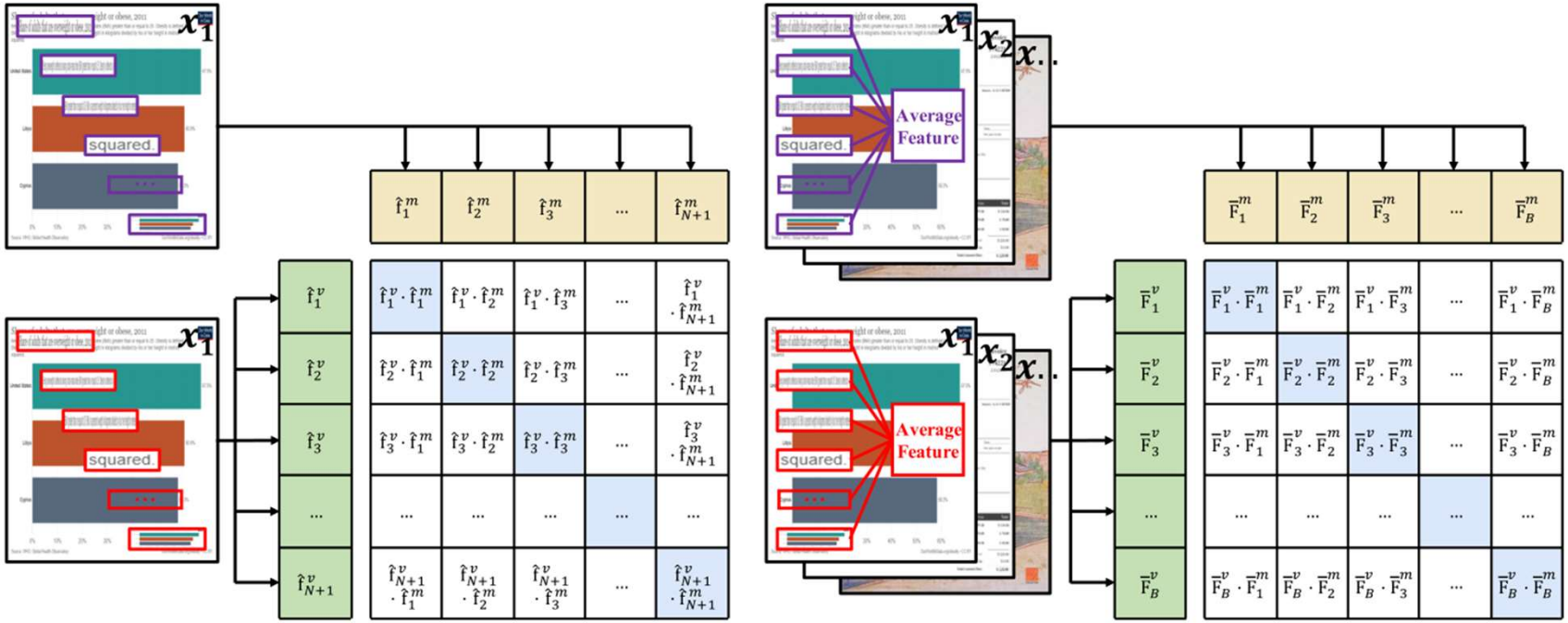
$$\mathbf{M}_i = Overlap(\mathbf{b}_i, \mathbf{b}^p)/Area(\mathbf{b}^p),$$

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

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

Illustrative View of DoCo

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

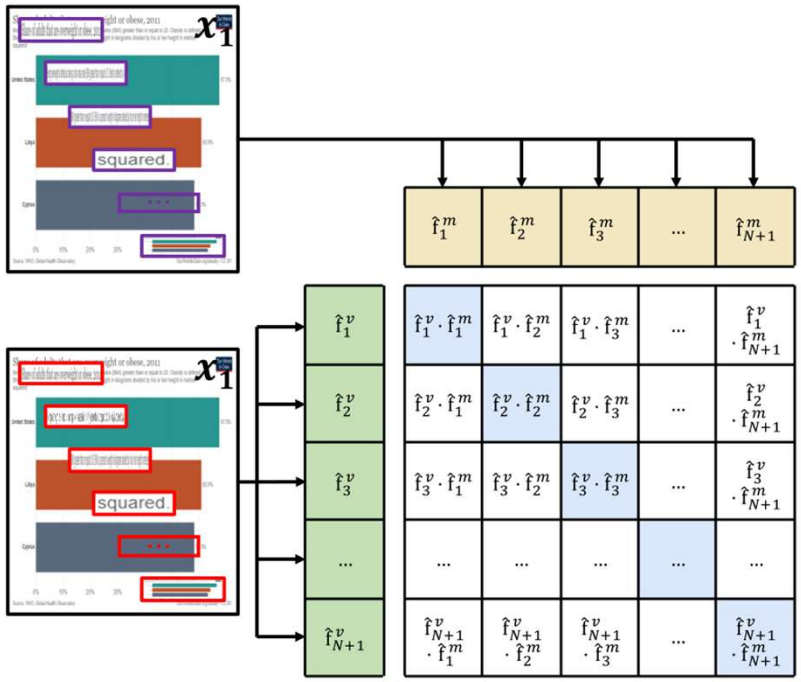


(a) Intra-DoCo (b) Inter-DoCo

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

Intra-DoCo

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



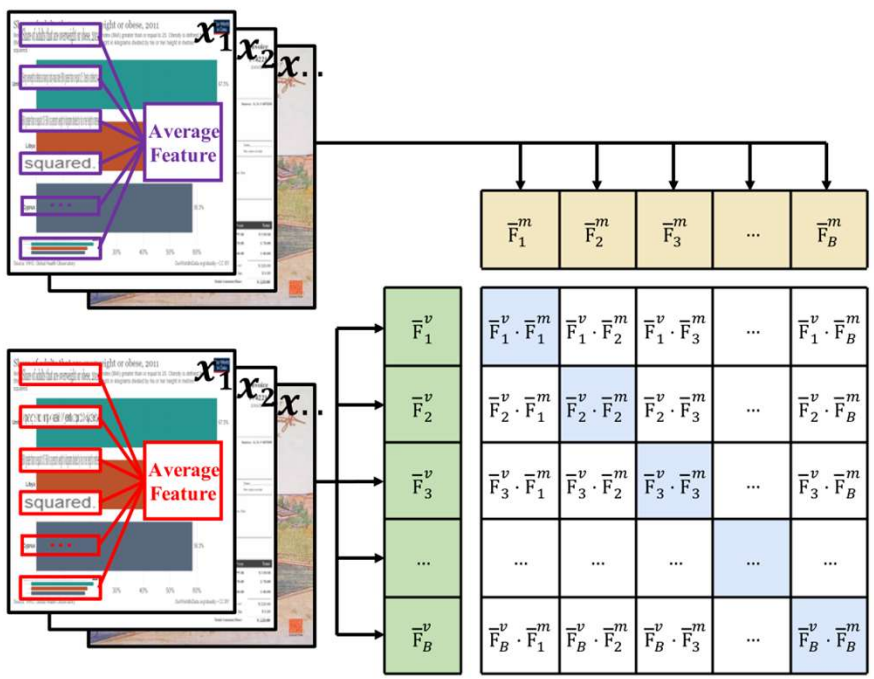
(a) Intra-DoCo

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

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

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

Inter-DoCo



(b) Inter-DoCo

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

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

$$\mathcal{F}_{inter}^v = \left\{\overline{\mathbf{F}}_1^v, \overline{\mathbf{F}}_2^v, ..., \overline{\mathbf{F}}_B^v\right\} \in \mathbb{R}^{B\times d_v}$$

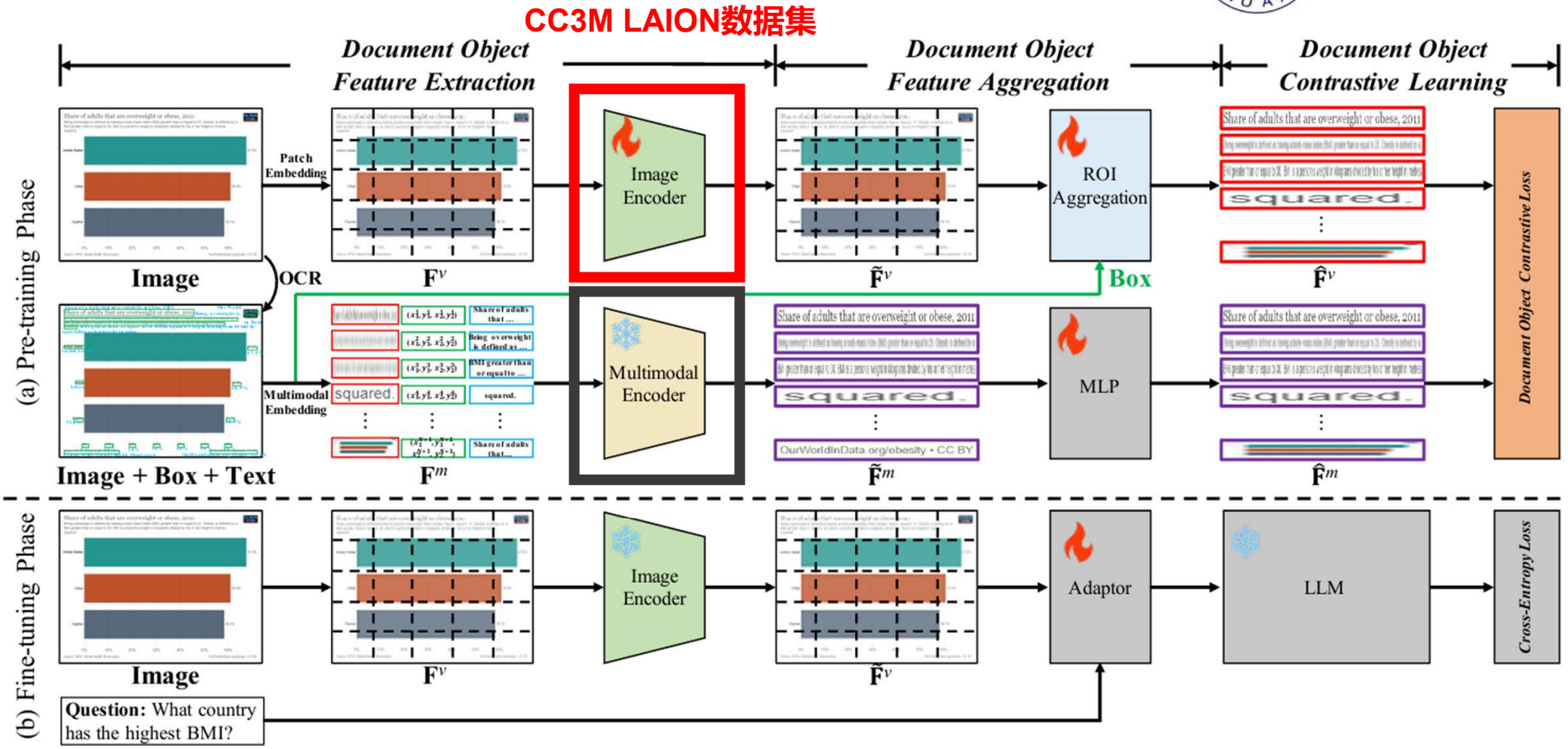
$$\mathcal{F}_{inter}^m = \left\{\overline{\mathbf{F}}_1^m, \overline{\mathbf{F}}_2^m, ..., \overline{\mathbf{F}}_B^m\right\} \in \mathbb{R}^{B\times d_v}$$

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

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

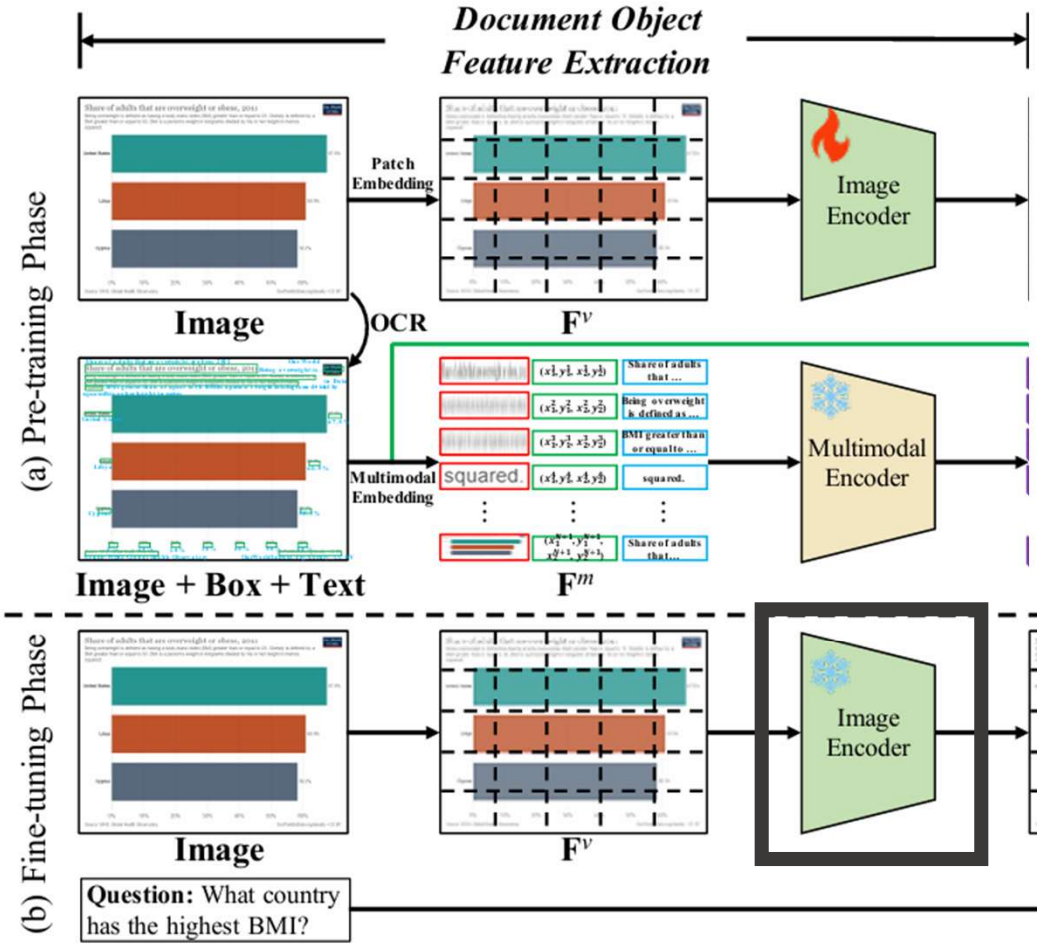
Training Strategies: Pre-training

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



Training Strategies: Fine-tuning datasets

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



Dataset	Task	Fine-tuning (Amt)	Evaluation (Amt)
TextVQA [30]	VQA	35K	5.7k
DocVQA [21]	VQA	39K	5.2k
ChartQA [20]	VQA	28K	2.5k
OCRVQA* [23]	VQA	200K	100k
InfoVQA [22]	VQA	24K	3.3k
KLC [31]	KIE	14k	4.9k
WTQ [24]	Table	14k	4.3k
TextCaps [29]	Image Caption	91K	16k

Outcomes

南京航空航天大學
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

南京航空航天大學
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

南京航空航天大學
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

南京航空航天大學
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

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

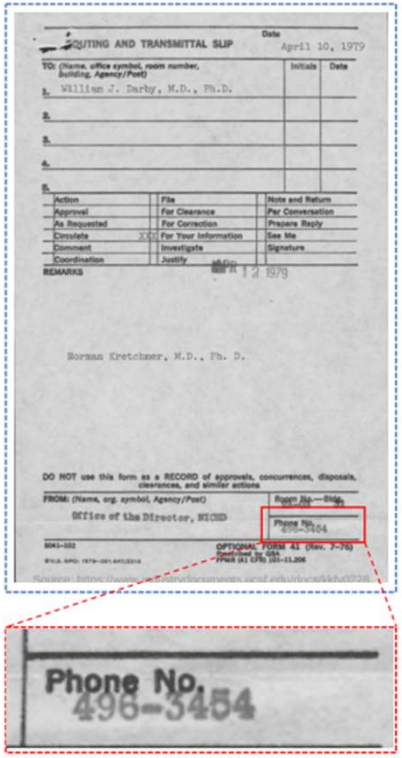
Method	Resolution	OCRVQA	TextVQA	DocVQA	InfoVQA	ChartQA	KLC	WTQ	TextCaps
MiniGPT-4‡ [43]	224^2	11.5	18.7	3.0	13.3	4.3	-	-	-
mPLUG-Owl‡ [40]	224^2	28.6	40.2	6.9	16.5	9.5	-	-	-
Qwen-VL [2]	448^2	75.7	63.8	65.1	-	65.7	-	-	-
Qwen-VL-Chat [2]	448^2	70.5	61.5	62.6	-	66.3	-	-	-
mPLUG-DocOwl [38]	-	-	52.6	62.2	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.

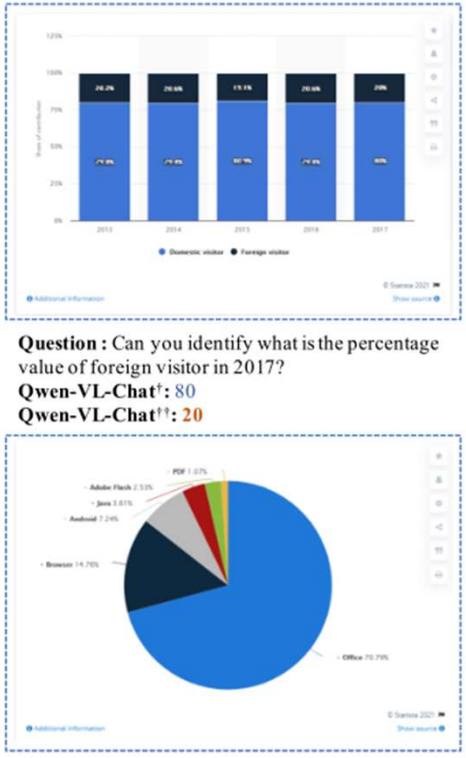
Consensus-based Image Description Evaluation (CIDEr)---基于共识的图像描述评估

Outcomes

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

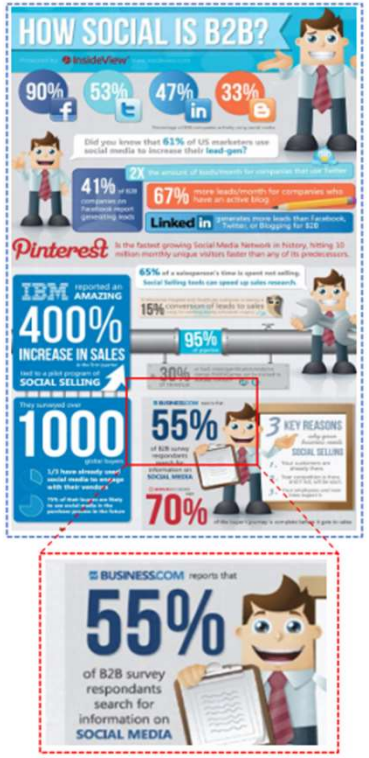


Question: What is the "phone number" given on the slip?
Qwen-VL-Chat†: 5041-102
Qwen-VL-Chat††: 496-3454

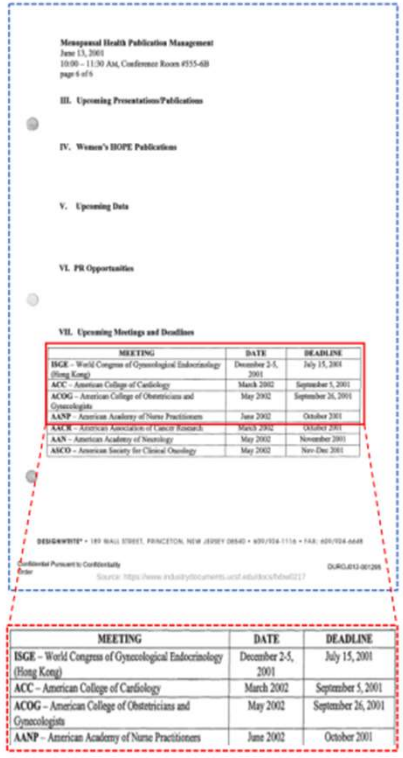


Question : Can you identify what is the percentage value of foreign visitor in 2017?
Qwen-VL-Chat†: 80
Qwen-VL-Chat††: 20

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%



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

南京航空航天大學
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**

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

恳请各位老师批评指正

汇报人：赵翔鸽