CLIP の Modality Gap を考慮した RAG 検索手法の改良

山下 修平 [1]　白藤 大幹 [1]　斉藤 辰彦 [1]
[1] 三菱電機株式会社　情報技術総合研究所
{Yamashita.Shuhei@bc, Shirafuji.Daiki@ay, Saito.Tatsuhiko@db}
.MitsubishiElectric.co.jp

概要

CLIP の埋め込み空間はテキストと画像で大きく分離している (modality gap) ため，異なるモーダル間の類似度が低いことが問題となる．本研究では，事前実験として画像とテキストのデータを外部知識に持つ RAG の検索器に CLIP を用い，画像が必要なケースにおいて正解画像データが検索上位に現れないことを確認する．そこで，外部知識のモーダルに応じて検索スコアを標準化することで，modality gap を考慮した検索手法を提案する．評価実験により，画像の外部知識を必要とするケースの検索精度は 0%から 56%に改善し，RAG の生成精度は ROUGE-1 で 0.26 ポイント向上したことを確認した．

1　はじめに

近年大規模言語モデル (LLM) の発達は著しく，質問応答 (QA) など多様なタスクに活用されている．しかし，LLM は学習データに含まれていない知識を問われた際に，事実と異なる回答を生成する問題が広く知られている [1].

Retrieval-Augmented Generation (RAG) [2] は LLM に外部知識を参照させることで，LLM の回答精度を向上させるアプローチであり，主に QA タスクで用いられる．RAG では質問文に関連する外部知識を検索し，LLM に検索結果の外部知識を参照させることで回答を生成する．

さらに，LLM の実世界応用が重要視されつつあり，テキストのみでなく画像を扱える Large Vision-Lauguage Models (LVLMs) の研究が進んでいる [3]．LVLMs はテキストと画像を外部知識として扱う RAG (以下，マルチモーダル RAG) に活用され始めている [4, 5, 6].

マルチモーダル RAG の検索手法の 1 つとして，マルチモーダルな埋め込みモデルを用いたベクトル検索がある．Contrastive Language-Image Pre-training (CLIP) [7] はテキスト・画像を共通の潜在空間に埋め込むことができるモデルとして知られており，CLIP による埋め込みベクトルを利用した検索システムが広く構築されている．

しかし，CLIP には “modality gap” [8] と呼ばれる，テキストと画像の埋め込み空間がほぼ完全に分離する現象が知られている．そのため，CLIP の埋め込みベクトルに基づく検索では，クエリと同じモーダルの外部知識が検索上位を独占し，RAG の回答生成に悪影響を及ぼすと考えられる (図 1 (a)).

この問題に対し，Eslami ら [9] は CLIP を同じモーダル同士についても対象損失をとる学習設計に改良することで，modality gap を縮小した．また，Talmor ら [10] は質問文に応じて正答に必要なモーダルを判別する手法を提案した．しかし，いずれの手法も学習データの作成やモデルの学習が必要となる．

そこで本研究では CLIP の追加学習などを必要とせず，modality gap の影響を軽減した検索手法を提案する．提案手法は質問文と外部知識の埋め込みベクトル同士の cos 類似度を，外部知識のモーダルに応じた標準化を施すことでスケールを調整する (図 1 (b)). 実験により，提案手法はクエリと異なるモーダルの外部知識に対する検索精度，および RAG の回答文の精度を向上させたことを確認した．

2　関連研究

CLIP の Modality Gap　CLIP [7] はテキストと画像を共通の潜在空間に埋め込むように対照学習したマルチモーダルモデルである．Liang ら [8] は，CLIP の埋め込み空間がテキストと画像で分離する，modality gap の問題を提起した．この問題に対し，Eslami ら [9] は異なるモーダル間に加えて，同一モーダル間の対照学習を行う AlignCLIP を提案した．また Fahim ら [11] は埋め込みが潜在空間上で均等に分布するように，埋め込み間の距離を大きくさせる学習項を CLIP の損失関数に追加することを提

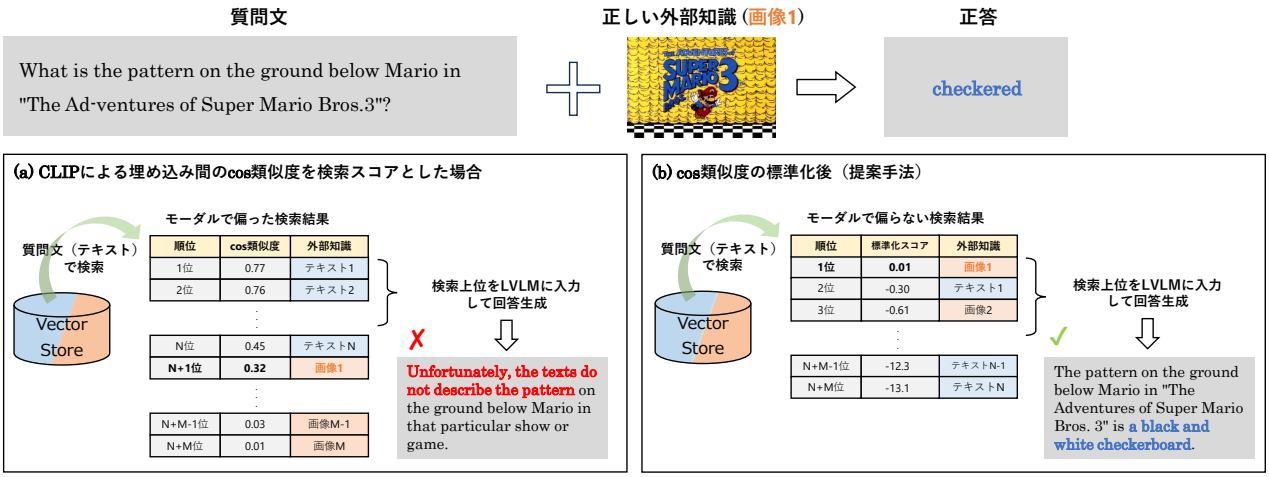


図 1: 本研究の概略図．(a) CLIP による埋め込み空間はテキストと画像でほぼ完全に分離しているため (modality gap)，検索上位をクエリと同じテキストモーダルの外部知識が占める．(b) 提案手法は cos 類似度を外部知識のモーダルに応じて標準化した値を検索スコアとすることで，modality gap の影響を軽減した検索を可能とする．

案した．これらの先行研究は CLIP の下流タスクにおける性能を劣化させることなく，modality gap を縮小させた．

マルチモーダル RAG　RAG [2] はベクトル検索などでクエリに関連する外部知識を抽出し，関連する知識とクエリを組み合わせて LLM に与えて回答を生成する手法である．外部知識としてテキストに加えて画像を扱える，マルチモーダル RAG の検索手法に CLIP が活用されている [5, 6]．Riedler と Langer [5] は，テキストと画像に対してそれぞれ別の手法で検索することで実現した．一方，本研究はテキストと画像を同一の手法によって検索するマルチモーダル RAG の構築を行う．

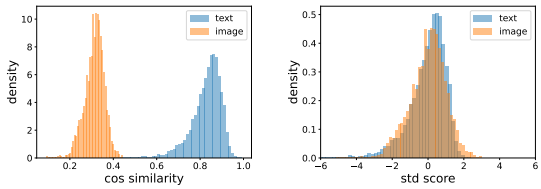
3　事前実験

本節では，modality gap が QA データにおいても存在するかを確認するための事前実験を実施する．

3.1　実験設定

データセットとして Multi Modal QA (MMQA) [10] を用いて実験を行う．MMQA は Wikipedia から抽出されたテキスト・画像・表モーダルのデータを外部知識として，その内容を問う QA データセットである．各 QA データには正答と，正答に必要な正しい外部知識が与えられている (詳細は付録 A 参照)．

本実験では画像とテキスト間の modality gap を確認することを目的とするため，正しい外部知識のモーダルがテキストである TextQ，画像である ImageQ に絞って使用する (表 1 にサンプルを示す)．



(a) 標準化前　(b) 標準化後

図 2: (a) MMQA 訓練データ中の質問文と正しい外部知識から計算される cos 類似度のヒストグラム．(b) cos 類似度をモーダル別に標準化したスコアのヒストグラム．

MMQA の訓練データにおける各 QA の質問文と正しい外部知識について，CLIP[1] による埋め込みベクトル同士の cos 類似度を計算する．

3.2　実験結果と考察

各 QA の質問文と正しい外部知識の cos 類似度を，外部知識のモーダル別にヒストグラムとして図 2a に示す．またその平均と分散を表 5 に示す．

正しい外部知識が質問文と同じモーダルのテキストである場合の方が，画像である場合と比べて cos 類似度が大きい値をとる．すなわち QA データにおいても，modality gap が確認された．この結果から，CLIP をそのまま RAG の検索器に用いると，画像を必要とする質問に対してもテキストを LVLM に入力して，生成精度を劣化させることが示唆される．

一方で，ヒストグラムはいずれも似た形状をしており，標準化してスケールを調整することで統一的な尺度で扱うことができると考えられる．

1)　https://huggingface.co/openai/clip-vit-base-patch32

表 1: MMQA の QA サンプル

QA の種類	質問文	正答	正しい外部知識
TextQ	When was the last Nintendo 64 game released?	August 20, 2002	The Nintendo 64 was first launched in Japan on June 23, 1996 ···. The last game to be published for the system was the North American-locked Tony Hawk's Pro Skater 3 on **August 20, 2002**.
ImageQ	What is the pattern on the ground below Mario in "The Adventures of Super Mario Bros. 3"?	checkered	図 1 の画像 1

4 提案手法

事前実験の結果と考察を踏まえて，本研究では CLIP による埋め込みベクトル間の cos 類似度を，外部知識のモーダルに応じて標準化することで，CLIP の modality gap を軽減した検索手法を提案する．本手法では，式 (1) に示すように，質問文 Q と外部知識 K の cos 類似度を K のモーダル M に応じて標準化し，その値を検索スコアとして用いる．

$$score(Q,K) = \frac{\cos(Q,K) - \mu_M}{\sqrt{\sigma_M^2}} \tag{1}$$

ここで，μ_M と σ_M^2 はモーダル M の外部知識を正答に必要とする質問文とその外部知識の，CLIP による埋め込み間の cos 類似度の平均と分散であり，訓練データから事前に計算する (値は付録 B を参照)．

図 2a の各 cos 類似度に対して，提案手法を適用した結果のヒストグラムを図 2b に示す．標準化によって，テキストと画像でギャップがあった cos 類似度が同じスケールに調整されている．

本検索手法によりスコアリングされた外部知識から，上位 N 件を取得し LVLM に入力してマルチモーダル RAG を構築する．

5 実験設定

5.1 データセット

3 節の事前実験で用いた MMQA [10] の，テスト用データの TextQ (721 件) と ImageQ (230 件) を用いる．外部知識はテキストが約 210K 件，画像が約 60K 件存在し，全体で約 270K 件の中から正しい外部知識を検索する必要がある．

5.2 検索精度の評価

提案手法の質問文に対して正しい外部知識を検索する能力を評価する．ベースラインとして，以下の 2 つの検索手法を評価する．

naive CLIP の埋め込みベクトル間の cos 類似度に基づいて検索する．

modal-clf [10] 質問文を入力として必要な外部知識のモーダルを予測するモーダル判別器を利用して，検索対象とする外部知識のモーダルを指定する手法．モーダル判別器は，BERT[2] [12] によるテキスト分類モデルとして実装した．

検索精度の評価指標として，正しい外部知識が検索上位 k 件に含まれている割合を評価する，Recall@k を用いる．

5.3 RAG 生成文の評価

提案手法を用いた RAG の生成文を評価する．5.2 節で述べた naive, modal-clf に加えて，外部知識の検索を行わない LVLM のみによる回答生成 (**w/o RAG**) を比較手法の 1 つとする．

質問文に対して生成文が正しく回答しているかを評価するために，自動評価指標に ROUGE-1 を用いる．ROUGE-1 を採用した理由は，QA データの正答の多くが単語形式であるためである．

加えて，人手評価を，以下の評価基準に基づき 1 人の作業者により実施する．

- 1：質問に対して，正しい答えを出力している
- 0.5：複数の回答を列挙することを必要とする場合に，その一部のみを出力している
- 0：正しい答えを出力できていない

人手評価はコストの観点から，テストデータ全体の一部 (100 件) を対象として実施した．

提案手法，ベースライン手法いずれにおいても，LVLM に GPT-4o (2024-08-06) [13] を使用する．また RAG においては，検索上位 10 件の外部知識を，付録 C 記載のプロンプトで LVLM に入力して回答を

2) https://huggingface.co/google-bert/bert-base-uncased

表 2: 各手法の Recall@k

QA	手法	検索件数 k 1	3	5	10	50	100
TextQ	naive	0.09	0.24	0.29	0.36	0.49	0.54
	Ours	0.08	0.20	0.24	0.32	0.45	0.50
	modal-clf	0.08	0.24	0.29	0.35	0.48	0.54
ImageQ	naive	0.00	0.00	0.00	0.00	0.00	0.00
	Ours	**0.35**	**0.48**	**0.51**	**0.56**	**0.66**	**0.69**
	modal-clf	0.42	0.54	0.60	0.64	0.75	0.78
Overall	naive	0.06	0.18	0.22	0.27	0.37	0.41
	Ours	**0.14**	**0.27**	**0.31**	**0.38**	**0.50**	**0.55**
	modal-clf	0.17	0.31	0.36	0.42	0.55	0.60

生成する．

6 実験結果と考察

6.1 検索精度の評価結果

各手法の検索精度を評価した結果を表 2 に示す．naive は ImageQ において検索件数を 100 件まで増やしても，Recall@k は 0 ポイントのままであった．この結果は，modality gap の影響で正しい外部知識を全く検索できなかったことを示す．

一方で提案手法は TextQ で精度を維持しつつ，ImageQ では Recall@10 で 56% 精度を向上させており，modality gap の影響を大きく軽減している．

また modal-clf と比較すると，TextQ と ImageQ いずれにおいても精度は劣り，Recall@10 で全体として 4%劣った．この差異は，modal-clf は訓練データでモーダル判別モデルを学習しているのに対して，提案手法は学習フリーであることに起因する．

6.2 RAG 生成文の評価結果

各生成文の評価結果と，ROUGE-1 (100 件) と人手評価 (100 件) の相関係数を表 3 に示す．ROUGE-1 と人手評価の間には正の相関があり，ROUGE-1 は適切な自動評価の指標であるといえる．

naive は ImageQ において，外部知識の検索を行わない w/o RAG よりも ROUGE-1 (全件) が 0.18 ポイント劣化している．これはプロンプト中の質問文に関連しない外部知識が，回答生成のノイズになったためだと考えられる．

提案手法を用いた RAG は，ImageQ において w/o RAG よりも ROUGE-1 (全件) が 0.08 ポイント，naive よりも 0.26 ポイント精度を改善した．よって，CLIP の modality gap を考慮した提案手法によって生成文の精度が向上したといえる．

表 3: 各手法による生成文の評価結果

QA	手法	人手 (100 件)	ROUGE-1 (100 件)	ROUGE-1 (全件)	相関係数
TextQ	w/o RAG	0.67	0.65	0.67	0.68
	naive	0.65	0.63	0.71	0.95
	Ours	**0.69**	**0.67**	0.70	0.95
	modal-clf	0.64	0.64	0.70	0.91
ImageQ	w/o RAG	0.34	0.44	0.39	0.63
	naive	0.16	0.28	0.21	0.57
	Ours	**0.58**	**0.58**	**0.47**	0.75
	modal-clf	0.70	0.64	0.54	0.78
Overall	w/o RAG	0.50	0.54	0.60	0.67
	naive	0.40	0.45	0.59	0.81
	Ours	**0.63**	**0.62**	**0.64**	0.84
	modal-clf	0.67	0.64	0.66	0.84

6.3 エラー分析

提案手法によって modality gap の影響を軽減できた ImageQ には，質問文に固有表現が含まれる傾向が見られた (サンプルは付録 E に記載)．例えば，固有表現である “The Wolverine” (映画名) や “Mississippi” (地名) を含む質問文では検索が成功した．これは，CLIP が固有表現のテキストと画像とを強く関連付けて学習したことで，高い検索スコアが得られたためだと考えられる．

逆に，固有表現を含まない質問文 “What is over the ... Hip hop music?” (全文は表 7 に記載) では，正解外部知識の画像内順位は上位 2 位と高いものの，テキストを含めた順位は 6,048 位であり，依然として modality gap の影響が残っていることを確認した．

今後は，固有表現以外のテキストと画像の関連性を強く捉えられる手法が精度改善に必要である．

7 おわりに

本研究では，外部知識のモーダルに応じて標準化を施すことで，CLIP の追加学習なしに modality gap を考慮可能な検索手法を提案し，マルチモーダルな QA データセットで評価した．実験の結果，提案手法により質問文と異なるモーダル (画像) の外部知識を必要とするケースの検索精度は 0%から 56%に改善し，RAG の生成精度も ROUGE-1 で 0.26 ポイント向上した．今後は，提案手法では対応できなかった固有表現を含まないテキストと画像との modality gap を埋める手法を検討する．

参考文献

[1] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. **arXiv preprint arXiv:2309.01219**, 2023.

[2] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474. Curran Associates, Inc., 2020.

[3] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the Frontier of Vision-Language Models: A Survey of Current Methodologies and Future Directions. **arXiv preprint arXiv:2404.07214**, 2024.

[4] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 5558–5570, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[5] Monica Riedler and Stefan Langer. Beyond text: Optimizing rag with multimodal inputs for industrial applications. **arXiv preprint arXiv:2410.21943**, 2024.

[6] Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 1081–1093, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[8] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning. In **NeurIPS**, 2022.

[9] Sedigheh Eslami and Gerard de Melo. Mitigate the Gap: Investigating Approaches for Improving Cross-Modal Alignment in CLIP. **arXiv preprint arXiv:2406.17639**, 2024.

[10] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. MultiModalQA: Complex Question Answering over Text, Tables and Images. In **International Conference on Learning Representations**, 2021.

[11] Abrar Fahim, Alex Murphy, and Alona Fyshe. It's Not a Modality Gap: Characterizing and Addressing the Contrastive Gap. **arXiv preprint arXiv:2405.18570**, 2024.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] OpenAI. GPT-4o System Card. **arXiv preprint arXiv:2410.21276**, 2024.

A　データセットの詳細

Multi Modal QA (MMQA) [10] データセットは，テキスト・画像・表モーダルのデータを外部知識とする QA データセットである．本研究では，このうちテキストを外部知識とする TextQ，画像を外部知識とする ImageQ を用いて実験を行った．

MMQA の元々のテストデータは正答と正しい外部知識が公開されていないため，評価実験に用いることができなかった．そのため訓練データの一部を分割し，テストデータとして扱った．結果として得られた，QA の内訳を表 4 に示す．

外部知識の検索には LanceDB[3] を使用した．テキストモーダルの外部知識は文単位で区切り，各文の埋め込みベクトルの平均との cos 類似度を検索スコアの算出に用いた．

表 4: MMQA データセットの QA の内訳

QA	訓練データ	検証データ	テストデータ
TextQ	6,736	748	721
ImageQ	1,889	210	230
合計	8,625	958	951

B　cos 類似度の統計量

MMQA 訓練データ中の質問文と正しい外部知識から計算される cos 類似度の平均と分散を表 5 に示す．モーダルに応じて，cos 類似度の尺度が大きく異なる．

表 5: MMQA 訓練データ中の質問文と正しい外部知識の cos 類似度の平均・分散

統計量	値
μ_{text}	0.83
μ_{image}	0.31
σ^2_{text}	0.004
σ^2_{image}	0.001

C　GPT-4o へのプロンプト

表 6 に RAG による回答生成時に使用したプロンプトを示す．w/o RAG の場合は質問文のみを与えて回答を生成させた．

D　生成文の人手評価

人手評価は，非英語ネイティブで十分な英語力を持つ著者以外の 1 人に依頼した．評価者には質問文・正答・各生成文のみを提示した．また各手法による生成文の順番は QA ごとにランダムにシャッフルした．

3) https://lancedb.github.io/lancedb/

表 6: RAG の回答生成時のプロンプト

system	You are a helpful assistant.
user	Answer the following question referencing the following texts or images. Question: {question} Reference text 1: {retrieved text 1} ⋮

E　エラー分析のサンプル

ImageQ のいくつかのサンプルについて提案手法を用いて検索した際の，正しい外部知識の画像内順位とテキストを含めた順位を表 7 に示す．6.3 節で述べたように，提案手法によって modality gap の影響を軽減できた ImageQ には，質問文に固有表現が含まれる傾向が見られた．質問文 “What body part ... As Real as It Gets” (全文は表 7 に記載) は固有表現 “As Real as It Gets” (楽曲名) を含むものの，一般的な単語から成ることから，CLIP による質問文と画像の紐づけが弱く働いたと考えられる．

表 7: 提案手法使用時の ImageQ の質問文と正しい外部知識の検索順位．質問文に含まれる固有表現を太字で示している．

質問文	画像内順位	全体順位
How many women are on the poster for **The Wolverine** (film)?	1	1
Are the colors of the American flag present in the logo of **Club Atlético River Plate Puerto Rico?**	1	1
How many colors are on **the Mississippi** flag?	1	2
What type of ball is in **the Liga Nacional de Fútbol de Puerto Rico** logo?	3	3
What is over the ears of the man on the left who is playing Hip hop music?	2	6,048
How many people are playing the board game on the table?	3	7,934
What body part is on the front cover of **As Real as It Gets**?	47	17,030