The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives

Mohit Iyyer*[1] Varun Manjunatha*[1] Anupam Guha[1] Yogarshi Vyas[1]
Jordan Boyd-Graber[2] Hal Daumé III[1] Larry Davis[1]
[1]University of Maryland, College Park [2]University of Colorado, Boulder
{miyyer,varunm,aguha,yogarshi,hal,lsd}@umiacs.umd.edu jordan.boyd.graber@colorado.edu

Abstract

*Visual narrative is often a combination of explicit information and judicious omissions, relying on the viewer to supply missing details. In comics, most movements in time and space are hidden in the "gutters" between panels. To follow the story, readers logically connect panels together by inferring unseen actions through a process called "**closure**". While computers can now describe what is explicitly depicted in natural images, in this paper we examine whether they can understand the closure-driven narratives conveyed by stylized artwork and dialogue in comic book panels. We construct a dataset,* COMICS*, that consists of over 1.2 million panels (120 GB) paired with automatic textbox transcriptions. An in-depth analysis of* COMICS *demonstrates that neither text nor image alone can tell a comic book story, so a computer must understand both modalities to keep up with the plot. We introduce three cloze-style tasks that ask models to predict narrative and character-centric aspects of a panel given* n *preceding panels as context. Various deep neural architectures underperform human baselines on these tasks, suggesting that* COMICS *contains fundamental challenges for both vision and language.*

1. Introduction

Comics are fragmented scenes forged into full-fledged stories by the imagination of their readers. A comics creator can condense anything from a centuries-long intergalactic war to an ordinary family dinner into a single panel. But it is what the creator *hides* from their pages that makes comics truly interesting: the unspoken conversations and unseen actions that lurk in the spaces (or gutters) between adjacent panels. For example, the dialogue in Figure 1 suggests that between the second and third panels, Gilda commands her snakes to chase after a frightened Michael in some sort of strange cult initiation. Through a process called *closure* [40], which involves (1) understanding individual panels and (2) making connective inferences across panels, readers form coherent storylines from seemingly disparate panels such as these. In this paper, we study whether computers can do the same by collecting a dataset of comic books (COMICS) and designing several tasks that require closure to solve.

Figure 1. Where did the snake in the last panel come from? Why is it biting the man? Is the man in the second panel the same as the man in the first panel? To answer these questions, readers form a larger meaning out of the narration boxes, speech bubbles, and artwork by applying closure across panels.

Section 2 describes how we create COMICS,[1] which contains ∼1.2 million panels drawn from almost 4,000 publicly-available comic books published during the "Golden Age" of American comics (1938–1954). COMICS is challenging in both style and content compared to natural images (e.g., photographs), which are the focus of most existing datasets and methods [32, 56, 55]. Much like painters, comic artists can render a single object or concept in multiple artistic styles to evoke different emotional responses from the reader. For example, the lions in Figure 2 are drawn with varying degrees of realism: the more cartoon-

* Authors contributed equally

[1]Data, code, and annotations to be made available after blind review.

Figure 2. Different artistic renderings of lions taken from the COMICS dataset. The left-facing lions are more cartoonish (and humorous) than the ones facing right, which come from action and adventure comics that rely on realism to provide thrills.

ish lions, from humorous comics, take on human expressions (e.g., surprise, nastiness), while those from adventure comics are more photorealistic.

Comics are not just visual: creators push their stories forward through text—speech balloons, thought clouds, and narrative boxes—which we identify and transcribe using optical character recognition (OCR). Together, text and image are often intricately woven together to tell a story that neither could tell on its own (Section 3). To understand a story, readers must connect dialogue and narration to characters and environments; furthermore, the text must be read in the proper order, as panels often depict long scenes rather than individual moments [10]. Text plays a much larger role in COMICS than it does for existing datasets of visual stories [25].

To test machines' ability to perform closure, we present three novel cloze-style tasks in Section 4 that require a deep understanding of narrative and character to solve. In Section 5, we design four neural architectures to examine the impact of multimodality and contextual understanding via closure. All of these models perform significantly worse than humans on our tasks; we conclude with an error analysis (Section 6) that suggests future avenues for improvement.

2. Creating a dataset of comic books

Comics, defined by cartoonist Will Eisner as *sequential art* [13], tell their stories in sequences of *panels*, or single frames that can contain both images and text. Existing comics datasets [19, 39] are too small to train data-hungry machine learning models for narrative understanding; additionally, they lack diversity in visual style and genres. Thus,

# Books	3,948
# Pages	198,657
# Panels	1,229,664
# Textboxes	2,498,657
Text cloze instances	89,412
Visual cloze instances	587,797
Char. coherence instances	72,313

Table 1. Statistics describing dataset size (top) and the number of total instances for each of our three tasks (bottom).

we build our own dataset, COMICS, by (1) downloading comics in the public domain, (2) segmenting each page into panels, (3) extracting textbox locations from panels, and (4) running OCR on textboxes and post-processing the output. Table 1 summarizes the contents of COMICS. The rest of this section describes each step of our data creation pipeline.

2.1. Where do our comics come from?

The "Golden Age of Comics" began during America's Great Depression and lasted through World War II, ending in the mid-1950s with the passage of strict censorship regulations. In contrast to the long, world-building story arcs popular in later eras, Golden Age comics tend to be small and self-contained; a single book usually contains multiple different stories sharing a common theme (e.g., crime or mystery). While the best-selling Golden Age comics tell of American superheroes triumphing over German and Japanese villains, a variety of other genres (such as romance, humor, and horror) also enjoyed popularity [18]. The Digital Comics Museum (DCM)[2] hosts user-uploaded scans of many comics by lesser-known Golden Age publishers that are now in the public domain due to copyright expiration. To avoid off-square images and missing pages, as the scans vary in resolution and quality, we download the 4,000 highest-rated comic books from DCM.[3]

2.2. Breaking comics into their basic elements

The DCM comics are distributed as compressed archives of JPEG page scans. To analyze closure, which occurs from panel-to-panel, we first extract panels from the page images. Next, we extract textboxes from the panels, as both location and content of textboxes are important for character and narrative understanding.

Panel segmentation: Previous work on panel segmentation uses heuristics [34] or algorithms such as density gradients and recursive cuts [52, 43, 48] that rely on pages with uniformly white backgrounds and clean gutters. Unfortunately, scanned images of eighty-year old comics do

[2] http://digitalcomicmuseum.com/

[3] Some of the panels in COMICS contain offensive caricatures and opinions reflective of that period in American history.

not particularly adhere to these standards; furthermore, many DCM comics have non-standard panel layouts and/or textboxes that extend across gutters to multiple panels.

After our attempts to use existing panel segmentation software failed, we turned to deep learning. We annotate 500 randomly-selected pages from our dataset with rectangular bounding boxes for panels. Each bounding box encloses both the panel artwork and the textboxes within the panel; in cases where a textbox spans multiple panels, we necessarily also include portions of the neighboring panel. After annotation, we train a region-based convolutional neural network to automatically detect panels. In particular, we use Faster R-CNN [45] initialized with a pretrained VGG_CNN_M_1024 model [9] and alternatingly optimize the region proposal network and the detection network. In Western comics, panels are usually read left-to-right, top-to-bottom, so we also have to properly order all of the panels within a page after extraction. We compute the midpoint of each panel and sort them using Morton order [41], which gives incorrect orderings only for rare and complicated panel layouts.

Textbox segmentation: Since we are particularly interested in modeling the interplay between text and artwork, we need to also convert the text in each panel to a machine-readable format.[4] As with panel segmentation, existing comic textbox detection algorithms [22, 47] could not accurately localize textboxes for our data. Thus, we resort again to Faster R-CNN: we annotate 1,500 panels for textboxes,[5] train a Faster-R-CNN, and sort the extracted textboxes within each panel using Morton order.

2.3. OCR

The final step of our data creation pipeline is applying OCR to the extracted textbox images. We unsuccessfully experimented with two trainable open-source OCR systems, Tesseract [50] and Ocular [6], as well as Abbyy's consumer-grade FineReader.[6] The ineffectiveness of these systems is likely due to the considerable variation in comic fonts as well as domain mismatches with pretrained language models (comics text is always capitalized, and dialogue phenomena such as dialects may not be adequately represented in training data). Google's Cloud Vision OCR[7] performs much better on comics than any other system we tried. While it sometimes struggles to detect short words or punctuation marks, the quality of the transcriptions is good considering the image domain and quality. We use the Cloud Vision API to run OCR on all 2.5 million textboxes for a cost of $3,000. We post-process the transcriptions by removing systematic spelling errors (e.g., failing to recognize the first letter of a word). Finally, each book in our dataset contains three or four full-page product advertisements; since they are irrelevant for our purposes, we train a classifier on the transcriptions to remove them.[8]

3. Data Analysis

In this section, we explore what makes understanding narratives in COMICS difficult, focusing specifically on *intrapanel* behavior (how images and text interact within a panel) and *interpanel* transitions (how the narrative advances from one panel to the next). We characterize panels and transitions using a modified version of the annotation scheme in Scott McCloud's "Understanding Comics" [40]. Over 90% of panels rely on both text and image to convey information, as opposed to just using a single modality. Closure is also important: to understand most transitions between panels, readers must make complex inferences that often require common sense (e.g., connecting jumps in space and/or time, recognizing when new characters have been introduced to an existing scene). We conclude that any model trained to understand narrative flow in COMICS will have to effectively tie together multimodal inputs through closure.

To perform our analysis, we manually annotate 250 randomly-selected pairs of consecutive panels from COMICS. Each panel of a pair is annotated for intrapanel behavior, while an interpanel annotation is assigned to the transition between the panels. Two annotators independently categorize each pair, and a third annotator makes the final decision when they disagree. We use four intrapanel categories (definitions from McCloud, percentages from our annotations):

1. **Word-specific, 4.4%**: The pictures illustrate, but do not significantly add to a largely complete text.
2. **Picture-specific, 2.8%**: The words do little more than add a soundtrack to a visually-told sequence.
3. **Parallel, 0.6%**: Words and pictures seem to follow very different courses without intersecting.
4. **Interdependent, 92.1%**: Words and pictures go hand-in-hand to convey an idea that neither could convey alone.

We group interpanel transitions into five categories:

1. **Moment-to-moment, 0.4%**: Almost no time passes between panels, much like adjacent frames in a video.
2. **Action-to-action, 34.6%**: The same subjects progress through an action within the same scene.

[4] Alternatively, modules for text spotting and recognition [27] could be built into architectures for our downstream tasks, but since comic dialogues can be quite lengthy, these modules would likely perform poorly.

[5] We make a distinction between *narration* and *dialogue*; the former usually occurs in strictly rectangular boxes at the top of each panel and contains text describing or introducing a new scene, while the latter is usually found in speech balloons or thought clouds.

[6] http://www.abbyy.com

[7] http://cloud.google.com/vision

[8] See supplementary material for specifics about our post-processing.

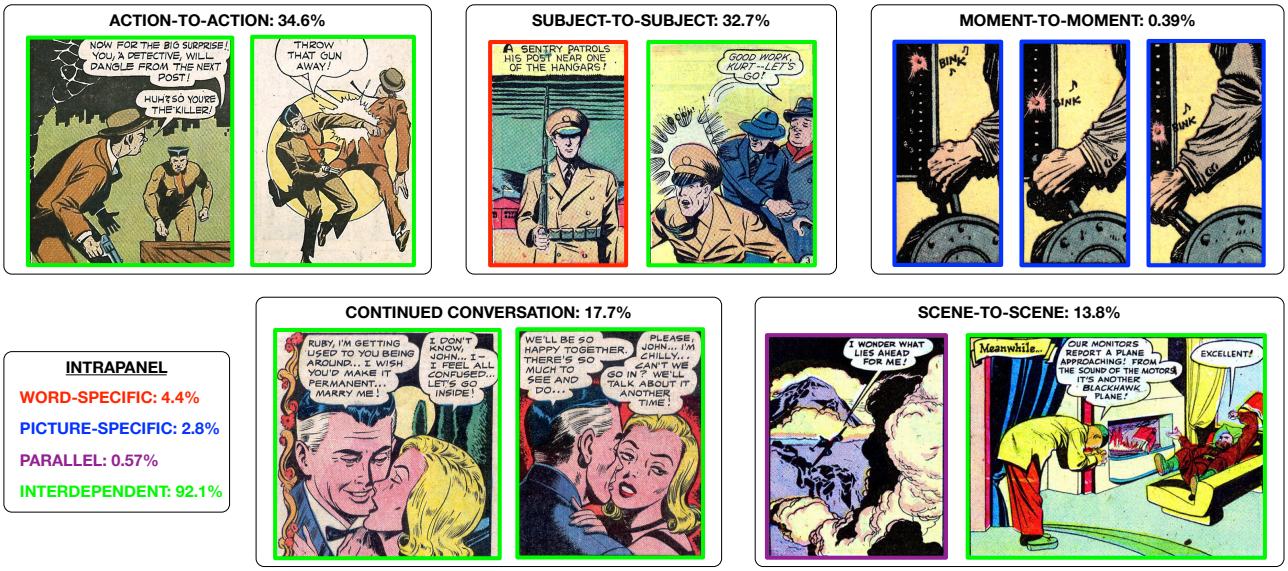

Figure 3. Five example panel sequences from COMICS, one for each type of interpanel transition. Individual panel borders are color-coded to match their intrapanel categories (legend in bottom-left). Moment-to-moment transitions unfold like frames in a movie, while scene-to-scene transitions are loosely strung together by narrative boxes. Percentages are the relative prevalance of the transition or panel type in an annotated subset of COMICS.

3. **Subject-to-subject, 32.7%**: New subjects are introduced while staying within the same scene or idea.
4. **Scene-to-scene, 13.8%**: Significant changes in time or space between the two panels.
5. **Continued conversation, 17.7%**: Subjects continue a conversation across panels without any other changes.

The two annotators agree on 96% of the intrapanel annotations (Cohen's $\kappa = 0.657$), which is unsurprising because almost every panel is interdependent. The interpanel task is significantly harder: agreement is only 68% (Cohen's $\kappa = 0.605$). Panel transitions are more diverse, as all types except moment-to-moment are relatively common (Figure 3); interestingly, moment-to-moment transitions require the least amount of closure as there is almost no change in time or space between the panels. Multiple transition types may occur in the same panel, such as simultaneous changes in subjects and actions, which also contributes to the lower interpanel agreement.

4. Tasks that test closure

To explore closure in COMICS, we design three novel tasks (*text cloze*, *visual cloze*, and *character coherence*) that test a model's ability to understand narratives and characters given a few panels of context. As shown in the previous section's analysis, a high percentage of panel transitions require non-trivial inferences from the reader; to successfully solve our proposed tasks, a model must be able to make the same kinds of connections.

While their objectives are different, all three tasks follow the same format: given preceding panels $p_{i-1}, p_{i-2}, \ldots, p_{i-n}$ as context, a model is asked to predict some aspect of panel p_i. While previous work on visual storytelling focuses on *generating* text given some context [24], the dialogue-heavy text in COMICS makes evaluation difficult (e.g., dialects, grammatical variations, many rare words). We want our evaluations to focus specifically on closure, not generated text quality, so we instead use a cloze-style framework [53]: given c candidates—with a single correct option—models must use the context panels to rank the correct candidate higher than the others. The rest of this section describes each of the three tasks in detail; Table 1 provides the total instances of each task with the number of context panels $n = 3$.

Text Cloze: In the *text cloze* task, we ask the model to predict what text out of a set of candidates belongs in a particular textbox, given both context panels (text and image) as well as the current panel image. While initially we did not put any constraints on the task design, we quickly noticed two major issues. First, since the panel images include textboxes, any model trained on this task could in principle learn to crudely imitate OCR by matching text candidates to the actual image of the text. To solve this problem, we "black out" the rectangle given by the bounding boxes for each textbox in a panel (see Figure 4).[9] Second, panels often have multiple textboxes (e.g., conversations between characters); to focus on interpanel transitions rather

[9] To reduce the chance of models trivially correlating candidate length to textbox size, we remove very short and very long candidates.

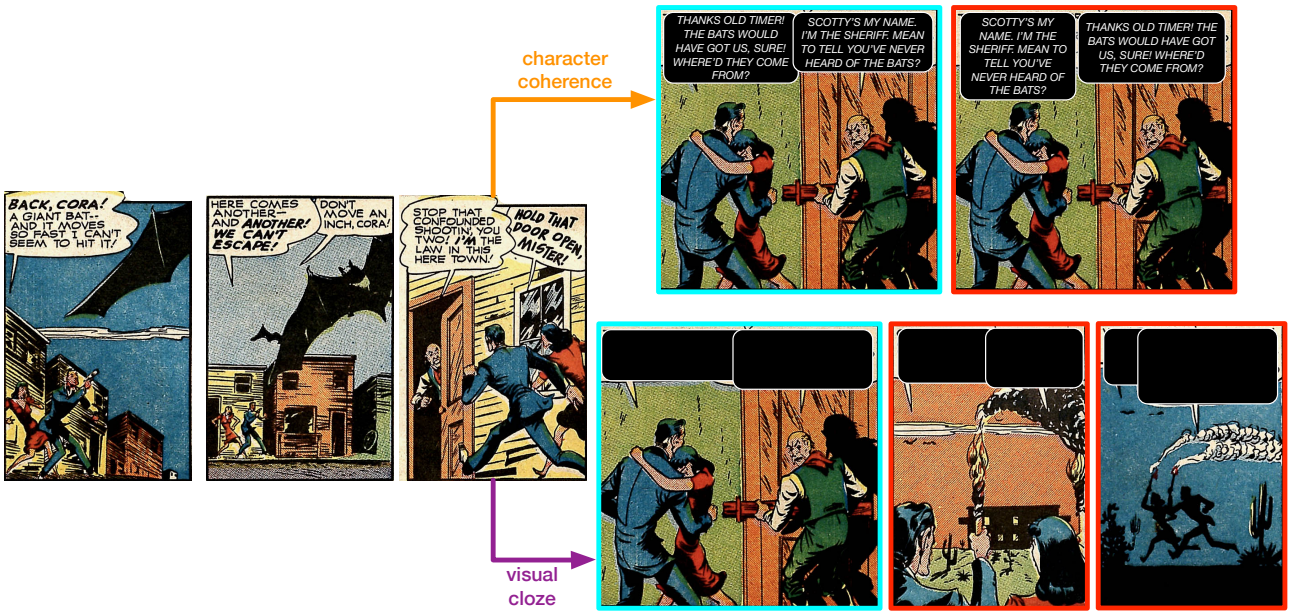

Figure 4. In the character coherence task (top), a model must order the dialogues in the final panel, while visual cloze (bottom) requires choosing the image of the panel that follows the given context. For visualization purposes, we show the original context panels; during model training and evaluation, textboxes are blacked out in every panel.

than intrapanel complexity, we restrict p_i to panels that contain only a single textbox. Thus, nothing from the current panel matters other than the artwork; the majority of the predictive information comes from previous panels.

Visual Cloze: We know from Section 3 that in most cases, text and image work interdependently to tell a story. In the *visual cloze* task, we follow the same set-up as in *text cloze*, but our candidates are images instead of text. A key difference is that models are not given text from the final panel; in *text cloze*, models are allowed to look at the final panel's artwork. This design is motivated by eyetracking studies in single-panel cartoons, which show that readers look at artwork before reading the text [7], although atypical font style and text length can invert this order [16].

Character Coherence: While the previous two tasks focus mainly on narrative structure, our third task attempts to isolate character understanding through a re-ordering task. Given a jumbled set of text from the textboxes in panel p_i, a model must learn to match each candidate to its corresponding textbox. We restrict this task to panels that contain exactly two dialogue boxes (narration boxes are excluded to focus the task on characters). While it is often easy to order the text based on the language alone (e.g., "how's it going" always comes before "fine, how about you?"), many cases require inferring which character is likely to utter a particular bit of dialogue based on both their previous utterances and their appearance (e.g., Figure 4, top).

4.1. Task Difficulty

For *text cloze* and *visual cloze*, we have two difficulty settings that vary in how cloze candidates are chosen. In the *easy* setting, we sample textboxes (or panel images) from the entire COMICS dataset at random. Most incorrect candidates in the easy setting have no relation to the provided context, as they come from completely different books and genres. This setting is thus easier for models to "cheat" on by relying on stylistic indicators instead of contextual information. With that said, the task is still non-trivial; for example, many bits of short dialogue can be applicable in a variety of scenarios. In the *hard* case, the candidates come from nearby pages, so models must rely on the context to perform well. For *text cloze*, all candidates are likely to mention the same character names and entities, while color schemes and textures become much less distinguishing for *visual cloze*.

5. Models & Experiments

To measure the difficulty of these tasks for deep learning models, we adapt strong baselines for multimodal language and vision understanding tasks to the comics domain. We evaluate four different neural models, variants of which were also used to benchmark the Visual Question Answering dataset [2] and encode context for visual storytelling [25]: *text-only*, *image-only*, and two *image-text* models. Our best-performing model encodes panels with a hierarchical LSTM architecture (see Figure 5).

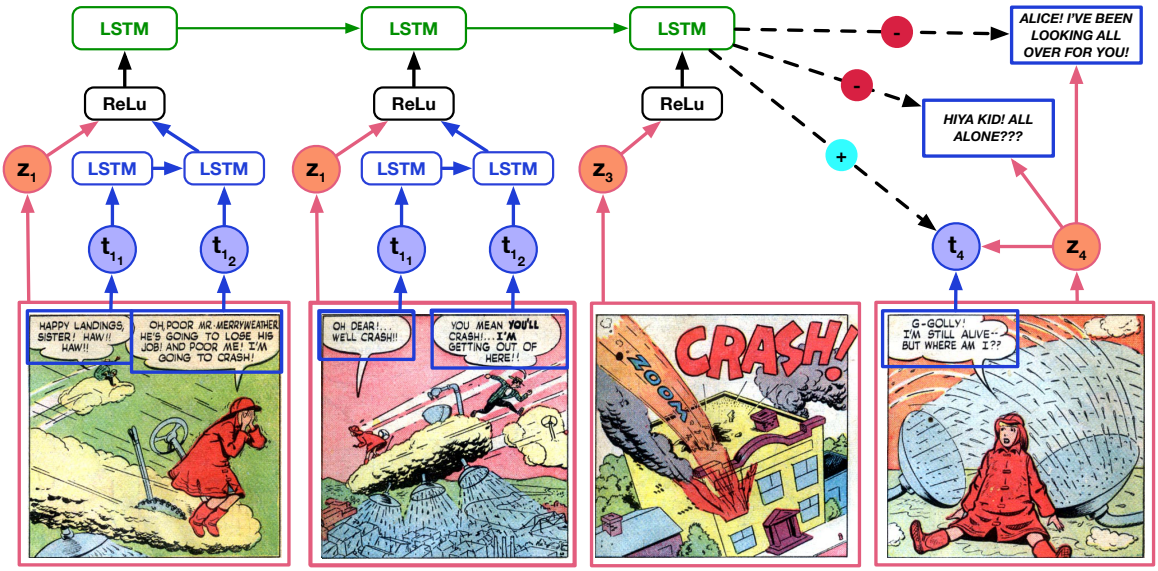

Figure 5. The **image-text** architecture applied to an instance of the *text cloze* task. Pretrained image features are combined with learned text features in a hierarchical LSTM architecture to form a context representation, which is then used to score text candidates.

On *text cloze*, accuracy increases when models are given images (in the form of pretrained VGG-16 features) in addition to text; on the other tasks, incorporating both modalities is less important. Additionally, for the *text cloze* and *visual cloze* tasks, models perform far worse on the *hard* setting than the *easy* setting, confirming our intuition that these tasks are non-trivial when we control for stylistic dissimilarities between candidates. Finally, none of the architectures outperform human baselines, which demonstrates the difficulty of understanding COMICS: image features obtained from models trained on natural images cannot capture the vast variation in artistic styles, and textual models struggle with the richness and ambiguity of colloquial dialogue highly dependent on visual contexts. In the rest of this section, we first introduce a shared notation and then use it to specify all of our models.

5.1. Model definitions

In all of our tasks, we are asked to make a prediction about a particular panel given the preceding n panels as context.[10] Each panel consists of three distinct elements: image, text (OCR output), and textbox bounding box coordinates. For any panel p_i, the corresponding image is z_i. Since there can be multiple textboxes per panel, we refer to individual textbox contents and bounding boxes as t_{i_x} and b_{i_x}, respectively. Each of our tasks has a different set of answer candidates A: *text cloze* has three text candidates $t_{a_{1\ldots3}}$, *visual cloze* has three image candidates $z_{a_{1\ldots3}}$, and *character coherence* has two combinations of text / bounding box pairs, $\{t_{a_1}/b_{a_1}, t_{a_2}/b_{a_2}\}$ and $\{t_{a_1}/b_{a_2}, t_{a_2}/b_{a_1}\}$. Our architectures differ mainly in the encoding function g that converts a sequence of context panels $p_{i-1}, p_{i-2}, \ldots, p_{i-n}$ into a fixed-length vector c. We score the answer candidates by taking their inner product with c and normalizing with the softmax function,

$$s = \text{softmax}(A^T c), \tag{1}$$

and we minimize the cross-entropy loss against the ground-truth labels.[11]

Text-only: The text-only baseline only has access to the text t_{i_x} within each panel. Our g function encodes this text on multiple levels: we first compute a representation for each t_{i_x} with a word embedding sum[12] and then combine multiple textboxes within the same panel using an *intra-panel* LSTM [23]. Finally, we feed the panel-level representations to an *interpanel* LSTM and take its final hidden state as the context representation (Figure 5). For *text cloze*, the answer candidates are also encoded with a word embedding sum; for *visual cloze*, we project the 4096-d `fc7` layer of VGG-16 down to the word embedding dimensionality with a fully-connected layer.[13]

[10] Test and validation instances for all tasks come from comic books that are unseen during training.

[11] Performance falters slightly on a development set with contrastive max-margin loss functions [51] in place of our softmax alternative.

[12] As in previous work for visual question answering [57], we observe no noticeable improvement with more sophisticated encoding architectures.

[13] For training and testing, we use three panels of context and three candidates. We use a vocabulary size of 30,000 words, restrict the maximum number of textboxes per panel to three, and set the dimensionality of word embeddings and LSTM hidden states to 256. Models are optimized using Adam [29] for ten epochs, after which we select the best-performing model on the dev set.

Model	Text Cloze		Visual Cloze		Char. Coheren.
	easy	*hard*	*easy*	*hard*	
Random	33.3	33.3	33.3	33.3	50.0
Text-only	63.4	52.9	55.9	48.4	68.2
Image-only	51.7	49.4	**85.7**	**63.2**	**70.9**
NC-image-text	63.1	59.6	-	-	65.2
Image-text	**68.6**	**61.0**	81.3	59.1	69.3
Human	–	84	–	88	87

Table 2. Combining image and text in neural architectures improves their ability to predict the next image or dialogue in COMICS narratives. The contextual information present in preceding panels is useful for all tasks: the model that only looks at a single panel (**NC-image-text**) always underperforms its context-aware counterpart. However, even the best performing models lag well behind humans.

Image-only: The image-only baseline is even simpler: we feed the `fc7` features of each context panel to an LSTM and use the same objective function as before to score candidates. For *visual cloze*, we project both the context and answer representations to 512-d with additional fully-connected layers before scoring. While the COMICS dataset is certainly large, we do not attempt learning visual features from scratch as our task-specific signals are far more complicated than simple image classification. We also try fine-tuning the lower-level layers of VGG-16 [4]; however, this substantially lowers task accuracy even with very small learning rates for the fine-tuned layers.

Image-text: We combine the previous two models by concatenating the output of the intrapanel LSTM with the `fc7` representation of the image and passing the result through a fully-connected layer before feeding it to the interpanel LSTM (Figure 5). For *text cloze* and *character coherence*, we also experiment with a variant of the image-text baseline that has no access to the context panels, which we dub **NC-image-text**. In this model, the scoring function computes inner products between the image features of p_i and the text candidates.[14]

6. Error Analysis

Table 2 contains our full experimental results, which we briefly summarize here. On *text cloze*, the image-text model dominates those trained on a single modality. However, text is much less helpful for *visual cloze* than it is for *text cloze*, suggesting that visual similarity dominates the former task. Having the context of the preceding panels helps across the board, although the improvements are lower in the *hard* setting. There is more variation across the models in the *easy* setting; we hypothesize that the *hard* case requires moving away from pretrained image features, and transfer learning methods may prove effective here. Differences between models on *character coherence* are minor; we suspect that more complicated attentional architectures that leverage the bounding box locations b_{i_x} are necessary to "follow" speech bubble tails to the characters who speak them.

We also compare all models to a human baseline, for which the authors manually solve one hundred instances of each task (in the *hard* setting) given the same preprocessed input that is fed to the neural architectures. Most human errors are the result of poor OCR quality (e.g., misspelled words) or low image resolution. Humans comfortably outperform all models, making it worthwhile to look at where computers fail but humans succeed.

The top row in Figure 6 demonstrates an instance (from *easy text cloze* where the image helps the model make the correct prediction. The text-only model has no idea that an airplane (referred to here as a "ship") is present in the panel sequence, as the dialogue in the context panels make no mention of it. In contrast, the image-text model is able to use the artwork to rule out the two incorrect candidates.

The bottom two rows in Figure 6 show *hard text cloze* instances in which the image-text model is deceived by the artwork in the final panel. While the final panel of the middle row does contain what looks to be a creek, "catfish creek jail" is more suited for a narrative box than a speech bubble, while the meaning of the correct candidate is obscured by the dialect and out-of-vocabulary token. Similarly, a camera films a fight scene in the last row; the model selects a candidate that describes a fight instead of focusing on the context in which the scene occurs. These examples suggest that the contextual information is overridden by strong associations between text and image, motivating architectures that go beyond similarity by leveraging external world knowledge to determine whether an utterance is truly appropriate in a given situation.

7. Related Work

Our work is related to three main areas: (1) multimodal tasks that require language and vision understanding, (2) computational methods that focus on non-natural images, and (3) models that characterize language-based narratives.

Deep learning has renewed interest in jointly reasoning about vision and language. Datasets such as MS COCO [35] and Visual Genome [31] have enabled image captioning [54, 28, 56] and visual question answering [37, 36]. Similar to our *character coherence* task, researchers have built models that match TV show characters with their visual attributes [15] and speech patterns [21].

Closest to our own comic book setting is the visual storytelling task, in which systems must generate [24] or reorder [1] stories given a dataset (SIND) of photos from

[14] We cannot apply this model to *visual cloze* because we are not allowed access to the artwork in panel p_i.

Figure 6. Three *text cloze* examples from the development set, shown with a single panel of context (boxed candidates are predictions by the **text-image** model). The airplane artwork in the top row helps the **image-text** model choose the correct answer, while the **text-only** model fails because the dialogue lacks contextual information. Conversely, the bottom two rows show the **image-text** model ignoring the context in favor of choosing a candidate that mentions something visually present in the last panel.

Flikr galleries of "storyable" events such as weddings and birthday parties. SIND's images are fundamentally different from COMICS in that they lack coherent characters and accompanying dialogue. Comics are created by skilled professionals, not crowdsourced workers, and they offer a far greater variety of character-centric stories that depend on dialogue to further the narrative; with that said, the text in COMICS is less suited for generation because of OCR errors.

We build here on previous work that attempts to understand non-natural images. Zitnick et al. [58] discover semantic scene properties from a clip art dataset featuring characters and objects in a limited variety of settings. Applications of deep learning to paintings include tasks such as detecting objects in oil paintings [11, 12] and answering questions about artwork [20]. Previous computational work on comics focuses primarily on extracting elements such as panels and textboxes [46]; in addition to the references in Section 2, there is a large body of segmentation research on manga [3, 44, 38, 30].

To the best of our knowledge, we are the first to computationally model *content* in comic books as opposed to just extracting their elements. We follow previous work in language-based narrative understanding; very similar to our *text cloze* task is the "Story Cloze Test" [42], in which models must predict the ending to a short (four sentences long) story. Just like our tasks, the Story Cloze Test proves difficult for computers and motivates future research into commonsense knowledge acquisition. Others have studied characters [14, 5, 26] and narrative structure [49, 33, 8] in novels.

8. Conclusion & Future Work

We present the COMICS dataset, which contains over 1.2 million panels from "Golden Age" comic books. We design three cloze-style tasks on COMICS to explore *closure*, or how readers connect disparate panels into coherent stories. Experiments with different neural architectures, along with a manual data analysis, confirm the importance of multimodal models that combine text and image for comics understanding. We additionally show that context is crucial for predicting narrative or character-centric aspects of panels.

However, for computers to reach human performance, they will need to become better at leveraging context. Readers rely on commonsense knowledge to make sense of dramatic scene and camera changes; how can we inject such knowledge into our models? Another potentially intriguing direction, especially given recent advances in generative adversarial networks [17], is generating artwork given dialogue (or vice versa). Finally, COMICS presents a golden opportunity for transfer learning; can we train models that generalize across natural and non-natural images much like humans do?

9. Acknowledgments

We thank the anonymous reviewers for their insightful comments and the UMIACS and Google Cloud Support staff for their help with OCR. Manjunatha and Davis were supported by Office of Naval Research grant N000141612713, while Iyyer, Boyd-Graber, and Daumé were supported by NSF grant IIS-1320538. Any opinions, findings, or conclusions expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

References

[1] H. Agrawal, A. Chandrasekaran, D. Batra, D. Parikh, and M. Bansal. Sort story: Sorting jumbled images and captions into stories. In *Proceedings of Empirical Methods in Natural Language Processing*, 2016. 7

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision*, 2015. 5

[3] Y. Aramaki, Y. Matsui, T. Yamasaki, and K. Aizawa. Interactive segmentation for manga. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference*, 2014. 8

[4] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba. Cross-modal scene networks. *arXiv*, 2016. 7

[5] D. Bamman, T. Underwood, and N. A. Smith. A Bayesian mixed effects model of literary character. In *Proceedings of the Association for Computational Linguistics*, 2014. 8

[6] T. Berg-Kirkpatrick, G. Durrett, and D. Klein. Unsupervised transcription of historical documents. In *Proceedings of the Association for Computational Linguistics*, 2013. 3

[7] P. J. Carroll, J. R. Young, and M. S. Guertin. Visual analysis of cartoons: A view from the far side. In *Eye movements and visual cognition*. Springer, 1992. 5

[8] N. Chambers and D. Jurafsky. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Association for Computational Linguistics*, 2009. 8

[9] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014. 3

[10] N. Cohn. The limits of time and transitions: challenges to theories of sequential image comprehension. *Studies in Comics*, 1(1), 2010. 2

[11] E. Crowley and A. Zisserman. The state of the art: Object retrieval in paintings using discriminative regions. In *British Machine Vision Conference*, 2014. 8

[12] E. J. Crowley, O. M. Parkhi, and A. Zisserman. Face painting: querying art with photos. In *British Machine Vision Conference*, 2015. 8

[13] W. Eisner. *Comics & Sequential Art*. Poorhouse Press, 1990. 2

[14] D. K. Elson, N. Dames, and K. R. McKeown. Extracting social networks from literary fiction. In *Proceedings of the Association for Computational Linguistics*, 2010. 8

[15] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... Buffy" – automatic naming of characters in TV video. In *Proceedings of the British Machine Vision Conference*, 2006. 7

[16] T. Foulsham, D. Wybrow, and N. Cohn. Reading without words: Eye movements in the comprehension of comic strips. *Applied Cognitive Psychology*, 30, 2016. 5

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of Advances in Neural Information Processing Systems*, 2014. 8

[18] R. Goulart. *Comic Book Encyclopedia: The Ultimate Guide to Characters, Graphic Novels, Writers, and Artists in the Comic Book Universe*. HarperCollins, 2004. 2

[19] C. Guérin, C. Rigaud, A. Mercier, F. Ammar-Boudjelal, K. Bertet, A. Bouju, J.-C. Burie, G. Louis, J.-M. Ogier, and A. Revel. eBDtheque: a representative database of comics. In *International Conference on Document Analysis and Recognition*, 2013. 2

[20] A. Guha, M. Iyyer, and J. Boyd-Graber. A distorted skull lies in the bottom center: Identifying paintings from text descriptions. In *NAACL Human-Computer Question Answering Workshop*, 2016. 8

[21] M. Haurilet, M. Tapaswi, Z. Al-Halah, and R. Stiefelhagen. Naming TV characters by watching and analyzing dialogs. In *IEEE Winter Conference on Applications of Computer Vision*, 2016. 7

[22] A. K. N. Ho, J.-C. Burie, and J.-M. Ogier. Panel and speech balloon extraction from comic books. In *IAPR International Workshop on Document Analysis Systems*, 2012. 3

[23] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997. 6

[24] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, et al. Visual storytelling. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2016. 4, 7

[25] T. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. B. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. Visual storytelling. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2016. 2, 5

[26] M. Iyyer, A. Guha, S. Chaturvedi, J. Boyd-Graber, and H. Daumé III. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2016. 8

[27] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1), 2016. 3

[28] A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015. 7

[29] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2014. 6

[30] S. Kovanen and K. Aizawa. A layered method for determining manga text bubble reading order. In *International Conference on Image Processing*, 2015. 8

[31] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 7

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In

Proceedings of Advances in Neural Information Processing Systems, 2012. 1

[33] W. G. Lehnert. Plot units and narrative summarization. *Cognitive Science*, 5(4), 1981. 8

[34] L. Li, Y. Wang, Z. Tang, and L. Gao. Automatic comic page segmentation based on polygon detection. *Multimedia Tools and Applications*, 69(1), 2014. 2

[35] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. 2014. 7

[36] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering, 2016. 7

[37] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Computer Vision and Pattern Recognition*, 2015. 7

[38] Y. Matsui. Challenge for manga processing: Sketch-based manga retrieval. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, 2015. 8

[39] Y. Matsui, K. Ito, Y. Aramaki, T. Yamasaki, and K. Aizawa. Sketch-based manga retrieval using manga109 dataset. *arXiv preprint arXiv:1510.04389*, 2015. 2

[40] S. McCloud. *Understanding Comics*. HarperCollins, 1994. 1, 3

[41] G. M. Morton. *A computer oriented geodetic data base and a new technique in file sequencing*. International Business Machines Co, 1966. 3

[42] N. Mostafazadeh, N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2016. 8

[43] X. Pang, Y. Cao, R. W. Lau, and A. B. Chan. A robust panel extraction method for manga. In *Proceedings of the ACM International Conference on Multimedia*, 2014. 2

[44] X. Pang, Y. Cao, R. W. H. Lau, and A. B. Chan. A robust panel extraction method for manga. In *Proceedings of the ACM International Conference on Multimedia*, 2014. 8

[45] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of Advances in Neural Information Processing Systems*, 2015. 3

[46] C. Rigaud. *Segmentation and indexation of complex objects in comic book images*. PhD thesis, University of La Rochelle, France, 2014. 8

[47] C. Rigaud, J.-C. Burie, J.-M. Ogier, D. Karatzas, and J. Van de Weijer. An active contour model for speech balloon detection in comics. In *International Conference on Document Analysis and Recognition*, 2013. 3

[48] C. Rigaud, C. Guérin, D. Karatzas, J.-C. Burie, and J.-M. Ogier. Knowledge-driven understanding of images in comic books. *International Journal on Document Analysis and Recognition*, 18(3), 2015. 2

[49] R. Schank and R. Abelson. *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. L. Erlbaum, 1977. 8

[50] R. Smith. An overview of the tesseract ocr engine. In *International Conference on Document Analysis and Recognition*, 2007. 3

[51] R. Socher, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2014. 6

[52] T. Tanaka, K. Shoji, F. Toyama, and J. Miyamichi. Layout analysis of tree-structured scene frames in comic images. In *International Joint Conference on Artificial Intelligence*, 2007. 2

[53] W. L. Taylor. Cloze procedure: a new tool for measuring readability. *Journalism and Mass Communication Quarterly*, 30(4), 1953. 4

[54] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*, 2015. 7

[55] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *Proceedings of the International Conference of Machine Learning*, 2016. 1

[56] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference of Machine Learning*, 2015. 1, 7

[57] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*, 2015. 6

[58] C. L. Zitnick, R. Vedantam, and D. Parikh. Adopting abstract images for semantic scene understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(4):627–638, 2016. 8