VKG-QA: Visual Knowledge Graph-based Question Answer for Multimodal Large Language Models

Supplementary Material

A. Details of Benchmark Construction

A.1. Subgraph Sampling

We first construct a directed KG from the raw datasets (i.e., the selected KG and KGQA corpora), yielding a global graph in which entities serve as nodes and relations form directed edges. In this stage, we perform no preprocessing or filtering operations, ensuring that the topology of the global KG remains fully consistent with the original datasets.

Based on this global KG, we extract controlled local k-hop neighborhood subgraphs centered on a designated entity. This process expands in both directions following a BFS-style layering scheme, collecting predecessors and successors at each step. To constrain structural complexity, we impose strict upper bounds on the number of neighbors sampled at each hop, thereby regulating the size of the local subgraph. When a node has more neighbors than the predefined limit in the current hop, a random subset is selected to avoid any manual selection bias. During sampling, we also record the hop distance of every node, which is subsequently used for color encoding in the visualization stage.

A.2. Edge Pruning for Enhanced Visual Clarity

In the Graph-specific Comprehension tasks, to enhance the readability of VKG images while preserving their structural complexity, we propose an edge refinement strategy guided by visual clarity constraints. For multi-relational edges, to mitigate visual ambiguity caused by edge overlaps, appropriate curvature parameters are assigned to differentiate relations while maintaining layout aesthetics.

To address the problem of high-degree nodes, we further introduce a degree-aware edge refinement strategy comprising two stages. In the first stage, high-degree nodes are statistically identified, and redundant edges are pruned from the subgraphs. During this process, node degrees are dynamically updated after each pruning iteration to avoid excessive local edge removal and preserve global connectivity. In the second stage, to maintain the structural centrality of the central node, high-degree nodes are pruned again while excluding the central node itself. This strategy effectively enhances visual clarity while retaining topological integrity and relational diversity of the subgraphs.

A.3. Question Answer generator

Each refined subgraph is then passed to a dedicated analysis module (except for images used in reasoning tasks, where we instead rely directly on QA pairs obtained from the KGQA dataset and manual annotation). This module automatically generates template questions and their corresponding ground truth. Specifically, for basic visual categories, it leverages the visual information assigned to nodes and edges during sampling to automatically construct questions and answers; for graph-specific comprehension categories, it directly computes various graph-theoretic properties of the subgraph. The output of this analysis module provides a deterministic and fully verifiable standard answer for each question–image pair in our dataset.

A.4. Details of the Visualization

To generate high-quality and interpretable images from the sampled subgraphs, we implement a customized visualization module based on pyvis and introduce several extensions tailored to the needs of the benchmark. Here, we supplement the implementation details that are not fully presented in the main paper.

In the node-rendering stage, all nodes are represented as fixed circular shapes with standardized font settings and colored according to their hop distance from the center. The color encodes only the distance level and does not reflect entity types, relation categories, or any semantic attributes, thereby avoiding implicit cues that could bias model evaluation. The node size, label position, and border style are uniformly set to "70×70 px" across all subgraphs to ensure visual consistency. In practice, however, nodes with long text labels may appear visually larger, and such cases are allowed in our benchmark.

For edge rendering, we extend the default pyvis behavior by handling three additional cases. First, when multiple relations exist between the same pair of nodes, each relation is displayed as an independent labeled edge, ensuring that all relation types are explicitly visible. Second, for bidirectional edges (i.e., when both $h \rightarrow t$ and $t \rightarrow h$ exist), we draw paired curved paths to separate the two directions, improving the readability of arrow orientation and avoiding overlap. Third, self-loops are removed since they rarely contribute meaningful information and typically reduce visual clarity.

During layout generation, the physics engine is disabled by default. Annotators manually adjust node positions and edge layouts when necessary to reduce visual clutter, avoid label overlap, and improve overall readability, without altering the original graph structure.

Due to the characteristics of pyvis and the manual adjustment workflow, the initial output consists of a set of HTML files. We embed a lightweight JavaScript snippet into each

file, providing a "Save as PNG" button that exports the rendered graph as a high-resolution PNG image with a pure white background. This ensures consistent resolution and style for all evaluation images after manual refinement.

This visualization pipeline preserves the original KG topology while producing high-definition VKG images, thereby offering a stable, fair, and interpretable foundation for multimodal reasoning evaluation.

B. Details of the Prompts

In our VLM evaluation, to ensure the model consistently produces parsable and alignable standardized outputs during reasoning, we designed dedicated prompts for tasks requiring structured outputs, including **triple extraction, color list extraction, and node text extraction**. For these tasks, we additionally required the model to generate outputs strictly adhering to a JSON-style format, preventing extraneous text, explanatory content, or any deviations from the ground-truth format.

For other tasks with simple answers that do not require structured formatting, such as counting, graph-structure QA, reasoning tasks and so on, we employed concise instructions, instructing the model to return only the final answers in JSON format.

All prompts used for evaluation are summarized in the accompanying Table 3.

Triple Extraction
Q: {....The image shows a directed knowledge graph. Please identify and extract all the triples from the graph in the image.} Return the extracted triples as a JSON-style list of lists. Each triple as [head, relation, tail]. Do not include any extra text or explanation. Example: [[A, related to, B], [B, connected to, C]].
Color List Extraction
Q: {....List all the colors used for the nodes in the given knowledge graph image. Only use basic color terms when responding, separated by commas, and sort them in alphabetical order. Do not include any extra words or symbols.} Return the color names strictly as a JSON-style list using double quotes. No extra text or explanation. Example: [red, blue, green].
Text Extraction
Q: {....Extract all text appearing on the nodes in the given image. Return only the node names, separated by commas, and sort them in alphabetical order. Do not include any extra words or symbols.} Return the answer strictly as a JSON-style list of strings using double quotes. No extra words or explanation. Example: [Cassius Clay, Muhammad Ali, Sonny Liston].
General Questions
Q: {....Based only on the given image, the films that the filmmakers share with the film Xanadu are written by whom?} Return the answer strictly as JSON. No explanation.

Table 3. Prompts used for evaluating VLMs across different task categories. List structured tasks require strict JSON-style output constraints, while other tasks use a unified and minimal JSON-only instruction.

C. Analysis of Knowledge Dependence.

To investigate the intrinsic mechanisms that VLMs rely on when addressing complex visual reasoning tasks, such as whether they depend on explicit knowledge during graph-based reasoning, we designed a set of isomorphic yet semantically varied QA pairs within specific sub-tasks of the Graph-Specific Comprehension task (including degree analysis, relation direction identification, cycle detection, and connectivity assessment).

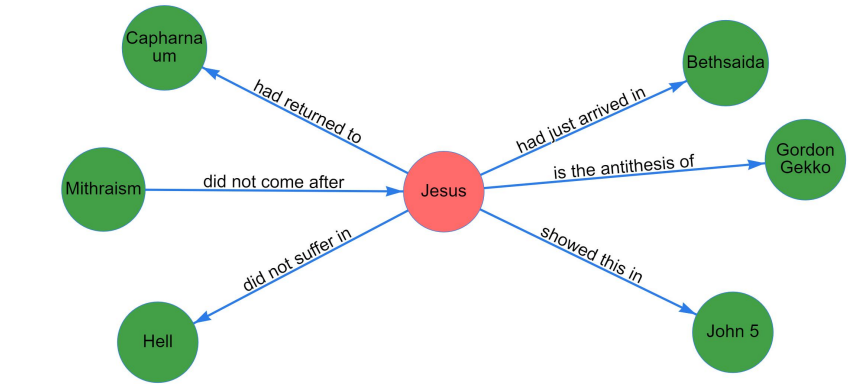
For the explicit knowledge condition (k1), models are required to perform reasoning based on abstract relational concepts. We directly introduce high-level semantic notions, with questions explicitly involving terms such as "cycle," "degree," or "strong connectivity." This setup evaluates the models' ability to align abstract linguistic concepts with concrete visual patterns.

For the implicit knowledge condition (k0), the questions require purely visual-structural reasoning. All concept, or domain, specific terminology is removed, compelling the models to infer the same properties solely from visual cues such as arrows, node connections, and directions. This directly tests capabilities including object localization, instruction execution, and holistic exploration and understanding of the graphical spatial layout.

We studied five closed-source models (GPT-5, Gemini 2.5-pro, GPT4o, QwenVLMax, QwenVLPlus) and four open-source models (Qwen2.5-vl-72b-instruct, Qwen2.5-vl-32b-instruct). Qwen 2.5-vL-7b-instruct, GLM-4.5V). The experimental results are presented in Table 4. The performance difference, denoted as $\Delta(k0 - k1)$, clearly delineates the models' strategic preferences. The results indicate significant divergence in model performance across the two question paradigms, which may reflect the varying degrees to which different MLLMs implicitly rely on conceptual cues versus visual patterns. Some top-performing models, such as Gemini-2.5-Pro (+2.46) and QwenVLMax (+7.13), perform better on tasks without knowledge prompts, demonstrating strong perceptual reasoning that directly leverages visual details like nodes and arrow directions, whereas introducing conceptual knowledge may introduce unreliable or inefficient cues. In contrast, knowledge-driven models such

Table 4. Knowledge prompt comparison results.

Knowledge	GPT-5	Gemini-2.5-pro	GPT4o	QwenVLMax	QwenVLPlus	Qwen2.5-vl-72b-instruct	Qwen2.5-vl-32b-instruct	Qwen2.5-vl-7b-instruct	GLM-4.5V
0	86.3	82.96	60.62	63.09	45.93	60	49.26	37.41	49.75
1	85.07	80.5	62.11	55.96	43.79	60.23	50.06	44.42	56.46
$\Delta(k0-k1)$	+1.23	+2.46	-1.49	+7.13	+2.14	-0.23	-0.8	-7.01	-6.71



Question : Extract all text appearing on the nodes in the given image.Return only the node names, separated by commas, and sort them in alphabetical order. Do not include any extra words or symbols.

Ground Truth : ['Bethsaida', 'Capharnaum', 'Gordon Gekko', 'Hell', 'Jesus', 'John 5', 'Mithraism']

Model Answer :

Reasoning:

- step 1: Identify all the nodes in the image.
- step 2: Extract the text appearing on each node.
- step 3: Sort the extracted node names in alphabetical order.

Final Answer: Bethsaida, Capharnaum, Gordon Gekko, Hell, Jesus, John 5, Mithraism ✔

Figure 9. Case study example-1 of VKG-QA on GPT-4.

as GLM-4.5V (-6.71) and Qwen2.5-7B (-7.01) benefit more from explicit knowledge cues and rely heavily on linguistic priors to interpret abstract relational concepts.

Notably, within the Qwen series, as model scale increases, the performance difference between the two question paradigms gradually diminishes, indicating decreasing sensitivity to prompt type and increasing stability. This trend further reflects that smaller models tend to rely on a single source of information, being more sensitive to either visual cues or explicit knowledge prompts, whereas larger models can efficiently extract graph-structural information from visuals while simultaneously leveraging knowledge cues and flexibly attending to the most informative source.

Additionally, we observed an interesting phenomenon: most closed-source models perform better under visual-only prompts, whereas open-source models tend to rely more on explicit knowledge cues. This phenomenon may stem from differences in training strategies and data distributions. Closed-source large models are typically trained on large-scale, diverse multimodal datasets with strong visual encoders and high instruction compliance, whereas open-source models, constrained by resources, tend to depend on linguistic prompts and instruction fine-tuning to accomplish reasoning tasks. This suggests that a model's reliance on visual versus knowledge information is determined not only by scale but also by its training methodology and data distribution.

D. Disentangling OCR and Graph Reasoning

Input	Model	Deg.	Dire.	Cyc.	Conn.	1-hop	Sup.	Avg.
Image-only	GLM-4.5V	57.6	70.2	97.3	31.6	94.8	68.0	64.1
	InternVL3-78B	46.0	42.9	89.3	50.9	91.4	50.0	56.6
	Qwen2.5-VL-7B	34.0	48.8	94.7	29.2	91.4	64.0	52.0
Image + Text	GLM-4.5V	67.0	84.5	96.0	37.5	96.6	88.0	71.3
	InternVL3-78B	54.7	79.7	89.3	44.6	100	82.0	68.7
Text-only	GLM-4.5-Air	72.6	79.8	94.7	53.0	87.9	86.0	74.6
	Qwen2.5-72B	43.4	92.9	97.3	58.3	93.1	86.0	71.1
	Qwen2.5-7B-Inst	24.5	73.8	96.0	55.4	93.1	78.0	60.5

Table 5. Performance across different input modalities.

To rigorously separate basic Optical Character Recognition (OCR) capabilities from genuine graph reasoning, we conducted a controlled experiment utilizing "OCR-perfect" inputs. In this setting, models were provided with both the original VKG images and their corresponding ground-truth textual triples. The experimental results are presented in

Table 5. Based on these findings, we derive two crucial insights:

1)The Reality of the Perceptual Bottleneck (Image + Text > Image Only): Compared to the image-only baseline, providing the ground-truth textual triples yields a slight performance improvement. This confirms that perceptual failure (e.g., accurately grounding text to specific nodes or identifying edges) is indeed a significant obstacle for current LMMs. This observation is highly consistent with our previous error analysis demonstrating that perceptual errors dominate the failure cases.

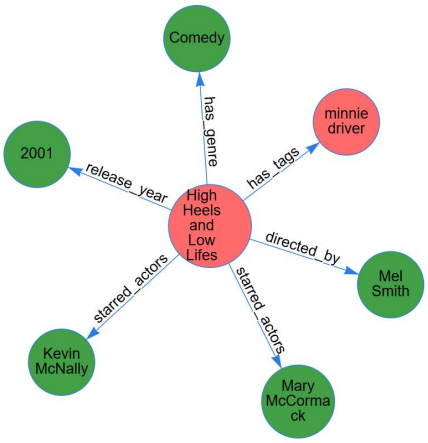
2)The Phenomenon of Visual Interference (Text Only > Image + Text): Intuitively, providing perfect text prompts should improve the performance of LMMs, enabling them to match pure text-based LLMs. Conversely, we observe a distinct performance drop when comparing the "Image + Text" setting to the "Text Only" setting. This decline indicates that the presence of the visual graph introduces complexities related to multimodal interference and structural alignment. Rather than simply reading text, the models are forced to execute a pixel-to-structure translation, attempting to establish a consistent mapping between the provided textual triples and the complex visual node layouts. This phenomenon provides strong, empirical evidence that VKG-QA evaluates multifaceted multimodal graph understanding, extending far beyond simple OCR tasks.

E. Additional Case Study Examples

To provide a more comprehensive view of model behaviors across different reasoning types, we include an extended set of case study examples in this section. For clarity and conciseness, we present only the input VKG images, corresponding questions, model predictions, and ground truth answers, as shown in Figure 9-18. The purpose of this section is to enable readers to directly inspect the representative successes and failures of different models, facilitating further investigation and reproducibility.

F. Specific Examples of the Benchmark

In this chapter, we present more clear image data, as shown in Figure 19-27. These examples are randomly selected from our benchmarks, and the corresponding partial Q&A pair matches are shown in Table 6.



Question : Based only on the given image, what is **High Heels and Low Lifes** about?

Ground Truth : minnie driver

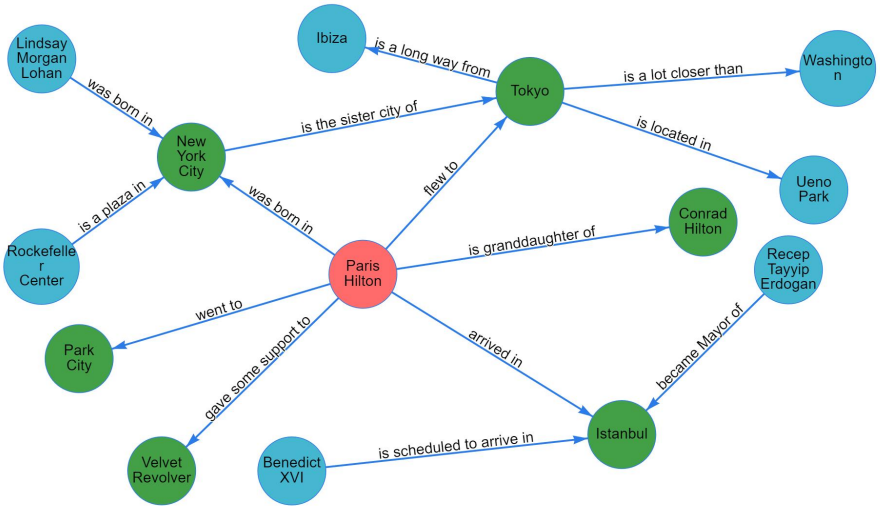
Model Answer :

Reasoning:

- step 1: Identify the genre of High Heels and Low Lifes from the image, which is labeled as Comedy.
- step 2: Note the release year of the film, which is 2001, as indicated in the image.
- step 3: Recognize the actors associated with the film, including Minnie Driver, Kevin McNally, and Mary McCormack, as well as the director, Mel Smith, from the image.

Final Answer: Based only on the given image, High Heels and Low Lifes is a comedy film released in 2001, featuring actors **Minnie Driver**, Kevin McNally, and Mary McCormack, and directed by Mel Smith. ✔

Figure 10. Case study example-2 of VKG-QA on GPT-4.



Question : Return the name of the node that is at the lower left of point 'Paris Hilton' and is blue. Only provide the node name. Do not include any extra words or symbols.

Ground Truth : Benedict XVI

Model Answer :

Reasoning:

- step 1: Identify the position of the node **'Paris Hilton'** in the image.
- step 2: Determine the nodes that are located to the lower left of 'Paris Hilton'.
- step 3: Identify the color of these nodes and select the one that is blue.

Final Answer: Velvet Revolver ✖

Figure 11. Case study example-3 of VKG-QA on GPT-4.

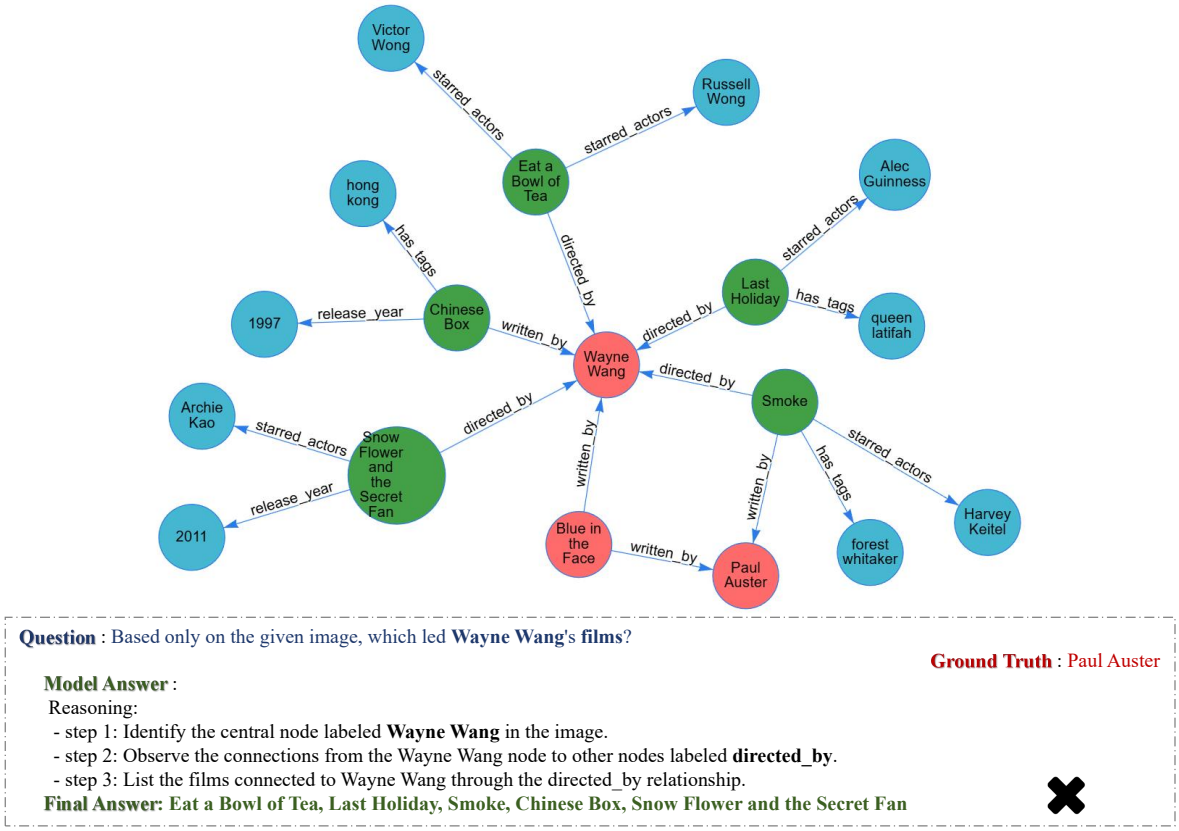
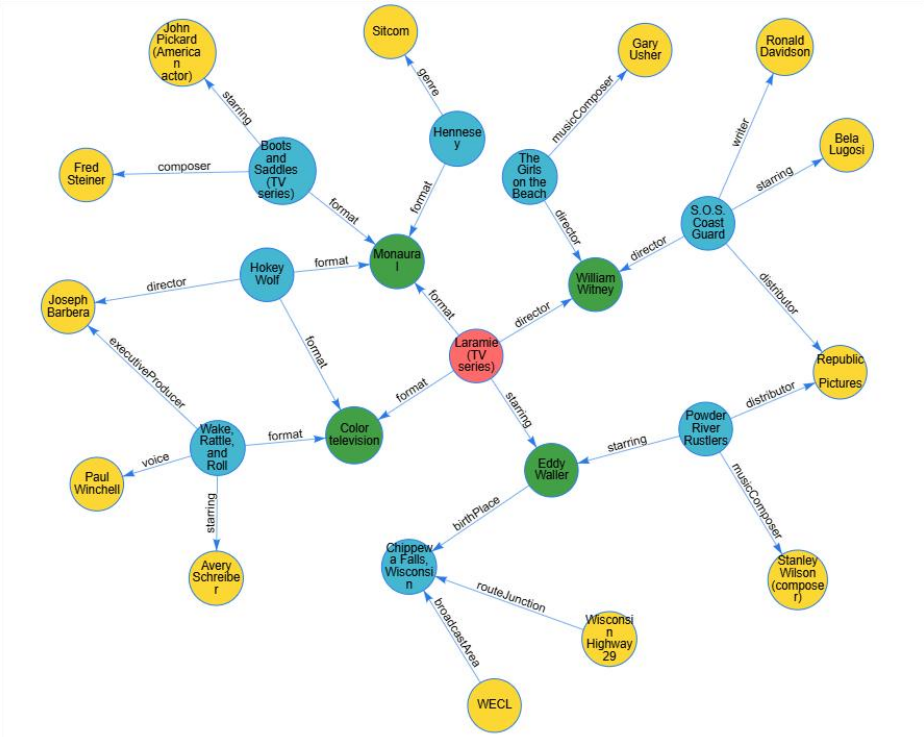


Figure 12. Case study example-4 of VKG-QA on GPT-4.

Table 6. VKG images examples and the corresponding partial question-answer pairs.

VKG	Question	Ground Truth
Figure 19	How many different colors are there in the nodes of the given image? Just provide the number. Do not include any extra words or symbols.	2
Figure 20	The image shows a directed knowledge graph, what is the maximum node degree in the subgraph? Please only answer the quantity.	5
Figure 21	Based only on the given image, who directed the film The Browning Version?	Mike Figgis
Figure 22	In the given image, find the nodes whose names start with 'Pr'. Return only the names of these nodes without including any extra words or symbols.	Private Browsing
Figure 23	The image shows a directed knowledge graph, please identify and extract all triples that involve the node 'An Inconvenient Truth' in the graph.	[["Gore", "won an Oscar for", "An Inconvenient Truth"]]
Figure 24	Based only on the given image, which films were also written by the screenwriter of The Beaver?	Scenic Route
Figure 25	Based only on the given image, which director has directed the largest number of movies? Only provide the director name, without any extra words or symbols.	Douglas Sirk
Figure 26	Based only on the given image, the films that the filmmakers share with the film Xanadu are written by whom?	Ernest Thompson
Figure 27	Based only on the given image, which city, Chicago or Illinois, is the birthplace of the rapper named Omen? Please answer directly the city name.	Illinois



Question : Based only on the given image, which work was directed by **William Witney** and also starred **Eddy Waller**? Please answer with the name only.

Ground Truth : Laramie

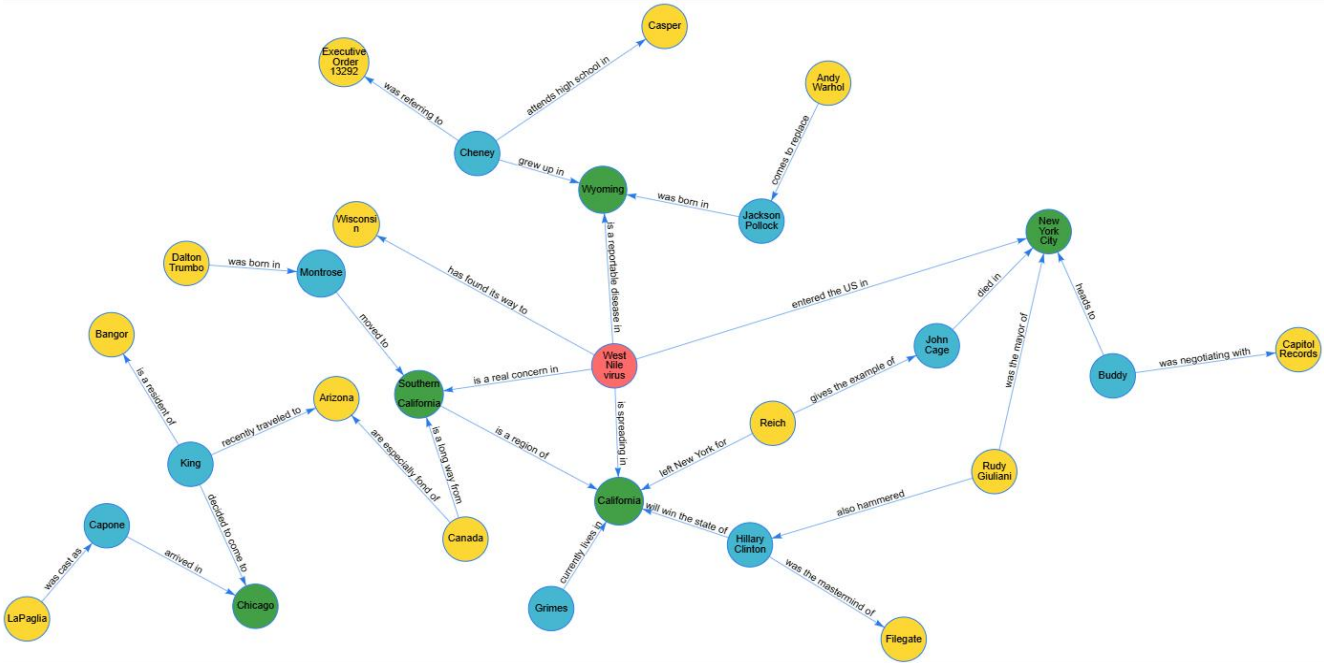
Model Answer :

Reasoning:

- step 1: Identify the works directed by **William Witney** in the image. The image shows **S.O.S. Coast Guard and Powder River Rustlers** as works directed by William Witney.
- step 2: Identify the works starring **Eddy Waller** in the image. The image shows **Powder River Rustlers** as a work starring Eddy Waller.
- step 3: Find the common work between the two lists from step 1 and step 2. Powder River Rustlers is the common work directed by William Witney and starring Eddy Waller.

Final Answer: Powder River Rustlers ✖

Figure 13. Case study example-5 of VKG-QA on GPT-4.



Question : The image shows a directed knowledge graph, does the set of nodes: **'{'Wyoming'}'** constitute a strongly connected subgraph? Please only answer with 'yes' or 'no'?

Ground Truth : yes

Model Answer :

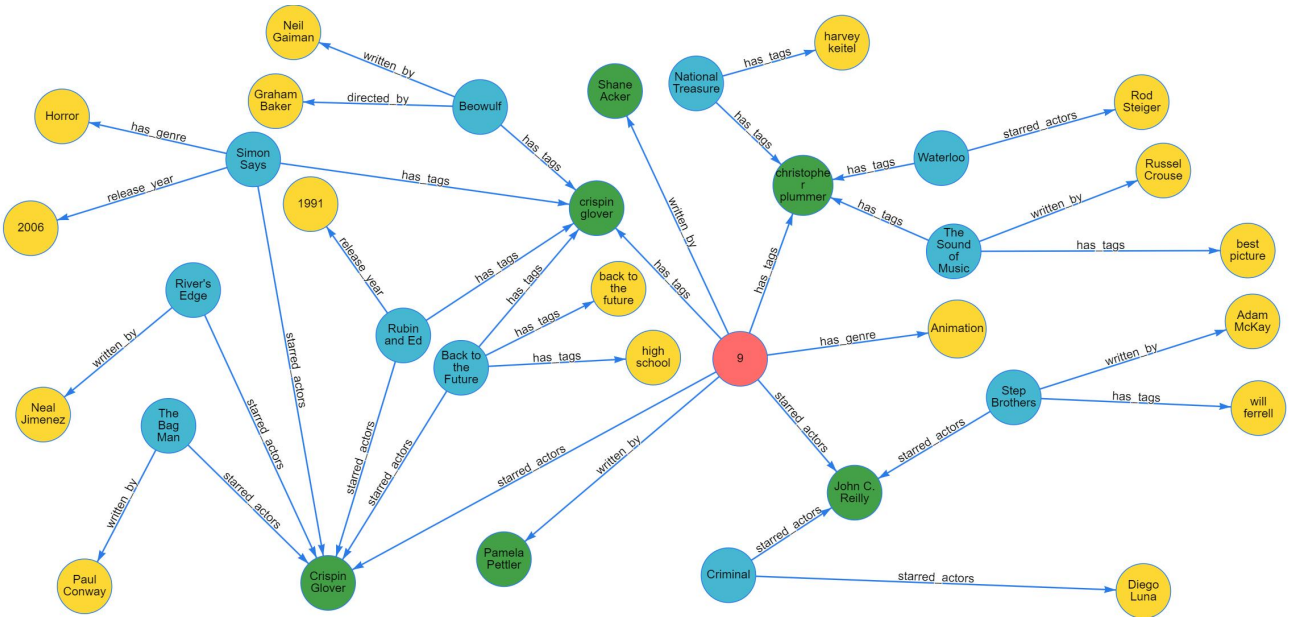
Reasoning:

- step 1: Identify the node in question, which is **'Wyoming'**.
- step 2: Determine if there are any **directed edges** starting from 'Wyoming' and ending at 'Wyoming'
- step 3: Check if 'Wyoming' can be **reached from itself** following the direction of the edges.

Final Answer: No

✖

Figure 14. Case study example-6 of VKG-QA on GPT-4.



Question : Based only on the given image, which **label** corresponds to a movie released in 1991? Only provide the label, without any extra words or symbols.

Ground Truth : crispin glover

Model Answer :

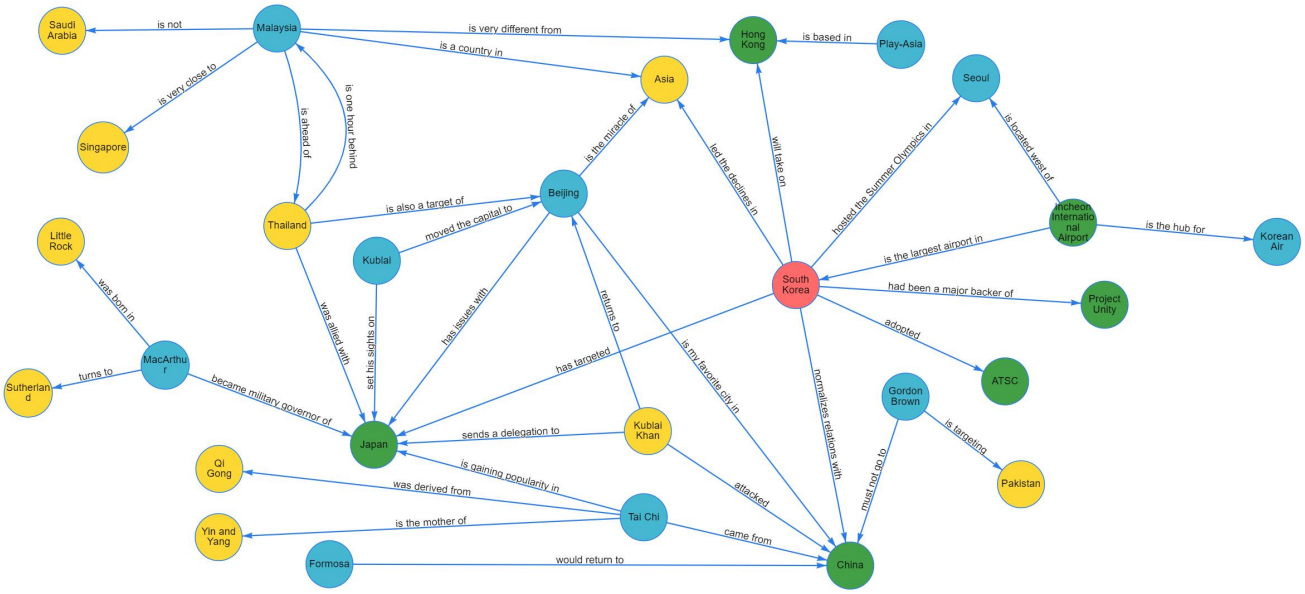
Reasoning:

- step 1: Identify the node labeled 1991 in the image.
- step 2: Follow the arrow labeled release year that points from 1991 to another node.
- step 3: Determine the label of the node that the release year arrow points to.

Final Answer: Rubin and Ed

✖

Figure 15. Case study example-7 of VKG-QA on GPT-4.



Question : Count the total number of edges in the given image. Output only a numeric value, without any extra words or symbols.

Ground Truth : 37

Model Answer :

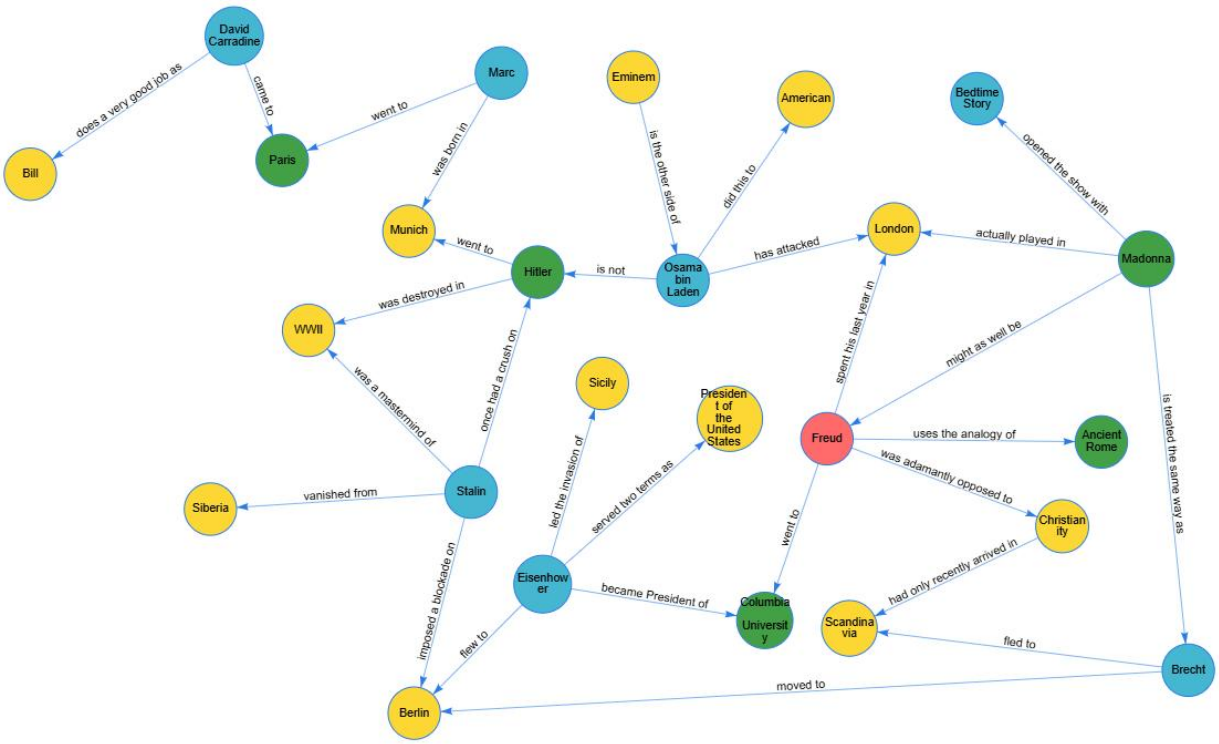
Reasoning:

- step 1: Identify all the nodes in the image.
- step 2: Count the number of edges (lines with arrows) connecting these nodes.
- step 3: Ensure each edge is counted only once, regardless of direction.

Final Answer: 38

✖

Figure 16. Case study example-8 of VKG-QA on GPT-4.



Question : In this image, List all the nodes in the simple cycle that contains node **'American'**. Please only answer node sequence in order if possible, otherwise output 'no'.

Ground Truth : no

Model Answer :

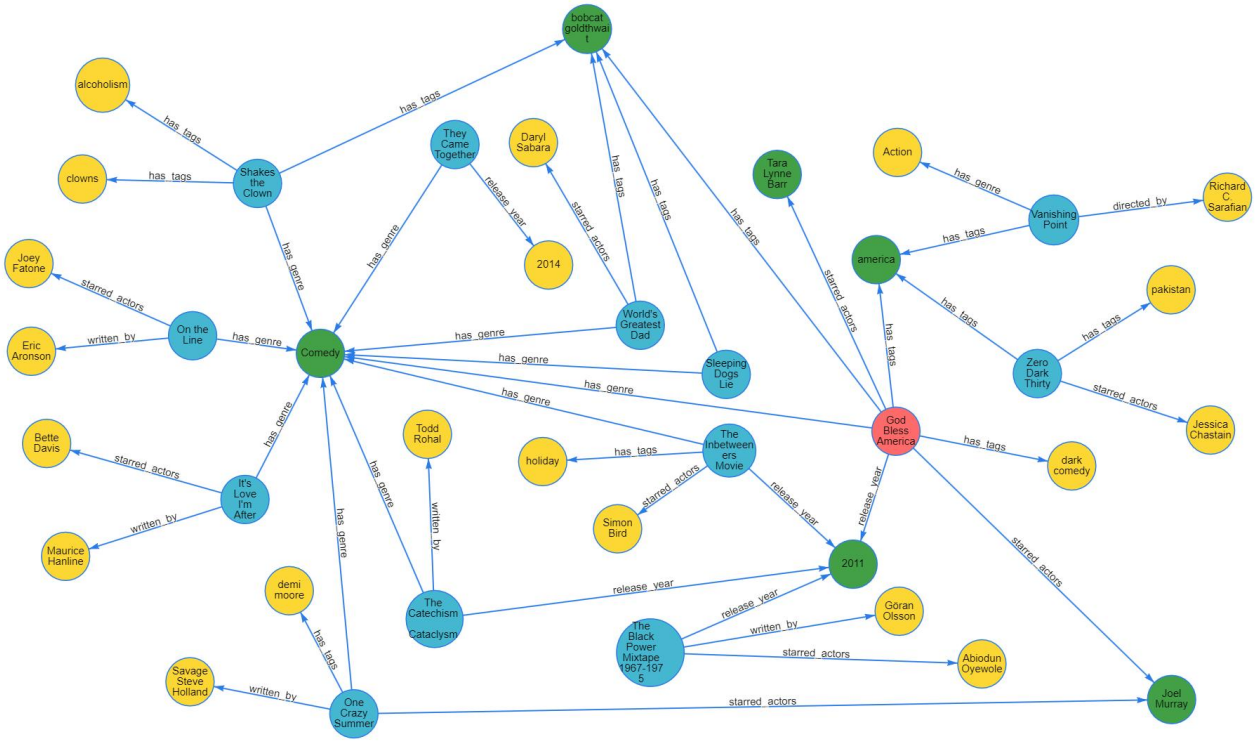
Reasoning:

- step 1: Identify the node **'American'** in the image.
- step 2: Trace the directed edges **starting** from 'American' to find a cycle.
- step 3: Verify if the cycle **returns** to 'American' and list the nodes in order.

Final Answer: American -> London -> Madonna -> Bedtime Story -> American

✖

Figure 17. Case study example-9 of VKG-QA on GPT-4.



Question : Based only on the given image, who is listed as director of the films starred by God Bless America actors.

Ground Truth : Savage Steve Holland

Model Answer :

Reasoning:

- step 1: Identify the actors starred in **God Bless America** from the image. The actors are **Joel Murray and Tara Lynne Barr**.
- step 2: Look for other **films** in the image that these actors have starred in.
- step 3: Determine the **director** of those films.

Final Answer: Based only on the given image, **Bobcat Goldthwait** is listed as the director of the films starred by God Bless America actors. ✖

Figure 18. Case study example-10 of VKG-QA on GPT-4.

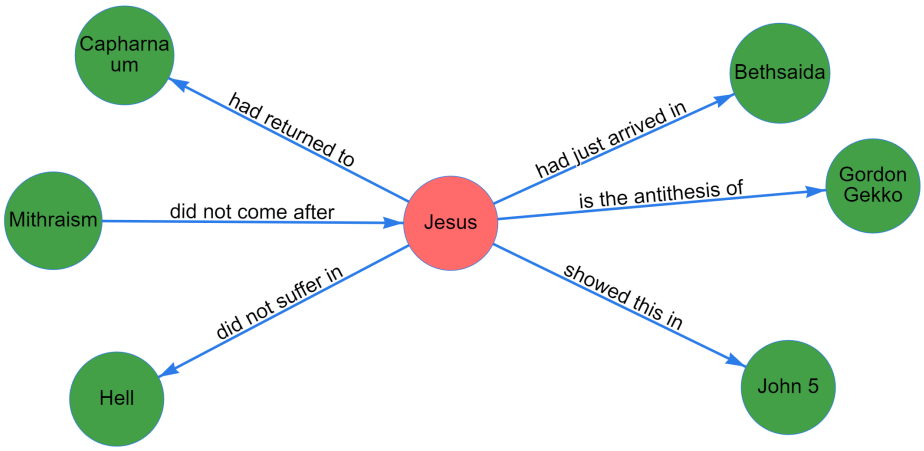


Figure 19. Example 1. The corresponding question is: ”How many different colors are there in the nodes of the given image? Just provide the number. Do not include any extra words or symbols.” The answer is 2.

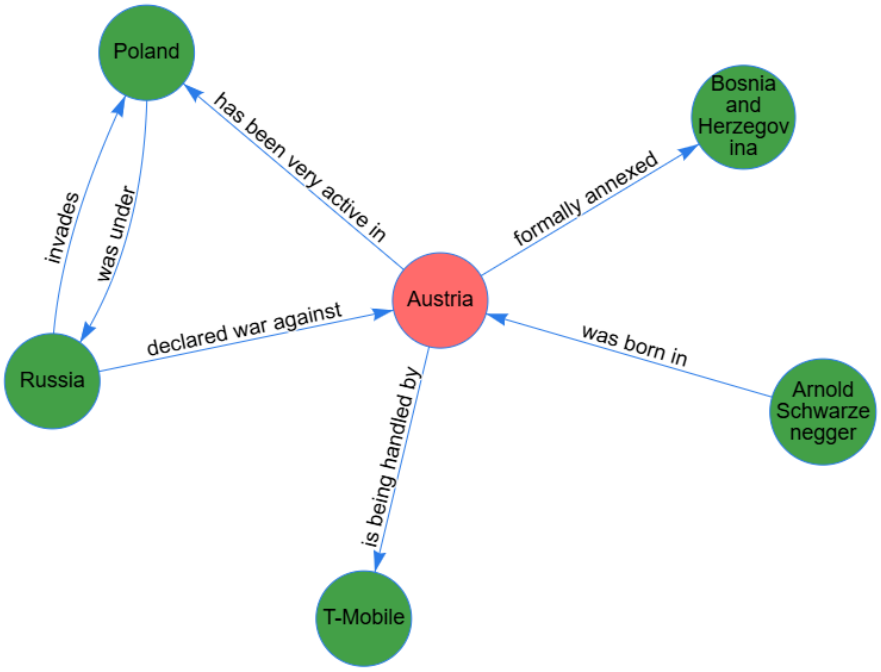


Figure 20. Example 2. The corresponding question is: "The image shows a directed knowledge graph, what is the maximum node degree in the subgraph? Please only answer the quantity." The answer is 5.

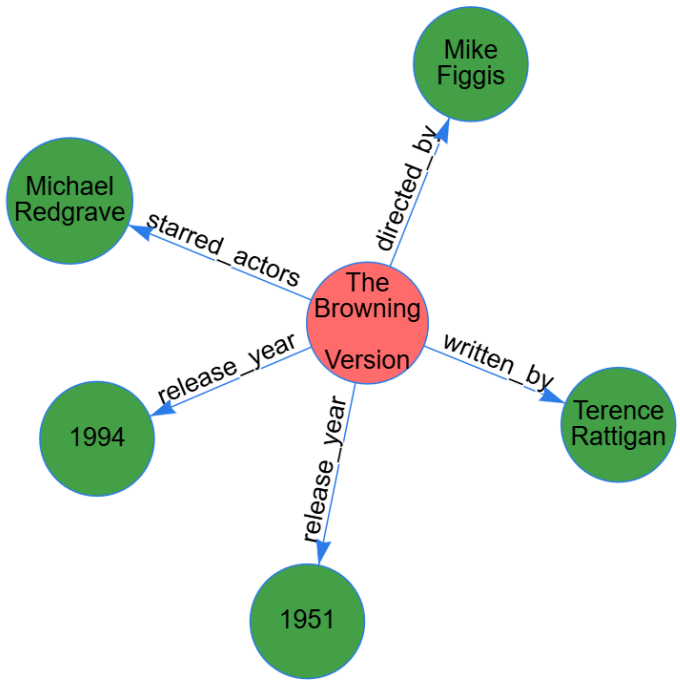


Figure 21. Example 3. The corresponding question is: "Based only on the given image, who directed the film The Browning Version?." The answer is Mike Figgis.

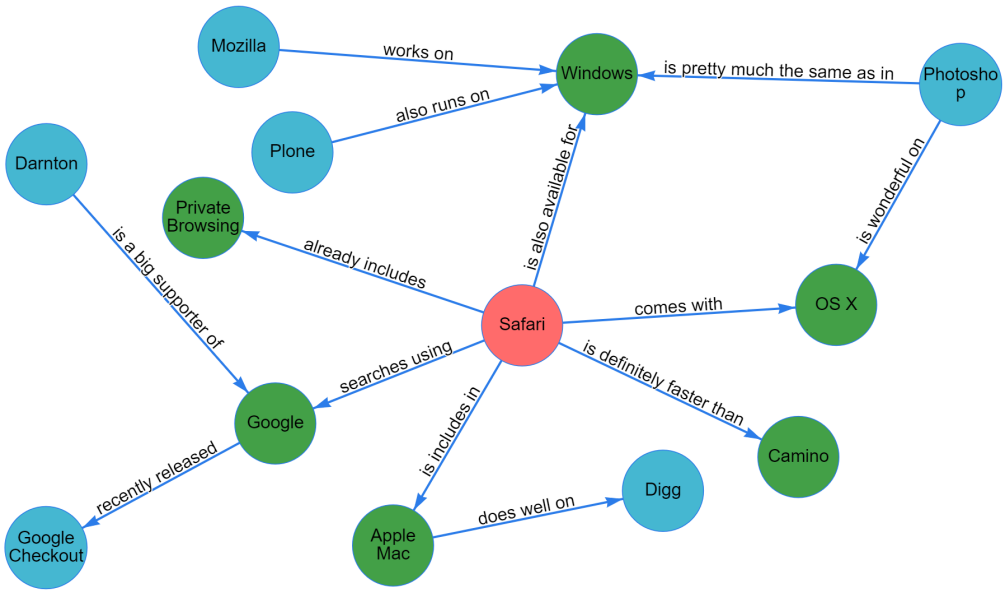


Figure 22. Example 4.The corresponding question is: ”In the given image, find the nodes whose names start with ’Pr’. Return only the names of these nodes without including any extra words or symbols.” The answer is Private Browsing.

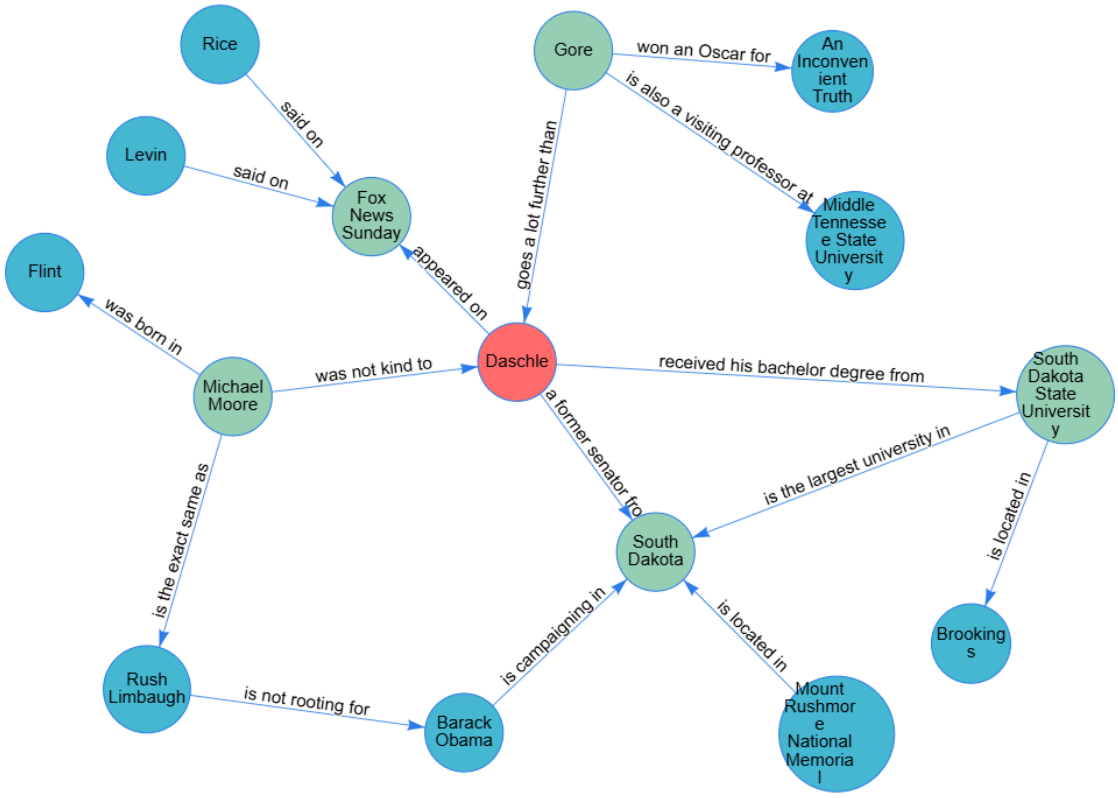


Figure 23. Example 5. The corresponding question is: ”The image shows a directed knowledge graph, Please identify and extract all triples that involve the node ’An Inconvenient Truth’ in the graph.” The answer is [[”Gore”, ”won an Oscar for”, ”An Inconvenient Truth”]].

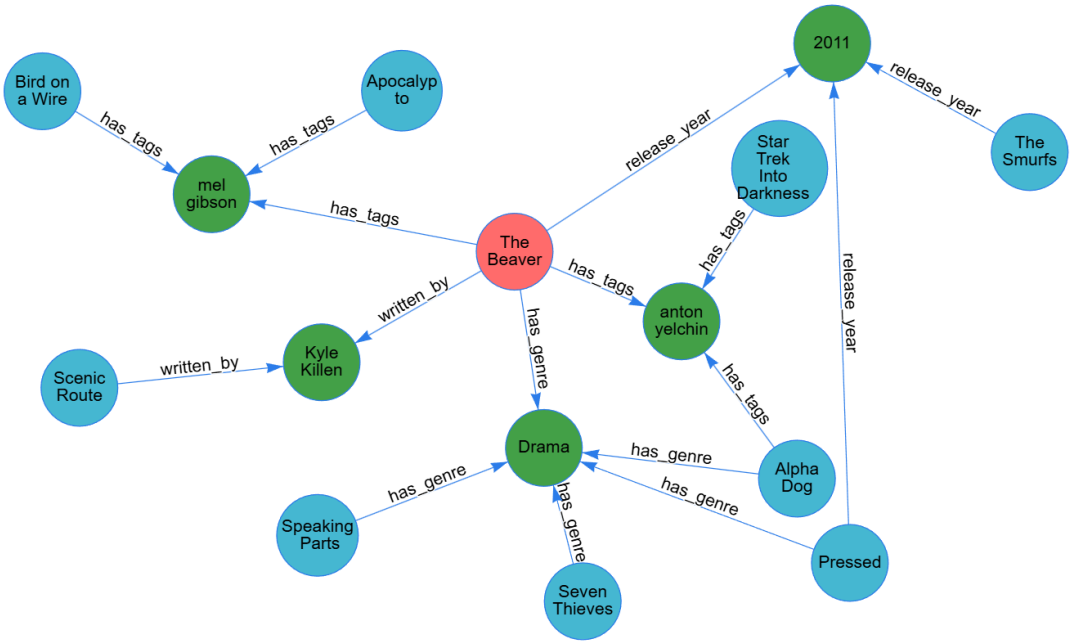


Figure 24. Example 6. The corresponding question is: ”Based only on the given image, which films were also written by the screenwriter of The Beaver?” The answer is Scenic Route.

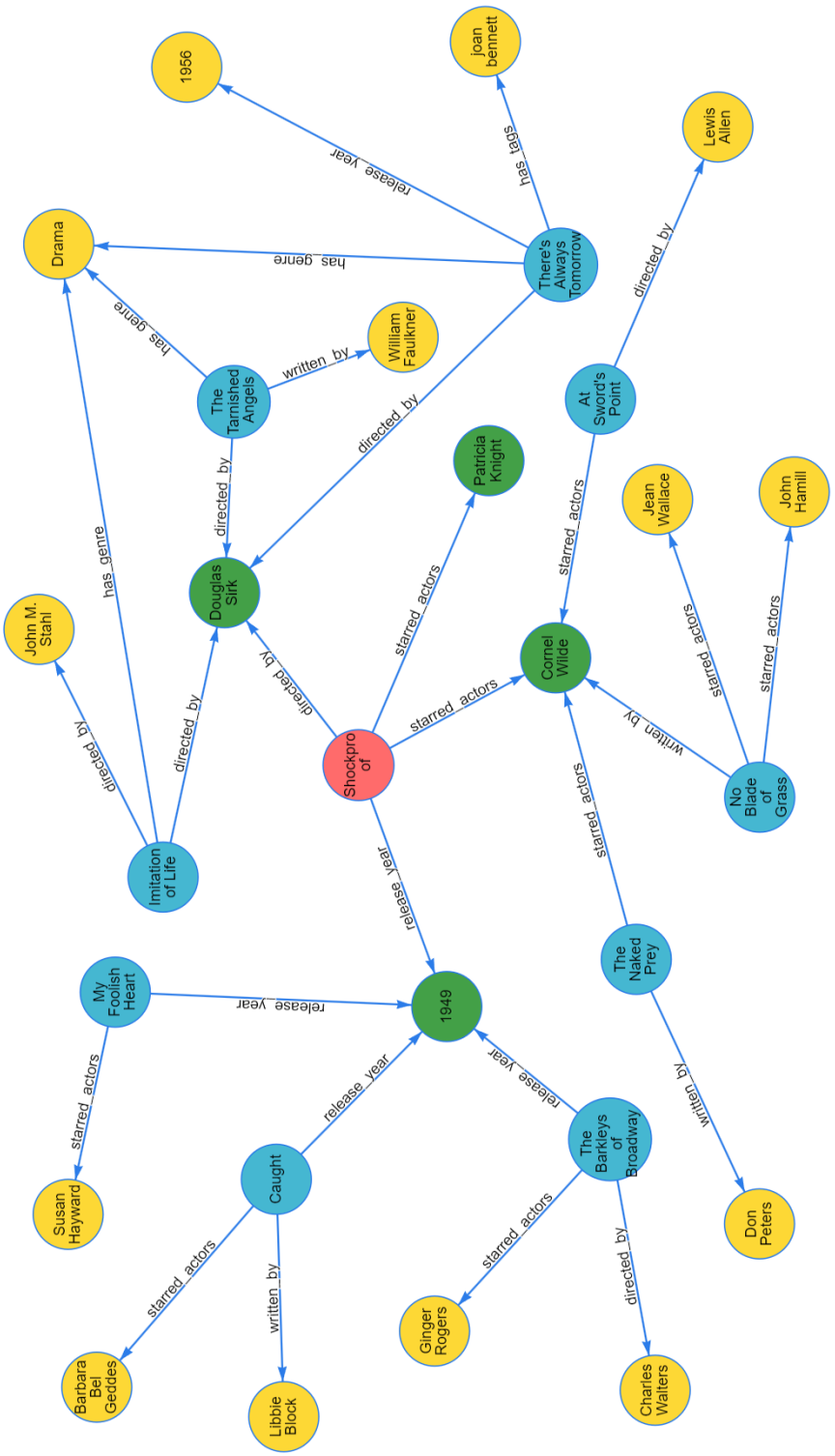


Figure 25. Example 7. The corresponding question is: ”Based only on the given image, which director has directed the largest number of movies? Only provide the director name, without any extra words or symbols.” The answer is Douglas Sirk.

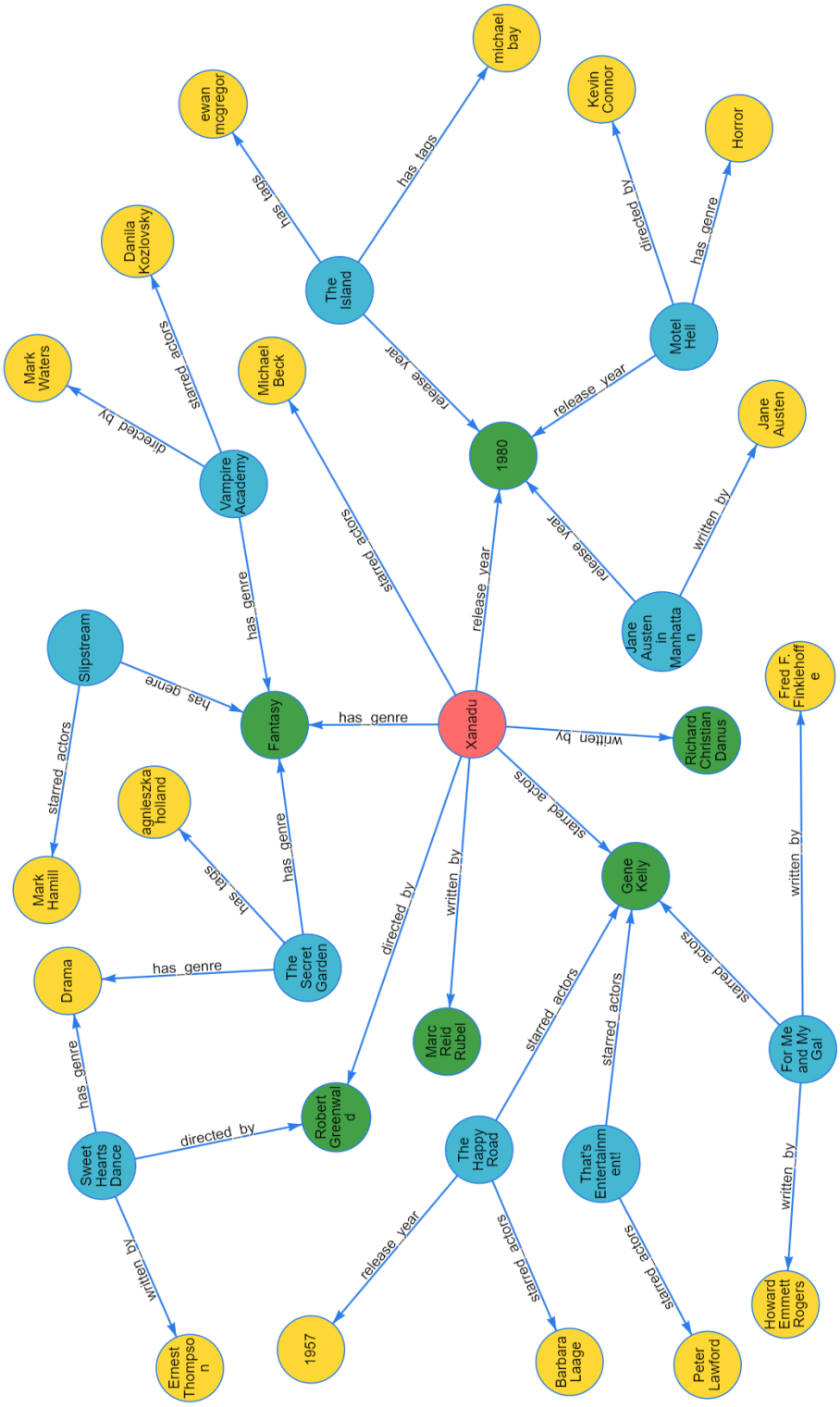


Figure 26. Example 8. The corresponding question is: ”Based only on the given image, the films that the filmmakers share with the film Xanadu are written by whom?” The answer is Ernest Thompson.

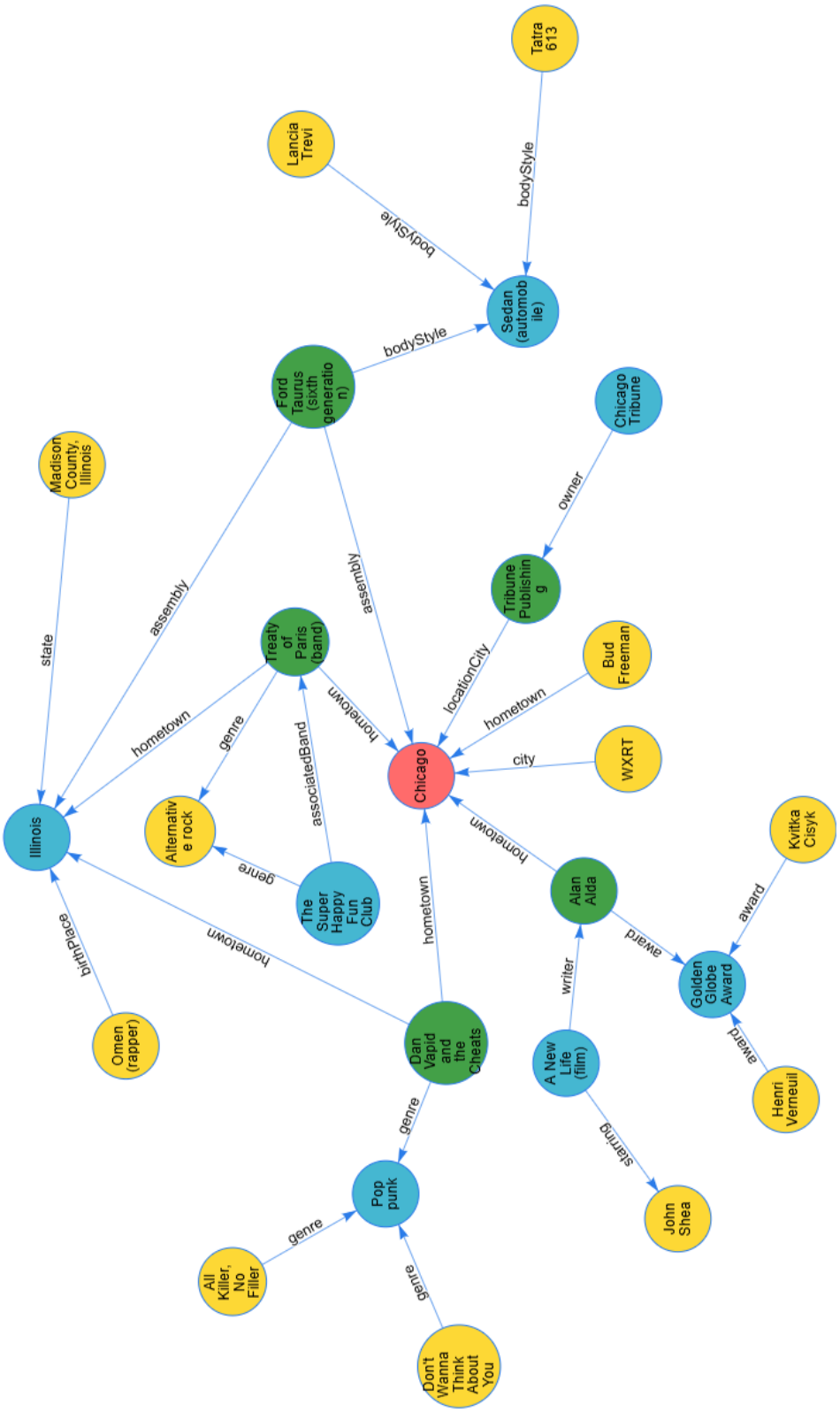


Figure 27. Example 9. The corresponding question is: ”Based only on the given image, which city, Chicago or Illinois, is the birthplace of the rapper named Omen? Please answer directly the city name.” The answer is Illinois.