In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, r2_score
```

In [2]:
```
dataset = pd.read_csv("Most popular 1000 Youtube videos.csv")
```

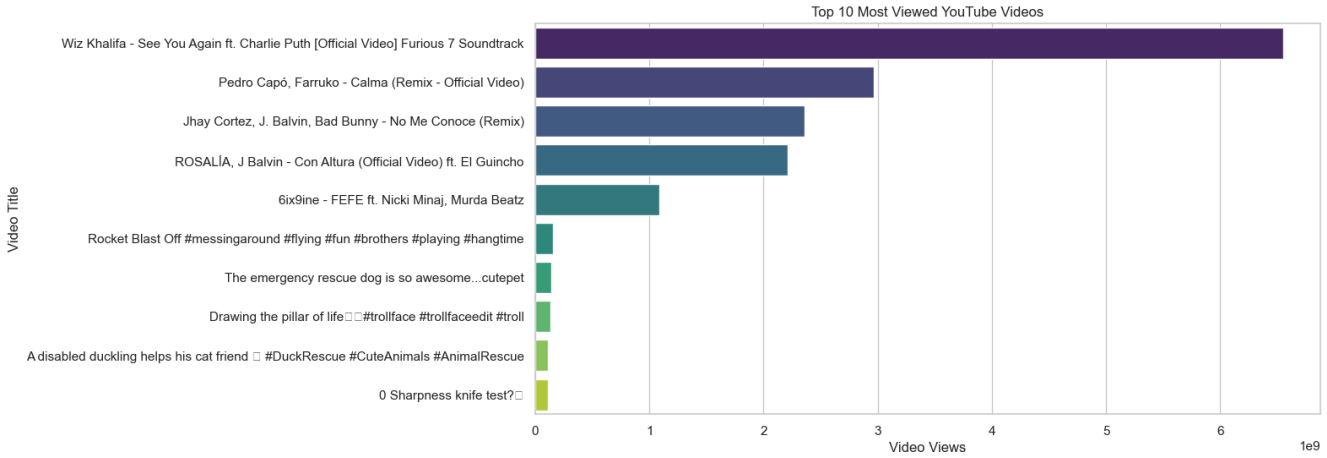
In [3]:
```
print(dataset.info())

# Convert numeric columns (remove commas)
dataset["Video views"] = dataset["Video views"].str.replace(",", "").astype(float)
dataset["Likes"] = dataset["Likes"].str.replace(",", "").astype(float)
dataset["Dislikes"] = dataset["Dislikes"].str.replace(",", "").astype(float)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   rank         1000 non-null   int64
 1   Video        1000 non-null   object
 2   Video views  1000 non-null   object
 3   Likes        1000 non-null   object
 4   Dislikes     527 non-null    object
 5   Category     982 non-null    object
 6   published    1000 non-null   int64
dtypes: int64(2), object(5)
memory usage: 54.8+ KB
None
```

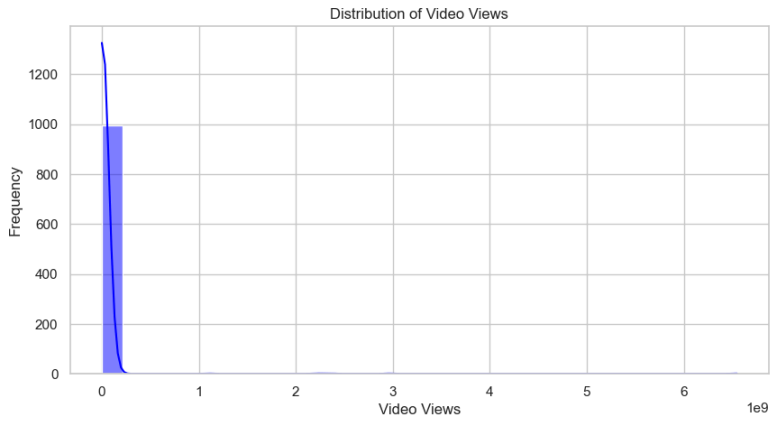
In [4]:
```
# Handle missing values
dataset["Dislikes"].fillna(0, inplace=True)
dataset["Category"].fillna("Unknown", inplace=True)
```

In [6]:
```
# --- EDA ---
sns.set_theme(style="whitegrid")

# Top 10 Most Viewed Videos
top_videos = dataset.nlargest(10, "Video views")
plt.figure(figsize=(12, 6))
sns.barplot(y=top_videos["Video"], x=top_videos["Video views"], palette="viridis")
plt.xlabel("Video Views")
plt.ylabel("Video Title")
plt.title("Top 10 Most Viewed YouTube Videos")
plt.show()
```

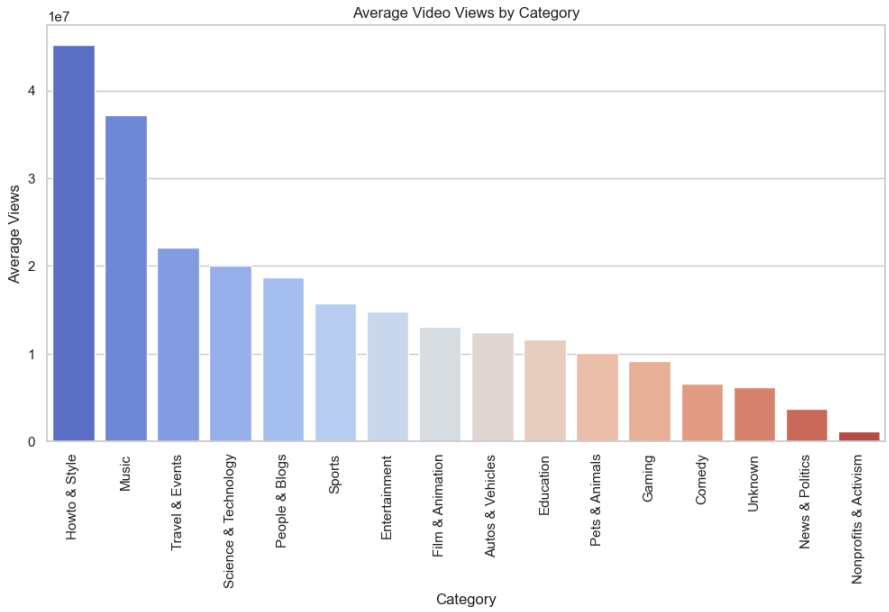
```
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 129398 (\N{FREEZING FACE}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 128511 (\N{MOYAI}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 128037 (\N{FRONT-FACING BABY CHICK}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 128128 (\N{SKULL}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
```



In [7]:
```
plt.figure(figsize=(10, 5))
sns.histplot(dataset["Video views"], bins=30, kde=True, color="blue")
plt.xlabel("Video Views")
plt.ylabel("Frequency")
plt.title("Distribution of Video Views")
plt.show()
```



In [8]:
```
# Category Popularity (Average Views per Category)
category_views = dataset.groupby("Category")["Video views"].mean().sort_values(ascending=False)
plt.figure(figsize=(12, 6))
sns.barplot(x=category_views.index, y=category_views.values, palette="coolwarm")
plt.xticks(rotation=90)
plt.xlabel("Category")
plt.ylabel("Average Views")
plt.title("Average Video Views by Category")
plt.show()
```



In [9]:
```
# --- Machine Learning Model ---

# Encode 'Category' as numeric
le = LabelEncoder()
dataset["Category_encoded"] = le.fit_transform(dataset["Category"])

# Features & Target
features = ["Likes", "Dislikes", "Category_encoded", "published"]
target = "Video views"

X = dataset[features]
y = dataset[target]
```

In [10]:
```
# Train-test split (80% train, 20% test)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

In [11]:
```
# Standardize features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

In [12]:
```
# Train Random Forest Model
model = RandomForestRegressor(n_estimators=100, random_state=42)
model.fit(X_train_scaled, y_train)
```

Out[12]:
```
RandomForestRegressor(random_state=42)
```

In [13]:
```
# Predictions
y_pred = model.predict(X_test_scaled)

# Model evaluation
mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Absolute Error: {mae:.2f}")
print(f"R² Score: {r2:.2f}")
```

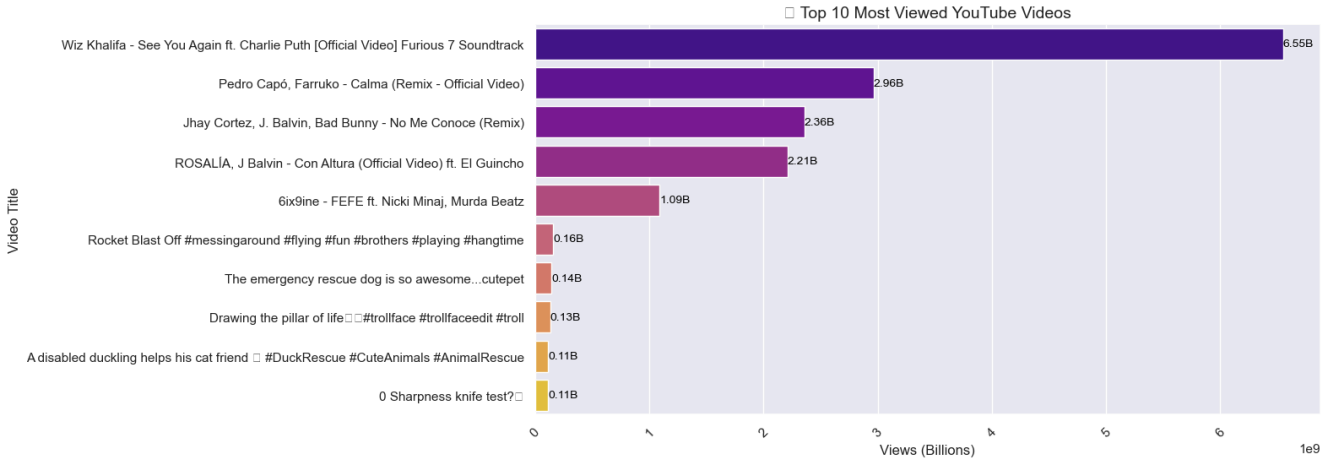
```
Mean Absolute Error: 17371732.29
R² Score: 0.74
```

In [14]:
```
# --- Future Predictions ---
# Example: Predict views for a new video
new_video = pd.DataFrame([[500000, 10000, le.transform(["Music"])[0], 2024]], columns=features)
new_video_scaled = scaler.transform(new_video)
predicted_views = model.predict(new_video_scaled)

print(f"Predicted Views for New Video: {predicted_views[0]:,.0f}")
```

```
Predicted Views for New Video: 19,727,978
```

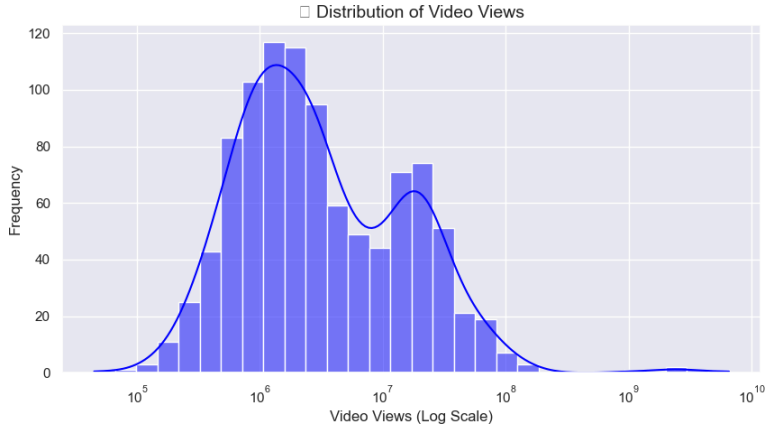
In [15]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np

# Set theme
sns.set_theme(style="darkgrid")
plt.rcParams["axes.titlesize"] = 14
plt.rcParams["axes.labelsize"] = 12

# --- Top 10 Most Viewed Videos ---
top_videos = dataset.nlargest(10, "Video views")

plt.figure(figsize=(12, 6))
ax = sns.barplot(y=top_videos["Video"], x=top_videos["Video views"], palette="plasma")
plt.xlabel("Views (Billions)")
plt.ylabel("Video Title")
plt.title("🔥 Top 10 Most Viewed YouTube Videos")
plt.xticks(rotation=45)

# Annotate views on bars
for i, v in enumerate(top_videos["Video views"]):
    ax.text(v, i, f"{v/1e9:.2f}B", va="center", fontsize=10, color="black")

plt.show()
```

```
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 129398 (\N{FREEZING FACE}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 128511 (\N{MOYAI}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 128037 (\N{FRONT-FACING BABY CHICK}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 128128 (\N{SKULL}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 128293 (\N{FIRE}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
```

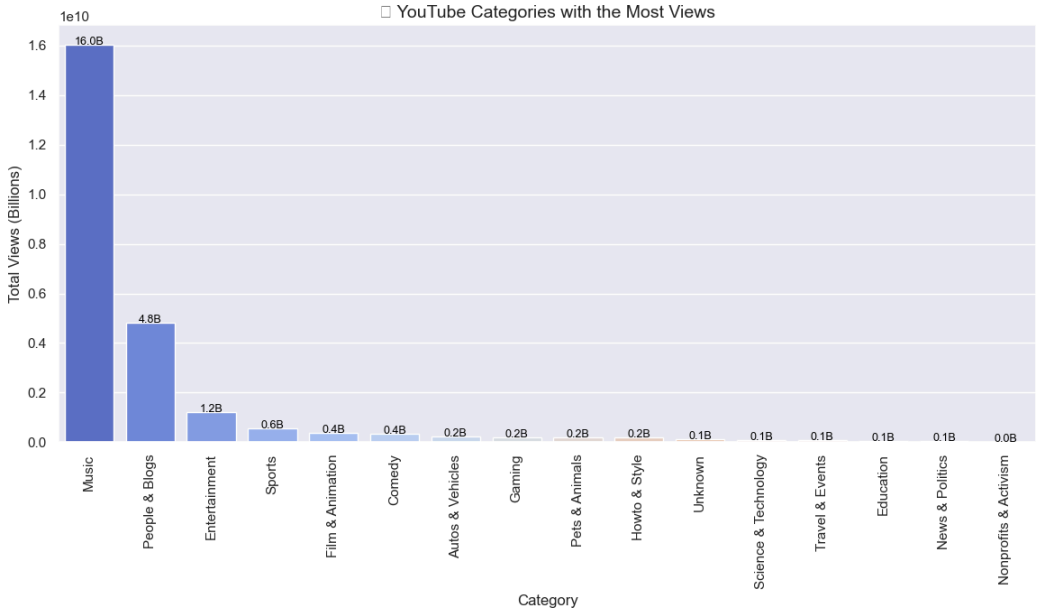


In [17]:
```
# --- Distribution of Video Views ---
plt.figure(figsize=(10, 5))
sns.histplot(dataset["Video views"], bins=30, kde=True, color="blue", log_scale=True)  # Log scale for better clarity
plt.xlabel("Video Views (Log Scale)")
plt.ylabel("Frequency")
plt.title("📊 Distribution of Video Views")
plt.show()

# --- Category Popularity (Total Views Per Category) ---
category_views = dataset.groupby("Category")["Video views"].sum().sort_values(ascending=False)

plt.figure(figsize=(14, 6))
ax = sns.barplot(x=category_views.index, y=category_views.values, palette="coolwarm")
plt.xticks(rotation=90)
plt.xlabel("Category")
plt.ylabel("Total Views (Billions)")
plt.title("🎬 YouTube Categories with the Most Views")

# Annotate total views
for i, v in enumerate(category_views.values):
    ax.text(i, v, f"{v/1e9:.1f}B", ha="center", fontsize=9, color="black")

plt.show()
```

```
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 128202 (\N{BAR CHART}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
```



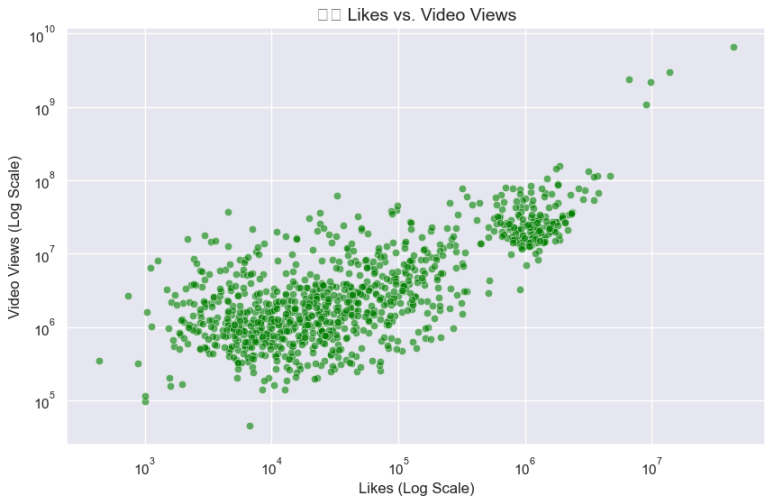

```
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 127916 (\N{CLAPPER BOARD}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
```



In [19]:
```
# --- Likes vs Views Scatter Plot ---
plt.figure(figsize=(10, 6))
sns.scatterplot(x=dataset["Likes"], y=dataset["Video views"], alpha=0.6, color="green")
plt.xscale("log")  # Log scale to spread data points
plt.yscale("log")
plt.xlabel("Likes (Log Scale)")
plt.ylabel("Video Views (Log Scale)")
```

```
plt.title("❤️ Likes vs. Video Views")
plt.show()
```

```
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 10084 (\N{HEAVY BLACK HEART}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
C:\Users\Deviare User\anaconda3\Lib\site-packages\IPython\core\pylabtools.py:152: UserWarning: Glyph 65039 (\N{VARIATION SELECTOR-16}) missing from current font.
  fig.canvas.print_figure(bytes_io, **kw)
```



In [ ]: